data.professional.self.assessment → Machine Learning Exam

courtesy: Ebazhanov - https://github.com/Ebazhanov

- https://github.com/Ebazhanov/linkedin-skill-assessments-quizzes/blob/845bc1a3107d201e7cd152f4143a18bc10d330d2/machine-learning/machine-learning-quiz.md

---

Machine Learning Study Exam

Q1. You are part of a data science team that is working for a national fast-food chain. You create a simple report that shows trend: Customers who visit the store more often and buy smaller meals spend more than customers who visit less frequently and buy larger meals. What is the most likely diagram that your team created?

- [ ] multiclass classification diagram
- [ ] linear regression and scatter plots
- [ ] pivot table
- [ ] K-means cluster diagram

Q2. You work for an organization that sells a spam filtering service to large companies. Your organization wants to transition its product to use machine learning. It currently a list Of 250,00 keywords. If a message contains more than few of these keywords, then it is identified as spam. What would be one advantage of transitioning to machine learning?

- [ ] The product would look for new patterns in spam messages.
- [ ] The product could go through the keyword list much more quickly.
- [ ] The product could have a much longer keyword list.
- [ ] The product could find spam messages using far fewer keywords.

Q3. You work for a music streaming service and want to use supervised machine learning to classify music into different genres. Your service has collected thousands of songs in each genre, and you used this as your training data. Now you pull out a small random subset of all the songs in your service. What is this subset called?

- [ ] data cluster
- [ ] Supervised set
- [ ] big data
- [ ] test data

Q4. In traditional computer programming, you input commands. What do you input with machine learning?

- [ ] patterns
- [ ] programs
- [ ] rules
- [ ] data

Q5. Your company wants to predict whether existing automotive insurance customers are more likely to buy homeowners insurance. It created a model to better predict the best customers contact about homeowners insurance, and the model had a low variance but high bias. What does that say about the data model?

- [ ] It was consistently wrong.
- [ ] It was inconsistently wrong.
- [ ] It was consistently right.
- [ ] It was equally right end wrong.

[Reference](https://medium.com/datadriveninvestor/bias-and-variance-in-machine-learning-51fdd38d1f86)

Q6. You want to identify global weather patterns that may have been affected by climate change. To do so, you want to use machine learning algorithms to find patterns that would otherwise be imperceptible to a human meteorologist. What is the place to start?

- [ ] Find labeled data of sunny days so that the machine will learn to identify bad weather.
- [ ] Use unsupervised learning have the machine look for anomalies in a massive weather database.
- [ ] Create a training set of unusual patterns and ask the machine learning algorithms to classify them.
- [ ] Create a training set of normal weather and have the machine look for similar patterns.

Q7. You work in a data science team that wants to improve the accuracy of its K-nearest neighbor result by running on top of a naive Bayes result. What is this an example of?

- [ ] regression
- [ ] boosting
- [ ] bagging
- [ ] stacking

Q8. `____` looks at the relationship between predictors and your outcome.

- [ ] Regression analysis
- [ ] K-means clustering
- [ ] Big data
- [ ] Unsupervised learning

Q9. What is an example of a commercial application for a machine learning system?

- [ ] a data entry system
- [ ] a data warehouse system
- [ ] a massive data repository

- [ ] a product recommendation system

Q10. What does this image illustrate?

![Machine Learning Q10](images/machine-learning_Q10.jpg)

- [ ] a decision tree
- [ ] reinforcement learning
- [ ] K-nearest neighbor
- [ ] a clear trendline

Q11. You work for a power company that owns hundreds of thousands of electric meters. These meters are connected to the internet and transmit energy usage data in real-time. Your supervisor asks you to direct project to use machine learning to analyze this usage data. Why are machine learning algorithms ideal in this scenario?

- [ ] The algorithms would help the meters access the internet.
- [ ] The algorithms will improve the wireless connectivity.
- [ ] The algorithms would help your organization see patterns of the data.
- [ ] By using machine learning algorithms, you are creating an IoT device.

Q12. To predict a quantity value. use `___`.

- [ ] regression
- [ ] clustering
- [ ] classification
- [ ] dimensionality reduction

Q13. Why is naive Bayes called naive?

- [ ] It naively assumes that you will have no data.
- [ ] It does not even try to create accurate predictions.
- [ ] It naively assumes that the predictors are independent from one another.
- [ ] It naively assumes that all the predictors depend on one another.

