# *Hands - On Machine Learning with Scikit - Learn and PyTorch 1st edition Geron Test Bank*

SKU: 9798341607989-TEST-BANK

# Hands-On Machine Learning with Scikit-Learn and PyTorch

ISBN 9798341607989 | Chapter 1: The Machine Learning Landscape

*Total Questions: 150*

## Multiple Choice

**Q1.**

Which definition best matches machine learning as presented in the chapter?

**A) The science and art of programming computers so they can learn from data ✓ Correct**
**B)** The process of storing large datasets so computers can retrieve them quickly
**C)** The design of hardware that can execute algorithms without human oversight
**D)** The automation of software testing through fixed rule sets

*Rationale: The chapter defines machine learning as programming computers to learn from data rather than relying only on explicit rules.*

**Q2.**

According to Tom Mitchell's definition, a program is said to learn when its performance on a task improves with what factor?

**A) Experience ✓ Correct**
**B)** Processor speed
**C)** Network bandwidth
**D)** Storage capacity

*Rationale: Mitchell's definition states that learning occurs when performance on task T improves with experience E as measured by P.*

**Q3.**

In a supervised spam-filtering example, what is the name for the examples used by the system to learn?

**A)** Validation set
**B) Training set ✓ Correct**
**C)** Feature map
**D)** Inference pool

*Rationale: The examples used to teach the system are called the training set.*

**Q4.**

What is each individual example in a training set called?

**A) Training instance ✓ Correct**
**B)** Hyperparameter
**C)** Decision rule
**D)** Deployment artifact

*Rationale: Each example in the training set is called a training instance or sample.*



Hands - On Machine Learning with Scikit - Learn and PyTorch 1st edition Geron Test Bank

**Q5.**

What is the part of a machine learning system that learns from data and makes predictions called?

**A)** Compiler
**B) Model ✓ Correct**
**C)** Kernel cache
**D)** Data schema

*Rationale: The model is the component that learns patterns and produces predictions.*

**Q6.**

In classification tasks such as spam detection, which performance measure is commonly used?

**A)** Latency
**B) Accuracy ✓ Correct**
**C)** Throughput
**D)** Compression ratio

*Rationale: Accuracy, the ratio of correctly classified instances, is commonly used in classification.*

**Q7.**

Which of the following is NOT, by itself, considered machine learning in the chapter?

**A)** Training a spam filter on labeled emails
**B)** Learning a regression model from examples
**C) Downloading a copy of all Wikipedia articles ✓ Correct**
**D)** Updating a classifier using new labeled data

*Rationale: Simply acquiring data does not mean the computer has learned to perform a task better.*

**Q8.**

A traditional rule-based spam filter is described as likely becoming what over time?

**A)** A short list of stable heuristics
**B) A long list of complex rules ✓ Correct**
**C)** A fully self-correcting architecture
**D)** A reinforcement policy network

*Rationale: The chapter notes that traditional solutions often become difficult-to-maintain long lists of rules.*

**Q9.**

What is digging into large amounts of data to discover hidden patterns called?

**A) Data mining ✓ Correct**
**B)** Batch normalization
**C)** Model serving
**D)** Feature hashing

*Rationale: The chapter explicitly refers to discovering hidden patterns in large datasets as data mining.*



Hands - On Machine Learning with Scikit - Learn and PyTorch 1st edition Geron Test Bank

**Q10.**

Which task is described as typically using convolutional neural networks or vision transformers to classify product images?

**A)** Clustering

**B) Image classification ✓ Correct**

**C)** Association rule learning

**D)** Speech synthesis

*Rationale: Automatically classifying product images is identified as image classification.*

**Q11.**

Detecting tumors in brain scans by classifying each pixel is an example of which task?

**A) Semantic image segmentation ✓ Correct**

**B)** Regression forecasting

**C)** Association rule learning

**D)** Instance retrieval

*Rationale: Classifying each pixel to locate tumor boundaries is semantic image segmentation.*

**Q12.**

Automatically classifying news articles is an example of which machine learning area?

**A)** Reinforcement learning

**B) Natural language processing ✓ Correct**

**C)** Dimensionality reduction

**D)** Federated learning

*Rationale: News article classification is described as an NLP task, specifically text classification.*

**Q13.**

Forecasting a company's revenue next year based on performance metrics is primarily what type of task?

**A)** Clustering

**B) Regression ✓ Correct**

**C)** Novelty detection

**D)** Association learning

*Rationale: Predicting a numeric value such as revenue is a regression task.*

**Q14.**

Detecting credit card fraud is presented as an example of which unsupervised task?

**A)** Regression

**B) Anomaly detection ✓ Correct**

**C)** Feature extraction

**D)** Transfer learning

*Rationale: Fraud detection is cited as a typical anomaly detection problem.*

**Q15.**

Segmenting clients based on purchases for different marketing strategies is what type of task?

**A) Clustering ✓ Correct**
**B)** Regression
**C)** Speech recognition
**D)** Question answering

*Rationale: Grouping similar clients into segments is a clustering task.*

**Q16.**

Building an intelligent game bot that learns actions to maximize rewards over time belongs to which branch of machine learning?

**A)** Semi-supervised learning
**B) Reinforcement learning ✓ Correct**
**C)** Association rule learning
**D)** Dimensionality reduction

*Rationale: Learning actions from rewards in an environment defines reinforcement learning.*

**Q17.**

How are machine learning systems classified according to the amount and type of supervision during training?

**A)** Centralized versus decentralized
**B)** Deterministic versus probabilistic
**C) Supervised, unsupervised, semi-supervised, self-supervised, and reinforcement learning ✓ Correct**
**D)** Linear versus nonlinear

*Rationale: The chapter lists these as the main supervision-based categories.*

**Q18.**

In supervised learning, the desired solutions included in the training set are called what?

**A)** Clusters
**B) Labels ✓ Correct**
**C)** Rewards
**D)** Embeddings

*Rationale: Supervised learning uses labeled examples containing the desired outputs.*

**Q19.**

Which pair represents the two most common supervised learning tasks?

**A)** Clustering and anomaly detection
**B) Classification and regression ✓ Correct**
**C)** Visualization and association rule learning
**D)** Policy learning and transfer learning

*Rationale: The chapter identifies classification and regression as the two most common supervised tasks.*



Hands - On Machine Learning with Scikit - Learn and PyTorch 1st edition Geron Test Bank

**Q20.**

In supervised learning terminology, which word is more commonly used in regression tasks than in classification tasks?

**A)** Label
**B) Target ✓ Correct**
**C)** Reward
**D)** Sample

*Rationale: The chapter notes that target is more common in regression, while label is more common in classification.*

**Q21.**

In unsupervised learning, the training data is best described as what?

**A)** Fully labeled
**B)** Paired with rewards
**C) Unlabeled ✓ Correct**
**D)** Always balanced across classes

*Rationale: Unsupervised learning works with unlabeled data.*

**Q22.**

Which unsupervised task aims to detect groups of similar instances without predefined labels?

**A)** Regression
**B) Clustering ✓ Correct**
**C)** Fine-tuning
**D)** Policy optimization

*Rationale: Clustering discovers groups in unlabeled data.*

**Q23.**

What is the goal of dimensionality reduction?

**A)** To increase the number of labels available for training
**B) To simplify data without losing too much information ✓ Correct**
**C)** To convert supervised tasks into reinforcement tasks
**D)** To guarantee perfect generalization

*Rationale: Dimensionality reduction seeks a simpler representation while preserving important information.*

**Q24.**

Merging several correlated features into one more useful feature is called what?

**A) Feature extraction ✓ Correct**
**B)** Data drift
**C)** Cross-validation
**D)** Sampling bias

*Rationale: Combining correlated features into a reduced representation is feature extraction.*



Hands - On Machine Learning with Scikit - Learn and PyTorch 1st edition Geron Test Bank

**Q25.**

Which unsupervised task is trained mostly on normal instances so it can identify unusual ones later?