Q14. You work for an ice cream shop and created the chart below, which shows the relationship between the outside temperature and ice cream sales. What is the best description of this chart?

![Machine Learning Q14](images/machine-learning_Q14.jpg)

- [ ] It is a linear regression chart.
- [ ] It is a supervised trendline chart.
- [ ] It is a decision tree.

- [ ] It is a clustering trend chart.

Q15. How is machine learning related to artificial intelligence?

- [ ] Artificial intelligence focuses on classification, while machine learning is about clustering data.
- [ ] Machine learning is a type of artificial intelligence that relies on learning through data.
- [ ] Artificial intelligence is form of unsupervised machine learning.
- [ ] Machine learning and artificial intelligence are the same thing.

Q16. How do machine learning algorithms make more precise predictions?

- [ ] The algorithms are typically run more powerful servers.
- [ ] The algorithms are better at seeing patterns in the data.
- [ ] Machine learning servers can host larger databases.
- [ ] The algorithms can run on unstructured data.

Q17. You work for an insurance company. Which machine learning project would add the most value for the company?

- [ ] Create an artificial neural network that would host the company directory.
- [ ] Use machine learning to better predict risk.
- [ ] Create an algorithm that consolidates all of your Excel spreadsheets into one data lake.
- [ ] Use machine learning and big data to research salary requirements.

Q18. What is the missing information in this diagram?

![Machine Learning Q18](images/machine-learning_Q18.jpg)

- [ ] Training Set
- [ ] Unsupervised Data
- [ ] Supervised Learning
- [ ] Binary Classification

Q19. What is one reason not to use the same data for both your training set and your testing set?

- [ ] You will almost certainly underfit the model.
- [ ] You will pick the wrong algorithm.
- [ ] You might not have enough data for both.
- [ ] You will almost certainly overfit the model.

Q20. Your university wants to use machine learning algorithms to help sort through incoming student applications. An administrator asks if the admissions decisions might be biased against any particular group, such as women. What would be the best answer?

- [ ] Machine learning algorithms are based on math and statistics, and so by definition will be unbiased.
- [ ] There is no way to identify bias in the data.
- [ ] Machine learning algorithms are powerful enough to eliminate bias from the data.
- [ ] All human-created data is biased, and data scientists need to account for that.

Explanation: While machine learning algorithms don't have bias, the data can have them.

Q21. What is stacking?

- [ ] The predictions of one model become the inputs another.
- [ ] You use different versions of machine learning algorithms.
- [ ] You use several machine learning algorithms to boost your results.
- [ ] You stack your training set and testing set together.

Q22. You want to create a supervised machine learning system that identifies pictures of kittens on social media. To do this, you have collected more than 100,000 images of kittens. What is this collection of images called?

- [ ] training data
- [ ] linear regression
- [ ] big data
- [ ] test data

Q23. You are working on a project that involves clustering together images of different dogs. You take image and identify it as your centroid image. What type machine learning algorithm are you using?

- [ ] centroid reinforcement
- [ ] K-nearest neighbor
- [ ] binary classification
- [ ] K-means clustering

Explanation: The problem explicitly states "clustering".

Q24. Your company wants you to build an internal email text prediction model to speed up the time that employees spend writing emails. What should you do?

- [ ] Include training email data from all employees.
- [ ] Include training email data from new employees.

- [ ] Include training email data from seasoned employees.
- [ ] Include training email data from employees who write the majority of internal emails.

Q25. Your organization allows people to create online professional profiles. A key feature is the ability to create clusters of people who are professionally connected to one another. What type of machine learning method is used to create these clusters?

- [ ] unsupervised machine learning
- [ ] binary classification
- [ ] supervised machine learning
- [ ] reinforcement learning

Q26. What is this diagram a good example of?

![Machine Learning Q26](images/machine-learning_Q26.jpg)

- [ ] K-nearest neighbor
- [ ] a decision tree
- [ ] a linear regression
- [ ] a K-means cluster

Note: there are centres of clusters (C0, C1, C2).

Q27. Random forest is modified and improved version of which earlier technique?

- [ ] aggregated trees
- [ ] boosted trees
- [ ] bagged trees
- [ ] stacked trees

Q28. Self-organizing maps are specialized neural network for which type of machine learning?