**A)** Association rule learning

**B) Anomaly detection ✓ Correct**

**C)** Regression

**D)** Transfer learning

*Rationale: Anomaly detection learns what normal data looks like and flags unusual cases.*

**Q26.**

What is the key distinction between novelty detection and anomaly detection in the chapter?

**A)** Novelty detection requires labeled anomalies during training

**B) Novelty detection requires a very clean training set without examples of what should later be detected ✓ Correct**

**C)** Anomaly detection can only be used for images

**D)** Anomaly detection always performs better on rare classes

*Rationale: Novelty detection assumes the training set is clean and free of instances the system should later identify as novel.*

**Q27.**

Association rule learning is primarily used to discover what?

**A)** Optimal neural architectures

**B) Interesting relations between attributes ✓ Correct**

**C)** The exact value of a target variable

**D)** The shortest training time

*Rationale: Association rule learning uncovers relationships such as items frequently purchased together.*

**Q28.**

What describes semi-supervised learning?

**A)** Training only on rewards and penalties

**B) Training on data that is partially labeled ✓ Correct**

**C)** Training on a fully labeled dataset generated by simulation only

**D)** Training without any data preprocessing

*Rationale: Semi-supervised learning uses a mix of labeled and unlabeled data.*

**Q29.**

What describes self-supervised learning in the chapter?

**A)** Learning only from user-provided class labels

**B) Generating labels from unlabeled data so supervised-style training can be performed ✓ Correct**

**C)** Training a model only once and never updating it

**D)** Using a policy to maximize cumulative reward

*Rationale: Self-supervised learning creates labels from unlabeled data, such as masked portions to predict.*

**Q30.**

Transferring knowledge learned for one task to help with another task is called what?

**A)** Association learning
**B) Transfer learning ✓ Correct**
**C)** Sampling correction
**D)** Out-of-core learning

*Rationale: The chapter defines transfer learning as reusing knowledge across tasks.*

**Q31.**

In reinforcement learning, what is the learning system called?

**A)** Estimator
**B) Agent ✓ Correct**
**C)** Validator
**D)** Sampler

*Rationale: The acting and learning entity in reinforcement learning is called an agent.*

**Q32.**

In reinforcement learning, what is the name for the strategy that specifies which action an agent should choose in a given situation?

**A)** Cluster
**B) Policy ✓ Correct**
**C)** Feature map
**D)** Regularizer

*Rationale: A policy defines the agent's action choices in different situations.*

**Q33.**

What is batch learning?

**A)** Learning incrementally from one instance at a time in production only
**B) Training from scratch using all available data ✓ Correct**
**C)** Learning only from unlabeled data
**D)** Training many models and averaging their outputs

*Rationale: Batch learning trains on the full dataset rather than updating incrementally.*

**Q34.**

When a batch learning system is trained and then runs without learning anymore in production, this is called what?

**A) Offline learning ✓ Correct**
**B)** Federated learning
**C)** Meta-learning
**D)** Self-supervised learning

*Rationale: A model trained first and then deployed without further learning is described as offline learning.*



Hands - On Machine Learning with Scikit - Learn and PyTorch 1st edition Geron Test Bank

**Q35.**

What term does the chapter use for the gradual decay in a model's performance as the world changes while the model remains unchanged?

**A)** Feature collapse
**B) Data drift ✓ Correct**
**C)** Label smoothing
**D)** Model averaging

*Rationale: The chapter refers to this degradation over time as data drift, also called model rot.*

**Q36.**

What is online learning?

**A)** Training only on internet-based datasets
**B) Training incrementally by feeding instances sequentially or in mini-batches ✓ Correct**
**C)** Running a model in a web browser
**D)** Evaluating a model only after deployment

*Rationale: Online learning updates the model incrementally as new data arrives.*

**Q37.**

What is out-of-core learning?

**A)** Training a model without any memory usage
**B) Using online learning techniques to train on datasets too large to fit in memory ✓ Correct**
**C)** Performing inference on data stored outside the organization
**D)** Removing outliers before model training

*Rationale: Out-of-core learning handles huge datasets by loading and training on parts of the data at a time.*

**Q38.**

In online learning, what does the learning rate control?

**A)** The amount of disk space used by the model
**B) How fast the system adapts to changing data ✓ Correct**
**C)** The number of labels in the dataset
**D)** The similarity measure used for prediction

*Rationale: The learning rate determines how quickly the model incorporates new information.*

**Q39.**

What is catastrophic forgetting in online learning?

**A)** The inability to process unlabeled data
**B) The tendency to rapidly forget old data when adapting quickly to new data ✓ Correct**
**C)** A failure caused by insufficient CPU speed
**D)** A test-set leakage problem during validation

*Rationale: With a high learning rate, the system may adapt fast but forget previously learned patterns.*

**Q40.**

Which type of learning generalizes by comparing new cases with remembered examples using a similarity measure?

**A)** Model-based learning

**B) Instance-based learning ✓ Correct**

**C)** Reinforcement learning

**D)** Batch learning

*Rationale: Instance-based learning stores examples and predicts by similarity to known instances.*

**Q41.**

Which type of learning generalizes by building a predictive model from the training data?

**A)** Instance-based learning

**B) Model-based learning ✓ Correct**

**C)** Association learning

**D)** Novelty detection

*Rationale: Model-based learning fits a model to training data and uses it for prediction.*

**Q42.**

In the linear life-satisfaction example, choosing a linear function of GDP per capita is an example of what?

**A) Model selection ✓ Correct**

**B)** Cross-validation

**C)** Feature scaling

**D)** Anomaly scoring

*Rationale: Selecting a linear relationship as the representation of the data is model selection.*

**Q43.**

What is a cost function in model-based learning?

**A) A measure of how bad a model's predictions are ✓ Correct**

**B)** A list of hyperparameter values to try

**C)** A method for reducing the number of features

**D)** A way to split data into clusters

*Rationale: A cost or loss function quantifies prediction error so training can minimize it.*

**Q44.**

What is the process of running an algorithm to find the parameter values that best fit the training data called?

**A)** Inference

**B) Training the model ✓ Correct**

**C)** Deployment

**D)** Data augmentation

*Rationale: Training is the step where the algorithm adjusts model parameters to fit the data.*

**Q45.**

Applying a trained model to make predictions on new cases is called what?

**A)** Regularization
**B) Inference ✓ Correct**
**C)** Sampling
**D)** Labeling

*Rationale: The chapter calls prediction on new data inference.*

**Q46.**

Which challenge refers to having too few examples for a learning algorithm to capture the underlying patterns reliably?

**A) Insufficient quantity of training data ✓ Correct**
**B)** Catastrophic forgetting
**C)** Holdout validation
**D)** Model serving drift

*Rationale: Too little data is explicitly identified as a major challenge in machine learning.*

**Q47.**

What is sampling bias?

**A)** The use of too many model parameters
**B) Nonrepresentative data caused by a flawed sampling method ✓ Correct**
**C)** A failure to regularize a model sufficiently
**D)** A mismatch between training and test metrics

*Rationale: Sampling bias occurs when the method used to collect data produces a nonrepresentative sample.*

**Q48.**

What is feature engineering?

**A)** The automatic replacement of all missing labels
**B) The process of selecting, extracting, and creating useful features ✓ Correct**
**C)** The deployment of a model to production servers
**D)** The use of only neural networks for prediction

*Rationale: Feature engineering includes selecting useful features, extracting better ones, and creating new ones from additional data.*

**Q49.**

What is overfitting?

**A)** When a model is too simple to capture the data structure
**B) When a model performs well on training data but generalizes poorly to new instances ✓ Correct**
**C)** When a model uses too few features and too much regularization
**D)** When a model is evaluated only once on a test set

*Rationale: Overfitting means the model has adapted too closely to training data and fails to generalize.*



Hands - On Machine Learning with Scikit - Learn and PyTorch 1st edition Geron Test Bank

**Q50.**

What is a hyperparameter?