- [ ] semi-supervised learning
- [ ] supervised learning
- [ ] reinforcement learning
- [ ] unsupervised learning

Q29. Which statement about K-means clustering is true?

- [ ] In K-means clustering, the initial centroids are sometimes randomly selected.
- [ ] K-means clustering is often used in supervised machine learning.
- [ ] The number of clusters are always randomly selected.
- [ ] To be accurate, you want your centroids outside of the cluster.

Q30. You created machine learning system that interacts with its environment and responds to errors and rewards. What type of machine learning system is it?

- [ ] supervised learning
- [ ] semi-supervised learning
- [ ] reinforcement learning
- [ ] unsupervised learning

Q31. Your data science team must build a binary classifier, and the number one criterion is the fastest possible scoring at deployment. It may even be deployed in real time. Which technique will produce a model that will likely be fastest for the deployment team use to new cases?

- [ ] random forest
- [ ] logistic regression
- [ ] KNN
- [ ] deep neural network

Q32. Your data science team wants to use the K-nearest neighbor classification algorithm. Someone on your team wants to use a K of 25. What are the challenges of this approach?

- [ ] Higher K values will produce noisy data.
- [ ] Higher K values lower the bias but increase the variance.
- [ ] Higher K values need a larger training set.
- [ ] Higher K values lower the variance but increase the bias.

Q33. Your machine learning system is attempting to describe a hidden structure from unlabeled data. How would you describe this machine learning method?

- [ ] supervised learning
- [ ] unsupervised learning
- [ ] reinforcement learning
- [ ] semi-unsupervised learning

Q34. You work for a large credit card processing company that wants to create targeted promotions for its customers. The data science team created a machine learning system that groups together customers who made similar purchases, and divides those customers based on customer loyalty. How would you describe this machine learning approach?

- [ ] It uses unsupervised learning to cluster together transactions and unsupervised learning to classify the customers.
- [ ] It uses only unsupervised machine learning.

- [ ] It uses supervised learning to create clusters and unsupervised learning for classification.
- [ ] It uses reinforcement learning to classify the customers.

Q35. You are using K-nearest neighbor and you have a K of 1. What are you likely to see when you train the model?

- [ ] high variance and low bias
- [ ] low bias and low variance
- [ ] low variance and high bias
- [ ] high bias and high variance

Q36. Are data model bias and variance a challenge with unsupervised learning?

- [ ] No, data model bias and variance are only a challenge with reinforcement learning.
- [ ] Yes, data model bias is a challenge when the machine creates clusters.
- [ ] Yes, data model variance trains the unsupervised machine learning algorithm.
- [ ] No, data model bias and variance involve supervised learning.

Q37. Which choice is best for binary classification?

- [ ] K-means
- [ ] Logistic regression
- [ ] Linear regression
- [ ] Principal Component Analysis (PCA)

Explanation: Logistic regression is far better than linear regression at binary classification since it biases the result toward one extreme or the other. K-means clustering can be used for classification but is not as accurate in most scenarios.
[Source:](https://www.kaggle.com/minc33/k-means-clustering-vs-logistic-regression)

Q38. With traditional programming, the programmer typically inputs commands. With machine learning, the programmer inputs

- [ ] supervised learning
- [ ] data
- [ ] unsupervised learning
- [ ] algorithms

Explanation: This one is pretty straightforward and a fundamental concept.
[Source:](https://towardsdatascience.com/machine-learning-for-beginners-d247a9420dab)

Q39. Why is it important for machine learning algorithms to have access to high-quality data?

- [ ] It will take too long for programmers to scrub poor data.
- [ ] If the data is high quality, the algorithms will be easier to develop.
- [ ] Low-quality data requires much more processing power than high-quality data.
- [ ] If the data is low quality, you will get inaccurate results.

Q40. In K-nearest neighbor, the closer you are to neighbor, the more likely you are to

- [ ] share common characteristics
- [ ] be part of the root node
- [ ] have a Euclidean connection
- [ ] be part of the same cluster

Q41. In the HBO show Silicon Valley, one of the characters creates a mobile application called Not Hot Dog. It works by having the user take a photograph of food with their mobile device. Then the app says whether the food is a hot dog. To create the app, the software developer uploaded hundreds of thousands of pictures of hot dogs. How would you describe this type of machine learning?