**A)** A parameter learned directly from the training data during model fitting

**B) A parameter of the learning algorithm set before training and kept constant during training ✓ Correct**

**C)** A measure of similarity between training instances

**D)** A subset of the test set used for deployment

*Rationale: A hyperparameter controls the learning process itself and is chosen prior to training rather than learned from the data.*

**Q51.**

A hospital wants to build a system that flags potentially fraudulent credit card payments made through its patient portal. Which machine learning task best matches this use case?

**A)** Clustering

**B) Anomaly detection ✓ Correct**

**C)** Regression

**D)** Dimensionality reduction

*Rationale: Fraud detection is a classic anomaly detection task because the goal is to identify unusual transactions that differ from normal behavior.*

**Q52.**

A health system has millions of unlabeled radiology images but only a small labeled set for disease classification. The data science team first trains a model to reconstruct masked portions of images, then fine-tunes it for diagnosis. Which approach are they using?

**A) Self-supervised learning followed by transfer learning ✓ Correct**

**B)** Reinforcement learning followed by online learning

**C)** Association rule learning followed by clustering

**D)** Instance-based learning followed by batch scoring

*Rationale: Training on masked-image recovery generates labels from unlabeled data, and fine-tuning for diagnosis is transfer learning.*

**Q53.**

A nurse informaticist wants to group patients into utilization profiles based on visit frequency, diagnoses, and medication burden, without predefined categories. Which method is most appropriate?

**A)** Classification

**B) Clustering ✓ Correct**

**C)** Regression

**D)** Regularization

*Rationale: Clustering is used to discover natural groups in unlabeled data.*



Hands - On Machine Learning with Scikit - Learn and PyTorch 1st edition Geron Test Bank

**Q54.**
A public health analyst downloads a massive archive of articles but does not train any algorithm or improve performance on a task. According to the chapter, this situation is best described as:

**A)** Machine learning because more data always implies learning
**B)** Machine learning because storage itself creates a model
**C) Not machine learning because no task performance improved from experience ✓ Correct**
**D)** Not machine learning because only labeled data counts

*Rationale: A system learns only if experience improves performance on a task, not merely by storing data.*

**Q55.**
A coding compliance team uses a model that updates every hour as new payer denials arrive. Each update is small and inexpensive. This is an example of:

**A)** Batch learning
**B)** Offline learning only
**C) Online learning ✓ Correct**
**D)** Association learning

*Rationale: Online learning updates the model incrementally as new data arrives.*

**Q56.**
A biomedical startup trains a model on a dataset too large to fit into memory by loading chunks sequentially and updating the model after each chunk. This approach is called:

**A) Out-of-core learning ✓ Correct**
**B)** Hierarchical clustering
**C)** Offline inference
**D)** Feature selection

*Rationale: Out-of-core learning handles huge datasets by training incrementally on subsets that fit in memory.*

**Q57.**
A clinical decision support team deploys a nearest-neighbors model for symptom triage. The model predicts by comparing each new case to similar past cases. This is best classified as:

**A)** Model-based learning
**B) Instance-based learning ✓ Correct**
**C)** Reinforcement learning
**D)** Semi-supervised learning

*Rationale: Instance-based learning generalizes by comparing new instances to stored examples using similarity.*

**Q58.**
A healthcare economist models patient satisfaction as a function of out-of-pocket spending using a straight-line equation with fitted coefficients. This is an example of:

**A) Model-based learning ✓ Correct**
**B)** Association rule learning
**C)** Novelty detection
**D)** Self-supervised learning

*Rationale: A fitted linear equation is a predictive model built from data, so it is model-based learning.*

**Q59.**
A machine learning team at a payer reports very low training error but poor performance on new claims. What problem is most likely occurring?

**A)** Underfitting
**B) Overfitting ✓ Correct**
**C)** Feature extraction
**D)** Dimensionality reduction

*Rationale: Low training error with poor generalization is the hallmark of overfitting.*

**Q60.**
A model predicting clinic no-shows performs poorly even on the training data. Which issue is most likely?

**A)** Overfitting
**B)** Sampling bias
**C) Underfitting ✓ Correct**
**D)** Catastrophic forgetting

*Rationale: Poor performance on the training set suggests the model is too simple and is underfitting.*

**Q61.**
A health app team repeatedly tries many hyperparameter values and chooses the one that performs best on the test set. Why is this problematic?

**A) The test set becomes part of model selection and gives an overly optimistic estimate ✓ Correct**
**B)** The training set becomes unlabeled and unusable
**C)** The model can no longer be regularized
**D)** The validation set automatically becomes biased

*Rationale: Using the test set for tuning leaks evaluation information and undermines its role as an unbiased estimate of generalization.*

**Q62.**
A telehealth company has abundant web images for training a skin-condition model, but only a small set of app-captured images that truly represent production data. Which datasets should be most representative of production use?

**A)** Only the training set
**B)** Only the train-dev set
**C) The validation set and test set ✓ Correct**
**D)** All sets must contain identical proportions from all sources

*Rationale: The validation and test sets should match production data as closely as possible because they guide selection and final evaluation.*



Hands - On Machine Learning with Scikit - Learn and PyTorch 1st edition Geron Test Bank

**Q63.**

A rehabilitation robot learns to choose movements that maximize successful walking across uneven terrain by receiving rewards and penalties. Which learning paradigm is this?

**A)** Supervised learning
**B) Reinforcement learning ✓ Correct**
**C)** Semi-supervised learning
**D)** Association rule learning

*Rationale: Reinforcement learning trains an agent to choose actions that maximize reward over time.*

**Q64.**

A hospital quality team wants to predict a continuous value: next month's emergency department volume. Which supervised task is this?

**A)** Classification
**B) Regression ✓ Correct**
**C)** Clustering
**D)** Anomaly detection

*Rationale: Predicting a numeric quantity is a regression task.*

**Q65.**

A privacy-conscious mobile keyboard trains language models directly on users' devices without sending raw text to a central server. Which additional ML category from the chapter best describes this approach?

**A) Federated learning ✓ Correct**
**B)** Batch learning
**C)** Association rule learning
**D)** Novelty detection

*Rationale: Federated learning trains across devices while keeping raw user data local.*

**Q66.**

A case management team wants to identify combinations of services that are often used together so care bundles can be redesigned. Which unsupervised task is most appropriate?

**A) Association rule learning ✓ Correct**
**B)** Regression
**C)** Classification
**D)** Reinforcement learning

*Rationale: Association rule learning discovers relationships among attributes or co-occurring items in large datasets.*



Hands - On Machine Learning with Scikit - Learn and PyTorch 1st edition Geron Test Bank

**Q67.**
A health insurer uses a clustering algorithm to group unlabeled members and then assigns each unlabeled member the most common label within its cluster before training a classifier. This workflow is an example of:

**A) Semi-supervised learning ✓ Correct**
**B)** Reinforcement learning
**C)** Pure supervised learning
**D)** Novelty detection

*Rationale: Semi-supervised learning combines a small amount of labeled data with unlabeled data, often using clustering to propagate labels.*

**Q68.**
A speech recognition team says their system improves because they keep writing more handcrafted rules for accents and background noise. Based on the chapter, what is the strongest reason to prefer ML here?

**A)** ML eliminates the need for evaluation metrics
**B) ML scales better for complex problems with no practical rule-based solution ✓ Correct**
**C)** ML guarantees perfect predictions with little data
**D)** ML avoids all maintenance once deployed

*Rationale: Speech recognition is too complex for handcrafted rules to scale well, making ML a better fit.*

**Q69.**
A hospital's spam filter must adapt quickly to new phishing phrases appearing daily. Which approach is most operationally suitable if rapid adaptation is essential?