- [ ] Reinforcement machine learning
- [ ] unsupervised machine learning
- [ ] supervised machine learning
- [ ] semi-supervised machine learning

Q42. You work for a large pharmaceutical company whose data science team wants to use unsupervised learning machine algorithms to help discover new drugs. What is an advantage to this approach?

- [ ] You will be able to prioritize different classes of drugs, such as antibiotics.
- [ ] You can create a training set of drugs you would like to discover.
- [ ] The algorithms will cluster together drugs that have similar traits.
- [ ] Human experts can create classes of drugs to help guide discovery.

Explanation: This one is similar to an example talked about in the Stanford Machine Learning course.
[Source:](https://www.coursera.org/learn/machine-learning/home/welcome)

Q43. In 2015, Google created a machine learning system that could beat a human in the game of Go. This extremely complex game is thought to have more gameplay possibilities than there are atoms of the universe. The first version of the system won by observing hundreds of thousands of hours of human gameplay; the second version learned how to play by getting rewards while playing against itself. How would you describe this transition to different machine learning approaches?

- [ ] The system went from supervised learning to reinforcement learning.

- [ ] The system evolved from supervised learning to unsupervised learning.
- [ ] The system evolved from unsupervised learnin9 to supervised learning.
- [ ] The system evolved from reinforcement learning to unsupervised learning.

Q44. The security company you work for is thinking about adding machine learning algorithms to their computer network threat detection appliance. What is one advantage of using machine learning?

- [ ] It could better protect against undiscovered threats.
- [ ] It would very likely lower the hardware requirements.
- [ ] It would substantially shorten your development time.
- [ ] It would increase the speed of the appliance.

Q45. You work for a hospital that is tracking the community spread of a virus. The hospital created a smartwatch application that uploads body temperature data from hundreds of thousands of participants. What is the best technique to analyze the data?

- [ ] Use reinforcement learning to reward the system when a new person participates.
- [ ] Use unsupervised machine learning to cluster together people based on patterns the machine discovers.
- [ ] Use Supervised machine learning to sort people by demographic data.
- [ ] Use Supervised machine learning to classify people by body temperature.

Q46. Many of the advances in machine learning have come from improved `___`.

- [ ] statistics
- [ ] structured data
- [ ] availability
- [ ] algorithms

Q47. What is this diagram a good example of?

![Machine Learning Q45](images/machine-learning_Q45.jpg)

- [ ] unsupervised learning
- [ ] complex cluster
- [ ] multiclass classification
- [ ] k-nearest neighbour

Q48. Naive Bayes looks at each **_** predictor and creates a probability that belongs in each class.

- [ ] conditional
- [ ] multiclass

- [ ] independent
- [ ] binary

[Reference](https://towardsdatascience.com/all-about-naive-bayes-8e13cef044cf)

Q49. Someone on your data science team recommends that you use decision trees, naive Bayes and K-nearest neighbor, all at the same time, on the same training data, and then average the results. What is this an example of?

- [ ] regression analysis
- [ ] unsupervised learning
- [ ] high-variance modeling
- [ ] ensemble modeling

Q50. Your data science team wants to use machine learning to better filter out spam messages. The team has gathered a database of 100,000 messages that have been identified as spam or not spam. If you are using supervised machine learning, what would you call this data set?

- [ ] machine learning algorithm
- [ ] training set
- [ ] big data test set
- [ ] data cluster

Q51. You work for a website that enables customers see all images of themselves on the internet by uploading one self-photo. Your data model uses 5 characteristics to match people to their foto: color, eye, gender, eyeglasses and facial hair. Your customers have been complaining that get tens of thousands of photos without them. What is the problem?

- [ ] You are overfitting the model to the data
- [ ] You need a smaller training set
- [ ] You are underfitting the model to the data
- [ ] You need a larger training set

Q52. Your supervisor asks you to create a machine learning system that will help your human resources department classify jobs applicants into well-defined groups. What type of system are you more likely to recommend?

- [ ] an unsupervised machine learning system that clusters together the best candidates.
- [ ] you would not recommend a machine learning system for this type of project.
- [ ] a deep learning artificial neural network that relies on petabytes of employment data.
- [ ] a supervised machine learning system that classifies applicants into existing groups.

Q53. You and your data science team have 1 TB of example data. What do you typically do with that data?

- [ ] you use it as your training set.
- [ ] You label it big data.
- [ ] You split it into a training set and test set.
- [ ] You use it as your test set.

Q54. Your data science team is working on a machine learning product that can act as an artificial opponent in video games. The team is using a machine learning algorithm that focuses on rewards: If the machine does some things well, then it improves the quality of the outcome. How would you describe this type of machine learning algorithm?

- [ ] semi-supervised machine learning
- [ ] supervised machine learning
- [ ] unsupervised machine learning
- [ ] reinforcement learning

Q55. The model will be trained with data in one single batch is known as?

- [ ] Batch learning
- [ ] Offline learning
- [ ] Both A and B
- [ ] None of the above

Q56. Which of the following is NOT supervised learning?

- [ ] Decision Tree
- [ ] Linear Regression
- [ ] PCA
- [ ] Naive Bayesian

Q57. Suppose we would like to perform clustering on spatial data such as the geometrical locations of houses. We wish to produce clusters of many different sizes and shapes. Which of the following methods is the most appropriate?

- [ ] Decision Trees
- [ ] K-means clustering
- [ ] Density-based clustering
- [ ] Model-based clustering

Q58. The error function most suited for gradient descent using logistic regression is

- [ ] The entropy function.
- [ ] The squared error.
- [ ] The cross-entropy function.

- [ ] The number of mistakes.

Q59. Compared to the variance of the Maximum Likelihood Estimate (MLE), the variance of the Maximum A Posteriori (MAP) estimate is `___`

- [ ] Higher
- [ ] same
- [ ] Lower
- [ ] it could be any of the above

Q60. `___` refers to a model that can neither model the training data nor generalize to new data.

- [ ] good fitting
- [ ] overfitting
- [ ] underfitting
- [ ] all of the above

Q61. How would you describe this type of classification challenge?

![Machine Learning Q58](images/machine-learning_Q58.jpg)

- [ ] This is a multiclass classification challenge.
- [ ] This is a multi-binary classification challenge.
- [ ] This is a binary classification challenge.
- [ ] This is a reinforcement classification challenge.

  Explanation: Shows data being classified into more than two categories or classes. Thus, this is a multi-class classification challenge.

Q62. What does it mean to underfit your data model?

- [ ] There is too little data in your training set.
- [ ] There is too much data in your training set.
- [ ] There is not a lot of variance but there is a high bias.
- [ ] Your model has low bias but high variance.

Explanation: Underfitted data models usually have high bias and low variance. Overfitted data models have low bias and high variance.

Q63. Asian user complains that your company's facial recognition model does not properly identify their facial expressions. What should you do?

- [ ] Include Asian faces in your test data and retrain your model.

- [ ] Retrain your model with updated hyperparameter values.
- [ ] Retrain your model with smaller batch sizes.
- [ ] Include Asian faces in your training data and retrain your model.

Explanation: The answer is self-explanatory: if Asian users are the only group of people making the complaint, then the training data should have more Asian faces.

Q64. You work for a website that helps match people up for lunch dates. The website boasts that it uses more than 500 predictors to find customers the perfect date, but many customers complain that they get very few matches. What is a likely problem with your model?

- [ ] Your training set is too large.
- [ ] You are underfitting the model to the data.
- [ ] You are overfitting the model to the data.
- [ ] Your machine is creating inaccurate clusters.

Explanation: // This question is very similar to Q49 but involves a polar opposite scenario.

`that answer somewhat vague and unsettled. Small number of matchings does not necessarily implies that the model overfits, especially given 500 (!) independent variables. To me, it sounds more reasonable that the threshold (matching) criterion might be too tight, thus allowing only a small number of matching to occur. So a solution can be either softening the threshold criterion or increasing the number of candidates.`

Q65. (Mostly) whenever we see kernel visualizations online (or some other reference) we are actually seeing:

- [ ] What kernels extract
- [ ] Feature Maps
- [ ] How kernels Look

Q66. The activations for class A, B and C before softmax were 10,8 and 3. The different in softmax values for class A and class B would be :

- [ ] 76%
- [ ] 88%
- [ ] 12%
- [ ] 0.0008%

![image](images/machine-learning_Q62.png?raw=png)

Q67. The new dataset you have just scraped seems to exhibit lots of missing values. What action will help you minimizing that problem?

- [ ] Wise fill-in of controlled random values
- [ ] Replace missing values with averaging across all samples
- [ ] Remove defective samples
- [ ] Imputation

Q68. Which of the following methods can use either as an unsupervised learning or as a dimensionality reduction technique?