**A)** Weekly batch retraining only
**B) Online learning with close monitoring ✓ Correct**
**C)** Manual rule writing only
**D)** Dimensionality reduction without retraining

*Rationale: Online learning can adapt quickly to rapidly changing patterns, though it requires careful monitoring.*

**Q70.**
An ML engineer raises the learning rate of an online model so it adapts rapidly to new insurance claim patterns. What is the main risk highlighted in the chapter?

**A)** Sampling bias
**B) Catastrophic forgetting ✓ Correct**
**C)** Validation leakage
**D)** Underfitting from excessive constraints

*Rationale: A high learning rate can cause the model to forget older but still important patterns too quickly.*



Hands - On Machine Learning with Scikit - Learn and PyTorch 1st edition Geron Test Bank

**Q71.**
A remote patient monitoring system updates continuously from live sensor feeds. A device bug begins sending corrupted values, and model performance starts dropping. According to the chapter, the best immediate operational response is to:

**A)** Increase the learning rate so the model adapts faster
**B)** Add more random features to absorb the noise
**C) Monitor closely and switch learning off, possibly reverting to a previous state ✓ Correct**
**D)** Move the validation set into training to stabilize the model

*Rationale: For online learning systems, bad incoming data can rapidly degrade performance, so learning should be halted and the system reverted if needed.*

**Q72.**
A health services researcher has a small, frequently changing dataset and needs a simple method that compares new records with similar prior cases. Which approach may shine in this setting?

**A) Instance-based learning ✓ Correct**
**B)** High-degree polynomial regression
**C)** Batch-only ensemble learning
**D)** Association rule mining

*Rationale: Instance-based learning often works well on small datasets, especially when the data changes over time.*

**Q73.**
A genomics lab needs to estimate disease risk from very long DNA sequences by learning spread-out patterns across the sequence. Which model family is specifically highlighted in the chapter for this kind of task?

**A)** k-means
**B) State space models ✓ Correct**
**C)** Isolation forests
**D)** Linear regression

*Rationale: State space models are noted as particularly strong for spread-out patterns in very long sequences.*

**Q74.**
A patient experience team wants to visualize a high-dimensional survey dataset in two dimensions to inspect possible patterns. Which approach best fits this goal?

**A)** Regression
**B) Dimensionality reduction or visualization ✓ Correct**
**C)** Classification
**D)** Reinforcement learning

*Rationale: Visualization and dimensionality reduction methods are used to represent complex high-dimensional data in 2D or 3D.*



Hands - On Machine Learning with Scikit - Learn and PyTorch 1st edition Geron Test Bank

**Q75.**

A machine learning specialist merges strongly correlated variables such as patient age and cumulative wear-related utilization into a single derived feature. This is an example of:

**A)** Regularization

**B) Feature extraction ✓ Correct**

**C)** Novelty detection

**D)** Holdout validation

*Rationale: Combining correlated features into a more useful one is feature extraction.*

**Q76.**

A hospital develops a model to classify MRI images as tumor or no tumor using scans that have expert-provided labels. Which category best describes the training setup?

**A)** Unsupervised learning

**B) Supervised learning ✓ Correct**

**C)** Reinforcement learning

**D)** Association learning

*Rationale: Expert-provided labels make this a supervised learning problem.*

**Q77.**

A quality improvement team notices that 5% of patient records are missing age. According to the chapter, which is a reasonable data-preparation decision?

**A)** Keep missing values unchanged because ML always handles them automatically

**B) Choose among dropping the feature, dropping records, imputing values, or building alternate models ✓ Correct**

**C)** Replace all missing values with zero regardless of context

**D)** Use the test set to estimate the missing ages

*Rationale: The chapter lists several valid strategies for missing features, including dropping, imputing, or training separate models.*

**Q78.**

A healthcare startup trains a very complex model on a noisy, relatively small dataset and discovers odd patterns that do not hold in new patients. Which explanation best fits?

**A)** The model has learned genuine causal mechanisms

**B)** The model is underfitting due to too much data

**C) The model is overfitting by learning noise in the training data ✓ Correct**

**D)** The model is performing dimensionality reduction

*Rationale: Complex models on small or noisy datasets often learn spurious patterns that fail to generalize.*

**Q79.**

A data scientist reduces the slope flexibility of a regression model by imposing a penalty that discourages large parameter values. What is the main purpose of this step?

**A)** To increase overfitting so training error falls

**B) To simplify the model and reduce overfitting risk ✓ Correct**

**C)** To convert regression into clustering

**D)** To replace labels with generated targets

*Rationale: Regularization constrains model complexity to improve generalization and reduce overfitting.*

**Q80.**

A hospital uses 80% of its dataset for training and 20% for testing. What is the primary purpose of the test set?

**A)** To tune hyperparameters repeatedly

**B) To estimate generalization error on unseen cases ✓ Correct**

**C)** To increase the number of training examples

**D)** To replace the need for a validation set in all cases

*Rationale: The test set provides an estimate of how well the model will perform on new, unseen data.*

**Q81.**

A pharmacy analytics team compares a linear model with a polynomial model and also tries several regularization strengths. Which dataset should be used to choose among these candidate models before final testing?

**A)** The test set

**B) The validation set ✓ Correct**

**C)** Only the full training set

**D)** The production stream only

*Rationale: The validation set is used for model selection and hyperparameter tuning before final evaluation on the test set.*

**Q82.**

A team's validation set is very small, and model rankings vary wildly from run to run. Which method from the chapter can provide a more reliable estimate of model performance, albeit at higher computational cost?

**A) Repeated cross-validation ✓ Correct**

**B)** Increasing the test set until it includes all data

**C)** Switching to online learning

**D)** Using only one hyperparameter value

*Rationale: Repeated cross-validation averages performance across multiple validation splits, improving estimate stability.*

**Q83.**

A clinical NLP group trains on abundant web text but evaluates poorly on clinician notes. The model performs well on a held-out subset of the web training source. What does this pattern most strongly suggest?

**A)** Underfitting to the web training set

**B) Data mismatch between training source and production-like data ✓ Correct**

**C)** The test set is too large

**D)** The model is instance-based rather than model-based

*Rationale: Good performance on train-dev but poor performance on dev indicates mismatch between training data and target production data.*

**Q84.**

A hospital operations analyst asks why a larger training sample did not fix poor model performance. Which answer best reflects the chapter's warning?

**A)** Large samples always remove bias completely

**B) A large sample can still be nonrepresentative if the sampling method is flawed ✓ Correct**

**C)** Bias only occurs in supervised learning

**D)** Sampling bias matters only for test data, not training data

*Rationale: Even very large datasets can be biased if the sampling process is flawed.*

**Q85.**

A research team trains a disease-risk model mostly on data from tertiary-care hospitals and then applies it to a rural outpatient population with different characteristics. Which challenge is most directly implicated?

**A) Sampling bias or nonrepresentative training data ✓ Correct**

**B)** Catastrophic forgetting

**C)** Association rule failure

**D)** Label generation from self-supervision

*Rationale: Training data that does not represent the target population can lead to poor generalization due to sampling bias.*

**Q86.**

A payer wants to predict whether a claim is fraudulent and receives output as a probability of fraud. Which statement is most accurate?

**A)** This must be regression because the output is numeric

**B) This is classification even if the model outputs class probabilities ✓ Correct**

**C)** This is unsupervised because fraud labels are noisy

**D)** This is dimensionality reduction because probabilities compress information

*Rationale: Classification models such as logistic regression can output probabilities while still solving a classification task.*



Hands - On Machine Learning with Scikit - Learn and PyTorch 1st edition Geron Test Bank

**Q87.**

A machine learning consultant says, 'We should inspect the trained model to discover predictors clinicians may not have considered.' Which benefit of ML is being emphasized?

**A) Data mining for human insight ✓ Correct**

**B)** Guaranteed causal inference

**C)** Elimination of deployment constraints

**D)** Replacement of validation procedures

*Rationale: ML can reveal hidden patterns and correlations in data, supporting data mining and human learning.*

**Q88.**

A chatbot development team pretrains a large language model by masking words in text and predicting the missing words. According to the chapter, this is best categorized as:

**A)** Association rule learning

**B) Self-supervised learning ✓ Correct**

**C)** Reinforcement learning

**D)** Novelty detection

*Rationale: Masking parts of text and predicting them creates labels from unlabeled text, which is self-supervised learning.*

**Q89.**

A hospital wants a system that can keep learning from incoming email examples after deployment. Which term best describes the opposite setup, in which the model is trained and then used without further learning?

**A)** Online learning

**B) Offline learning ✓ Correct**

**C)** Semi-supervised learning

**D)** Federated learning

*Rationale: Offline learning refers to a model that is trained and then deployed without continuing to learn in production.*

**Q90.**

A health app stores the entire training set on-device so it can compare each new user to past users. What is a major operational disadvantage of this approach mentioned in the chapter?

**A)** It cannot use any similarity measure

**B) It scales poorly because prediction may require searching many stored instances ✓ Correct**

**C)** It always underfits due to low parameter count

**D)** It cannot handle changing data at all

*Rationale: Instance-based methods often scale poorly because they must store and search through training instances at prediction time.*

**Q91.**

A data scientist says the model parameter values will be learned from the data, but the regularization strength must be chosen before training. Which distinction is being described?

**A)** Training set versus test set

**B)** Label versus feature

**C) Model parameter versus hyperparameter ✓ Correct**

**D)** Batch learning versus online learning

*Rationale: Model parameters are learned during training, whereas hyperparameters are set in advance and control the learning process.*

**Q92.**

A claims model uses diagnosis codes, service dates, and cost fields, but also includes a random identifier with no predictive value. According to the chapter, why is this problematic?

**A) Irrelevant features can impair learning and contribute to poor performance ✓ Correct**

**B)** Random identifiers automatically regularize the model

**C)** Identifiers convert supervised learning into unsupervised learning

**D)** Irrelevant features matter only in tiny datasets

*Rationale: Too many irrelevant features can hinder learning, reflecting the garbage-in, garbage-out principle.*

**Q93.**

A health economist assumes a straight-line relationship between income and well-being before training a regression model. In the chapter's workflow, this step is called:

**A)** Feature extraction

**B) Model selection ✓ Correct**

**C)** Cross-validation

**D)** Novelty detection

*Rationale: Choosing a linear relationship is selecting the model type and architecture, which is model selection.*

**Q94.**

A machine learning lead explains that after selecting a model, the algorithm searches for parameter values that minimize prediction error on the training data. This process is called:

**A)** Inference

**B) Training ✓ Correct**

**C)** Deployment

**D)** Sampling

*Rationale: Training is the process of fitting model parameters to the training data by optimizing a cost or utility function.*



Hands - On Machine Learning with Scikit - Learn and PyTorch 1st edition Geron Test Bank

**Q95.**

A hospital wants to estimate how well its readmission model will perform before exposing patients and clinicians to it in production. Which approach is best aligned with the chapter?

**A)** Deploy immediately and rely on complaints to guide revisions

**B) Split data into training and test sets and estimate generalization before deployment ✓ Correct**

**C)** Use the full dataset for training and report training accuracy only

**D)** Tune on the test set until performance looks stable

*Rationale: Evaluating on a held-out test set provides a safer estimate of generalization than immediate production deployment.*

**Q96.**

A healthcare organization is deciding between several algorithm families and asks which one is guaranteed to be best for all datasets. According to the No Free Lunch theorem, the best answer is:

**A)** Neural networks are always best for complex healthcare data

**B)** Linear models are always best when interpretability matters

**C) No single model is guaranteed best without assumptions about the data ✓ Correct**

**D)** The newest algorithm is usually the safest choice

*Rationale: The No Free Lunch theorem states that no model is universally best across all possible datasets.*

**Q97.**

A model for classifying pet photos is first trained to repair masked pet images and later adapted to predict species labels. What is the main reason this pretraining can help?

**A)** It guarantees zero need for labeled data in the final task

**B) It helps the model learn useful visual representations before fine-tuning ✓ Correct**

**C)** It converts the task into reinforcement learning

**D)** It eliminates the need for a test set

*Rationale: Self-supervised pretraining helps the model learn features and representations that transfer to the target task.*

**Q98.**

A hospital's batch-trained sepsis prediction model gradually loses accuracy over time because practice patterns and data collection processes change. Which concept best describes this phenomenon?

**A)** Catastrophic forgetting

**B) Data drift ✓ Correct**

**C)** Association bias

**D)** Feature extraction

*Rationale: Data drift refers to performance decay over time as the world changes while the model stays fixed.*

**Q99.**

A healthcare analytics team must choose between retraining a huge model nightly from scratch or using incremental updates. Which factor most strongly favors incremental learning according to the chapter?

**A)** The need to avoid all monitoring

**B)** The inability to define labels

**C) Rapidly changing data or limited computational resources ✓ Correct**

**D)** A desire to maximize training time

*Rationale: Online or incremental learning is favored when data changes quickly or compute and memory resources are constrained.*

**Q100.**

A hospital develops a model that predicts well on training data and reasonably on validation data, then retrains the chosen configuration on the full training set including the validation portion before final evaluation. What should happen next?

**A)** Tune hyperparameters again on the test set

**B) Evaluate once on the test set to estimate generalization ✓ Correct**

**C)** Discard the test set because the model is already finalized

**D)** Replace the model with a nearest-neighbors baseline automatically

*Rationale: After model selection and retraining on the full training data, the final step is a one-time evaluation on the test set.*

**Q101.**

A hospital analytics team has millions of unlabeled chest images but only a small labeled subset for disease classification. Which strategy best uses the chapter's concepts to maximize downstream performance while minimizing labeling burden?

**A) Train a self-supervised model to reconstruct masked image regions, then fine-tune it on the labeled subset ✓ Correct**

**B)** Use only k-nearest neighbors on the labeled subset because unlabeled data cannot contribute to supervised tasks

**C)** Discard the unlabeled images and train a fully supervised model from scratch on the labeled subset

**D)** Apply association rule learning to discover labels and deploy the resulting rules as the final classifier

*Rationale: Self-supervised pretraining can leverage large unlabeled datasets to learn useful representations that are later fine-tuned with limited labeled data.*

**Q102.**

A fraud detection system must adapt within minutes to new attack patterns, but leadership is concerned that recent malicious inputs could rapidly degrade model behavior. Which recommendation best balances responsiveness and risk?

**A)** Use batch learning retrained annually because infrequent updates eliminate data quality concerns

**B) Use online learning with close monitoring, anomaly detection on inputs, and the ability to disable learning or revert models ✓ Correct**

**C)** Use only instance-based learning because similarity methods are immune to poisoned data

**D)** Use a very high learning rate so the system forgets suspicious historical patterns quickly

*Rationale: The chapter notes that online learning is suitable for rapid adaptation but requires monitoring, anomaly detection, and rollback safeguards against bad data.*

**Q103.**

A data scientist compares 120 hyperparameter settings for a readmission model and chooses the one with the lowest error on the test set. After deployment, performance is much worse than expected. What is the most defensible explanation?

**A)** The model underfit because the test set was too representative of production data

**B) The model overfit the test set through repeated evaluation during model selection ✓ Correct**

**C)** The model failed because test sets should contain only unlabeled data

**D)** The model was invalid because hyperparameters must be learned during gradient descent

*Rationale: Repeatedly tuning hyperparameters against the test set leaks information and makes the chosen model overly specialized to that test set.*

**Q104.**

A public health researcher trains a life-satisfaction model using mostly middle-income countries and then applies it to very poor and very rich countries. The predictions are unreliable even though the sample size is moderate. Which issue is most central?

**A)** Catastrophic forgetting

**B) Sampling bias causing nonrepresentative training data ✓ Correct**

**C)** Regularization error

**D)** Feature extraction collapse

*Rationale: The key problem is that the training data does not represent the target population to which the model is being generalized.*

**Q105.**

Which scenario is the strongest example of machine learning helping humans learn through data mining rather than merely automating prediction?

**A)** A spam filter blocks junk mail more accurately after retraining

**B)** A tumor segmentation model labels each pixel of an MRI scan

**C) Inspection of a trained model reveals previously unnoticed word combinations associated with phishing campaigns ✓ Correct**

**D)** A speech recognizer converts dictated notes into text in real time

*Rationale: The chapter highlights that inspecting learned patterns can reveal hidden correlations and trends, which is a data mining benefit.*

**Q106.**

A mobile health app must update a patient-specific model directly on the device because connectivity is intermittent and raw data should remain local. Which learning setup is most aligned with the chapter's discussion?

**A)** Batch learning from scratch every night on the device using the full historical dataset

**B) Online learning because incremental updates are resource-efficient and suitable when systems must learn autonomously with limited resources ✓ Correct**

**C)** Only unsupervised learning because supervised learning cannot run on constrained devices

**D)** Only model-based offline learning because online learning requires constant network access

*Rationale: The chapter identifies online learning as appropriate when a system must learn incrementally under limited computational resources.*

**Q107.**

A team builds a sepsis model using abundant historical ICU data from one health system, but production will occur in a rural network with different devices and workflows. Validation and test performance must estimate real-world use. What split strategy is most appropriate?

**A)** Use random splits only from the historical ICU dataset because larger samples always dominate representativeness concerns

**B) Reserve representative rural-network data for validation and test, and optionally use a train-dev set from the historical source to diagnose overfitting versus mismatch ✓ Correct**

**C)** Use the rural data only for final test and tune on the historical source to avoid wasting scarce representative data

**D)** Merge all sources and use a single test set because mismatch disappears when the sample is larger

*Rationale: When training and production data differ, validation and test sets should be representative of production, and a train-dev set can help isolate mismatch from overfitting.*

**Q108.**

An executive asks whether a model with the lowest training error should always be preferred. Which response best reflects the chapter's perspective?

**A)** Yes, because training error is the most direct measure of learning success

**B)** Yes, unless the model uses unsupervised pretraining

**C) No, because the goal is low generalization error on unseen data, not merely low error on the training set ✓ Correct**

**D)** No, because training error is irrelevant once the model is deployed

*Rationale: The chapter emphasizes that success is determined by generalization to new instances rather than fit to the training set alone.*

**Q109.**

A clinical NLP team is deciding between a linear classifier and a deep neural network for note classification. According to the No Free Lunch theorem, what is the most defensible decision principle?

**A)** Always choose the simpler model because all complex models overfit

**B)** Always choose the neural network because text tasks are inherently nonlinear

**C) Evaluate a few reasonable models based on assumptions about the data because no model is universally best a priori ✓ Correct**

**D)** Avoid model comparison because theorem implies all models perform equally

*Rationale: The No Free Lunch theorem implies there is no universally best model without assumptions, so reasonable candidates must be evaluated empirically.*

**Q110.**

A quality-improvement team has a model that performs poorly on both the training set and the test set. Which interpretation is most likely?

**A)** The model is overfitting due to excessive complexity

**B) The model is underfitting because it is too simple or too constrained to capture structure in the data ✓ Correct**

**C)** The validation set is leaking into training

**D)** The test set is too large relative to the training set

*Rationale: Poor performance even on the training data suggests the model cannot capture the underlying patterns, which is underfitting.*

**Q111.**

A hospital wants to cluster patients into utilization groups before designing care-management pathways. Why would supervised learning be a poor first choice for this specific objective?

**A) Because clustering seeks hidden structure without predefined labels ✓ Correct**

**B)** Because supervised learning cannot process tabular utilization data

**C)** Because classification models always require online learning

**D)** Because clustering is a regression problem in disguise

*Rationale: Clustering is an unsupervised task used when the goal is to discover groups without labeled outcomes.*

**Q112.**

An online recommendation model is missing long-term user preferences after several rapid updates to recent clicks. Which chapter concept most directly explains this failure mode?

**A)** Sampling bias

**B) Catastrophic forgetting caused by overly rapid adaptation ✓ Correct**

**C)** Feature extraction

**D)** Underfitting due to low variance

*Rationale: The chapter describes catastrophic forgetting as a risk in online learning when models adapt too quickly and forget older patterns.*

**Q113.**

A researcher says, 'We downloaded a massive biomedical corpus, so our computer has learned medicine.' Based on the chapter, what is the best rebuttal?

**A)** True, because acquiring data alone constitutes experience in machine learning

**B) False, because learning requires improved performance on a task as measured by a performance metric, not mere storage of data ✓ Correct**

**C)** True, because unlabeled data automatically creates labels through self-supervision

**D)** False, because machine learning applies only to numerical data

*Rationale: The chapter distinguishes possessing data from learning, which requires improved task performance from experience.*

**Q114.**

A team uses k-nearest neighbors for a nationwide radiology image retrieval system and notices unacceptable latency and memory demands in production. Which explanation best fits the chapter?

**A) Instance-based methods can be slow at prediction time and require storing training instances, especially at scale ✓ Correct**

**B)** k-nearest neighbors is inherently an offline-only algorithm and cannot be deployed

**C)** Image data is too low-dimensional for similarity search to work well

**D)** Model-based methods always require more memory than instance-based methods

*Rationale: The chapter notes that instance-based learning scales poorly because it must retain examples and search among them when predicting.*

**Q115.**

Which option best distinguishes a model parameter from a hyperparameter in the chapter's framework?

**A)** A model parameter is chosen before training, whereas a hyperparameter is learned from data during training

**B)** A model parameter governs data splitting, whereas a hyperparameter governs label quality

**C) A model parameter is fit by the learning algorithm, whereas a hyperparameter is set in advance to control the learning process ✓ Correct**

**D)** A model parameter applies only to neural networks, whereas a hyperparameter applies only to linear models

*Rationale: Model parameters are learned from training data, while hyperparameters are preset choices that shape learning.*

**Q116.**

A machine learning engineer proposes using a highly regularized linear model for a complex genomics prediction task because it will avoid overfitting. What is the strongest critique?

**A) Avoiding overfitting alone is insufficient because too much regularization may cause underfitting and miss important structure ✓ Correct**

**B)** Regularization can only be applied to unsupervised models

**C)** Linear models are invalid for any biomedical data

**D)** Overfitting is preferable to underfitting when the dataset is large

*Rationale: The chapter stresses balancing simplicity and fit, since excessive regularization can make a model too simple to learn the signal.*

**Q117.**

A team reports excellent train-dev performance but poor dev-set performance when the dev set is drawn from the actual deployment environment. What problem is most likely?

**A)** The model is underfit on the training set

**B)** The model is overfit to the dev set

**C) There is a mismatch between training-source data and deployment data ✓ Correct**

**D)** The test set has contaminated the training set

*Rationale: Good train-dev but poor dev performance indicates the model generalizes within the training source but not to the target environment, suggesting data mismatch.*

**Q118.**

A department chair asks why machine learning may outperform hand-coded rules for speech recognition. Which answer best aligns with the chapter?

**A)** Because speech recognition is a simple rule-based task once pitch is measured

**B) Because ML can discover complex patterns in large, variable, noisy data where no practical explicit rule set is known ✓ Correct**

**C)** Because rule-based systems cannot process audio files

**D)** Because supervised learning is unnecessary for language tasks

*Rationale: The chapter uses speech recognition as an example of a problem too complex for traditional rule-based programming.*

**Q119.**

A team trains a model to classify pathology slides and then inspects the learned features to identify visual patterns associated with rare lesions. This use case primarily illustrates which added value of ML?

**A)** Transfer learning

**B) Data mining for insight generation ✓ Correct**

**C)** Offline learning

**D)** Association rule learning only

*Rationale: Beyond prediction, the chapter highlights that ML can reveal hidden patterns and improve human understanding of complex data.*

**Q120.**

A startup wants one universal algorithm that is guaranteed to be best for every future healthcare dataset. Which response is most consistent with the chapter?

**A)** A universal best algorithm exists if enough data are collected

**B)** A universal best algorithm exists for all supervised tasks but not unsupervised tasks

**C) No such guarantee exists; model choice depends on assumptions about the data and empirical evaluation ✓ Correct**

**D)** Neural networks are the universal best algorithm under the No Free Lunch theorem

*Rationale: The No Free Lunch theorem implies no algorithm is universally superior across all possible datasets without assumptions.*

**Q121.**

Which statement most accurately characterizes self-supervised learning as presented in the chapter?

**A)** It is identical to clustering because both use unlabeled data only

**B) It generates labels from unlabeled data and often serves as pretraining for later fine-tuning on a target task ✓ Correct**

**C)** It is a form of reinforcement learning with synthetic rewards

**D)** It can be used only for image restoration, not classification or regression

*Rationale: The chapter describes self-supervised learning as generating labels from unlabeled data, often for pretraining before fine-tuning.*

**Q122.**

A data scientist removes a strongly correlated pair of variables by combining them into a single 'wear-and-tear' score before training. This is best described as:

**A) feature extraction within dimensionality reduction ✓ Correct**

**B)** association rule learning

**C)** nonresponse bias correction

**D)** catastrophic interference

*Rationale: Combining correlated features into a more useful one is feature extraction, a common dimensionality reduction approach.*

**Q123.**

A hospital uses a model trained monthly from scratch on all available claims data, and the production model does not update between retraining cycles. How should this system be classified?

**A)** Online and instance-based

**B) Batch and offline during production use ✓ Correct**

**C)** Reinforcement-based and online

**D)** Semi-supervised and incremental

*Rationale: Training from scratch on the full dataset and then deploying a fixed model is batch learning with offline operation in production.*

**Q124.**

A team wants to detect unusual infusion pump telemetry patterns after training mostly on normal operating data. Which framing is most appropriate?

**A)** Regression because the output is a continuous anomaly score

**B) Unsupervised anomaly detection because the model learns normal patterns and flags deviations ✓ Correct**

**C)** Supervised clustering because telemetry groups are latent classes

**D)** Reinforcement learning because the pump receives penalties for errors

*Rationale: The chapter defines anomaly detection as learning from mostly normal instances and identifying unusual new cases.*

**Q125.**

A classifier's validation performance improves after removing obviously erroneous records and extreme measurement glitches from the training set. Which chapter principle does this result support most directly?

**A) Poor-quality data can obscure underlying patterns, so cleaning data may improve learning ✓ Correct**

**B)** Outliers always increase generalization error regardless of model type

**C)** Training data should never be altered once collected

**D)** Data cleaning mainly reduces sampling bias, not noise

*Rationale: The chapter emphasizes that errors, outliers, and noise can hinder learning and that cleaning data is often worthwhile.*

**Q126.**

A researcher claims that because logistic regression is called 'regression,' it cannot be used for classification. What is the best response?

**A)** Correct; regression models output only continuous targets

**B)** Correct; classification requires unsupervised labels

**C) Incorrect; some regression-named models, such as logistic regression, are commonly used for classification by outputting class probabilities ✓ Correct**

**D)** Incorrect; logistic regression is actually a clustering algorithm

*Rationale: The chapter explicitly notes that logistic regression is commonly used for classification because it can estimate class probabilities.*



Hands - On Machine Learning with Scikit - Learn and PyTorch 1st edition Geron Test Bank

**Q127.**

A robotics lab wants an agent to learn how to traverse unknown terrain by trying actions and receiving rewards over time. Which approach best matches the task?

**A)** Semi-supervised learning
**B) Reinforcement learning ✓ Correct**
**C)** Association rule learning
**D)** Dimensionality reduction

*Rationale: Reinforcement learning is designed for agents that learn policies through actions and rewards in an environment.*

**Q128.**

An epidemiology team has 10 million records and proposes an 80/20 train-test split by default. Which critique most closely follows the chapter?

**A)** A 20% test set is always too small for large datasets
**B) A fixed 80/20 rule is unnecessary; with very large datasets, a much smaller test proportion may still estimate generalization well ✓ Correct**
**C)** Large datasets eliminate the need for a test set entirely
**D)** The test set should always be larger than the training set for stable evaluation

*Rationale: The chapter notes that test-set proportion depends on dataset size, and very large datasets may need only a small holdout.*

**Q129.**

A healthcare chatbot is pretrained by masking words in a massive unlabeled corpus and predicting the missing words, then fine-tuned for triage support. This pipeline primarily exemplifies:

**A)** association rule learning followed by clustering
**B) self-supervised pretraining with transfer learning ✓ Correct**
**C)** instance-based learning with online updates
**D)** reinforcement learning with offline rewards

*Rationale: Masking words for prediction is a self-supervised objective, and reusing the resulting model for a new task is transfer learning.*

**Q130.**

A team evaluating candidate models uses repeated cross-validation instead of a single validation split. What is the main trade-off described in the chapter?

**A) Lower bias in performance estimates at the cost of more training time ✓ Correct**
**B)** Higher risk of test leakage at the cost of less memory use
**C)** Better online adaptation at the cost of more catastrophic forgetting
**D)** Less sampling bias at the cost of more label noise

*Rationale: Repeated cross-validation can provide a more accurate performance estimate, but it multiplies training time.*



Hands - On Machine Learning with Scikit - Learn and PyTorch 1st edition Geron Test Bank

**Q131.**

A model predicts inpatient mortality accurately on training data after using patient ID as a feature, but fails on new admissions. Which explanation best fits the chapter's warning about features?

**A)** Patient ID is a useful extracted feature that improves generalization

**B) The model has likely exploited an irrelevant feature, contributing to overfitting ✓ Correct**

**C)** The issue is necessarily due to insufficient training quantity

**D)** The problem is that supervised learning should not use identifiers

*Rationale: Irrelevant features can allow complex models to learn spurious patterns that do not generalize to new cases.*

**Q132.**

Which recommendation best addresses a model that underfits a complex clinical outcome dataset?

**A)** Increase regularization strength and remove informative variables

**B) Use a more powerful model, improve features, or reduce constraints ✓ Correct**

**C)** Evaluate only on the training set until error is minimized

**D)** Replace labels with cluster assignments

*Rationale: The chapter lists more expressive models, better features, and reduced regularization as common remedies for underfitting.*

**Q133.**

A hospital wants to identify combinations of medications and supply purchases that tend to occur together in procurement data. Which ML task is the best fit?

**A)** Semantic segmentation

**B) Association rule learning ✓ Correct**

**C)** Regression

**D)** Reinforcement learning

*Rationale: Association rule learning is used to discover interesting relationships among attributes in large datasets.*

**Q134.**

A team says, 'Our validation set is tiny, but that is ideal because it leaves nearly all data for training.' What is the best evaluation of this claim?

**A)** Correct, because smaller validation sets always improve model selection

**B) Incorrect, because a very small validation set can make model evaluations imprecise and lead to selecting a suboptimal model ✓ Correct**

**C)** Correct, because validation precision does not matter if test data are plentiful

**D)** Incorrect, because validation sets must always be larger than training sets

*Rationale: The chapter warns that too small a validation set yields noisy estimates and can mislead model selection.*

**Q135.**

An analytics lead asks why a batch-trained model for classifying pet photos may still need periodic retraining despite the underlying categories remaining stable. Which answer is best?

**A)** Because cats and dogs biologically evolve too quickly for static models

**B) Because input distributions can drift due to changes in cameras, formats, image quality, and user behavior ✓ Correct**

**C)** Because batch models must always retrain daily regardless of task

**D)** Because offline models cannot make predictions without retraining

*Rationale: The chapter explains that even stable concepts may require retraining because the data distribution and acquisition process change over time.*

**Q136.**

A data scientist uses abundant unlabeled family photos, clusters faces, then asks users to provide one or a few names per person before training a recognizer. This most directly illustrates:

**A) semi-supervised learning combining unsupervised grouping with limited labels ✓ Correct**

**B)** reinforcement learning with user rewards

**C)** batch learning with no labels

**D)** association rule learning over image metadata

*Rationale: The chapter uses photo services as an example of semi-supervised learning that combines clustering with a small amount of labeling.*

**Q137.**

A policy analyst trains a linear model of health expenditure and sees low training error and low test error, but stakeholders argue a neural network must be better because it is more sophisticated. Which response is most justified?

**A)** Use the neural network because more complex models are always preferable when they can fit more patterns

**B) Keep the linear model if it generalizes well, since model choice should be driven by performance on unseen data rather than complexity alone ✓ Correct**

**C)** Reject both models because only instance-based methods are interpretable

**D)** Prefer the neural network because the No Free Lunch theorem requires the most flexible model

*Rationale: The chapter emphasizes choosing models based on generalization performance, not on complexity or sophistication by itself.*

**Q138.**

A team plans to use similarity between new patient cases and stored prior cases to make predictions. Which category does this approach fall under?

**A)** Model-based learning

**B) Instance-based learning ✓ Correct**

**C)** Reinforcement learning

**D)** Self-supervised learning

*Rationale: Instance-based learning generalizes by comparing new cases to memorized examples using a similarity measure.*

**Q139.**

A group trains a model on a reduced training set, picks the best hyperparameters using a validation set, retrains the chosen model on the full training set including the validation data, and then evaluates once on the test set. Why is this workflow preferred?

**A)** It allows the test set to directly optimize hyperparameters

**B)** It avoids using any labels during model selection

**C) It separates model selection from final generalization estimation, reducing test-set overfitting ✓ Correct**

**D)** It guarantees the selected model is globally optimal

*Rationale: Holdout validation preserves the test set for unbiased final evaluation after model and hyperparameters are chosen.*

**Q140.**

A model trained on a small sample of survey respondents performs poorly because respondents were disproportionately affluent and politically engaged. Which type of bias is most specifically illustrated by low response participation among those contacted?

**A)** Feature extraction bias

**B) Nonresponse bias ✓ Correct**

**C)** Regularization bias

**D)** Inference bias

*Rationale: The chapter identifies low response rates causing systematic differences between respondents and nonrespondents as nonresponse bias.*

**Q141.**

A machine learning consultant says, 'We should spend more effort collecting quality labeled examples before inventing a new algorithm.' Which chapter discussion most strongly supports this position?

**A)** The section on reinforcement learning rewards

**B) The unreasonable effectiveness of data ✓ Correct**

**C)** The warning about catastrophic forgetting

**D)** The note on semantic segmentation

*Rationale: The chapter cites evidence that with enough data, different algorithms can perform similarly well on complex problems.*

**Q142.**

A hospital operations team wants a 2D map of thousands of high-dimensional patient records to inspect possible subgroups visually before designing interventions. Which objective best matches this need?

**A)** Regression

**B) Visualization via dimensionality reduction ✓ Correct**

**C)** Supervised classification

**D)** Online learning

*Rationale: The chapter describes visualization algorithms and dimensionality reduction as tools for projecting complex unlabeled data into 2D or 3D for insight.*

**Q143.**

A team chooses an extremely high learning rate for an online model because it wants immediate adaptation to new billing patterns. What is the most important downside identified in the chapter?

**A)** The model will necessarily become batch-only

**B) The model may rapidly forget older useful patterns and become overly sensitive to recent noise ✓ Correct**

**C)** The model will stop using labels during training

**D)** The model cannot be monitored once deployed

*Rationale: A high learning rate increases responsiveness but also raises the risk of catastrophic forgetting and sensitivity to nonrepresentative recent data.*

**Q144.**

A model with many parameters fits a noisy training dataset almost perfectly, including accidental correlations such as country names containing a certain letter. Which intervention is most directly aimed at reducing this specific problem?

**A)** Increase model complexity further to capture all noise more precisely

**B) Apply regularization or simplify the model to constrain degrees of freedom ✓ Correct**

**C)** Eliminate the test set so the model is judged only on training fit

**D)** Convert the problem from supervised to unsupervised learning

*Rationale: Overfitting from excessive complexity is commonly addressed by simplifying the model or adding regularization constraints.*

**Q145.**

A team evaluating a new diagnostic model says that because it achieved 98% accuracy on the training set, it has proven the system learned effectively. Which response is best?

**A)** Correct, because training accuracy is the standard performance measure for all tasks

**B) Partly correct, but learning must be judged by improved performance on the task and especially by generalization to unseen data ✓ Correct**

**C)** Incorrect, because accuracy is never used in classification tasks

**D)** Incorrect, because only unsupervised models can be evaluated quantitatively

*Rationale: High training accuracy alone is insufficient; the chapter stresses performance on new cases as the key evidence of useful learning.*

**Q146.**

A biomedical startup has a dataset too large to fit into one machine’s memory and wants to train by loading chunks sequentially. Which term from the chapter best describes this setup?

**A)** Transfer learning

**B) Out-of-core learning ✓ Correct**

**C)** Association learning

**D)** Novelty detection

*Rationale: Out-of-core learning refers to incremental training on datasets too large to fit in memory by processing portions sequentially.*

**Q147.**

A team is deciding whether to remove a feature with many missing values. According to the chapter, which is the most defensible stance?

**A)** Missingness has only one valid remedy: delete all affected records

**B) The team should consider multiple options, such as dropping the feature, dropping instances, imputing values, or training alternative models ✓ Correct**

**C)** Any feature with missing values must be kept unchanged to preserve representativeness

**D)** Missing features are relevant only in unsupervised learning

*Rationale: The chapter presents several reasonable strategies for handling missing features rather than a single mandatory approach.*

**Q148.**

A health system wants a recommendation about whether to use instance-based or model-based learning for a small but frequently changing appointment no-show dataset. Which recommendation is most aligned with the chapter?

**A) Prefer instance-based methods because they can shine on small datasets and changing data, despite scaling limitations ✓ Correct**

**B)** Prefer model-based methods because instance-based methods cannot generalize at all

**C)** Avoid both because only reinforcement learning handles changing data

**D)** Prefer batch model-based methods because frequent change makes similarity measures invalid

*Rationale: The chapter notes that instance-based learning often works well on small datasets, especially when the data changes often.*

**Q149.**

A clinical team wants to know whether a poor dev-set result comes from overfitting to web-sourced training images or from mismatch with smartphone-acquired production images. Which additional dataset would most directly help answer this?

**A)** A second test set drawn from the smartphone data

**B) A train-dev set drawn from the same source as the training data but not used for training ✓ Correct**

**C)** A larger validation set made by merging training and test data

**D)** An unlabeled clustering set from a third source

*Rationale: A train-dev set from the training distribution helps determine whether errors stem from overfitting rather than source mismatch.*

**Q150.**

An MLOps committee is reviewing a highly accurate model that is difficult to maintain, too large for deployment hardware, and vulnerable to stale performance over time. What is the best synthesis of the chapter's position?

**A)** These concerns are secondary because data and model fit fully determine project success

**B) Operational constraints such as maintainability, memory, speed, scalability, security, and updating are legitimate deployment issues beyond core modeling ✓ Correct**

**C)** Deployment issues matter only for online learning systems

**D)** Such issues disappear once a model passes the test set

*Rationale: The chapter explicitly states that successful ML projects must address operational deployment constraints in addition to data and model quality.*