- [ ] SVM
- [ ] PCA
- [ ] LDA
- [ ] TSNE

Q69. What is the main motivation for using activation functions in ANN?

- [ ] Capturing complex non-linear patterns
- [ ] Transforming continuous values into "ON" (1) or "OFF" (0) values
- [ ] Help avoiding the vanishing/exploding gradient problem
- [ ] Their ability to activate each neurons individually.

Q70. Which loss function would fit best in a categorical (discrete) supervised learning?

- [ ] kullback-leibler (KL) loss
- [ ] Binary Crossentropy
- [ ] Mean Squared Error (MSE)
- [ ] Any L2 loss

Q71. What is the correct option?

![image](images/machine-learning_Q67.png?raw=png)

no.	Red	Blue	Green
1.	Validation error	Training error	Test error
2.	Training error	Test error	Validation error
3.	Optimal error	Validation error	Test error
4.	Validation error	Training error	Optimal error

- [ ] 1
- [ ] 2
- [ ] 3
- [ ] 4

Q72. You create a decision tree to show whether someone decides to go to the beach. There are three factors in this decision: rainy, overcast, and sunny. What are these three factors called?

- [ ] tree nodes
- [ ] predictors
- [ ] root nodes
- [ ] deciders

`// these nodes decide whether the someone decides to go to beach or not, for example if its rainy people will mostly refrain from going to beach`

Q73. You need to quickly label thousands of images to train a model. What should you do?

- [ ] Set up a cluster of machines to label the images
- [ ] Create a subset of the images and label them yourself
- [ ] Use naive Bayes to automatically generate labels.
- [ ] Hire people to manually label the images

Q74. The fit line and data in the figure exhibits which pattern?

![image](images/machine-learning_Q70.png?raw=png)

- [ ] low bias, high variance
- [ ] high bias, low variance
- [ ] high bias, high variance
- [ ] low bias, low variance

`// since the data is accurately classified and is neither overfitting nor underfitting the dataset`

Q75. You need to select a machine learning process to run a distributed neural network on a mobile application. Which would you choose?

- [ ] Scikit-learn
- [ ] PyTorch
- [ ] Tensowflow Lite
- [ ] Tensorflow

Q76. Which choice is the best example of labeled data?

- [ ] a spreadsheet
- [ ] 20,000 recorded voicemail messages
- [ ] 100,000 images of automobiles

- [ ] hundreds of gigabytes of audio files

Q77. In statistics, what is defined as the probability of a hypothesis test of finding an effect - if there is an effect to be found?

- [ ] confidence
- [ ] alpha
- [ ] power
- [ ] significance

Q78. You want to create a machine learning algorithm to identify food recipes on the web. To do this, you create an algorithm that looks at different conditional probabilities. So if the post includes the word **flour**, it has a slightly stronger probability of being a recipe. If it contains both **flour** and _sugar_, it even more likely a recipe. What type of algorithm are you using?

- [ ] naive Bayes classifier
- [ ] K-nearest neighbor
- [ ] multiclass classification
- [ ] decision tree

Q79. What is lazy learning?

- [ ] when the machine learning algorithms do most of the programming
- [ ] when you don't do any data scrubbing
- [ ] when the learning happens continuously
- [ ] when you run your computation in one big instance at the beginning

Q80. What is Q-learning reinforcement learning?

- [ ] supervised machine learning with rewards
- [ ] a type of unsupervised learning that relies heavily on a well-established model
- [ ] a type of reinforcement learning where accuracy degrades over time
- [ ] a type of reinforcement learning that focuses on rewards

[Reference](<https://towardsdatascience.com/a-beginners-guide-to-q-learning-c3e2a30a653c#:~:text=Q%2Dlearning%20is%20a%20model,equation(particularly%20Bellman%20equation).&text=Means%20it%20learns%20the%20value,independently%20of%20the%20agent's%20actions>) **Explanation**:Q-learning is a model-free reinforcement learning algorithm.Q-learning is a values-based learning algorithm. Value based algorithms updates the value function based on an equation(particularly Bellman equation).

Q81. The data in your model has low bias and low variance. How would you expect the data points to be grouped together on the diagram?

- [ ] They would be grouped tightly together in the predicted outcome.
- [ ] They would be grouped tightly together but far from the predicted.
- [ ] They would be scattered around the predict outcome.
- [ ] They would be scattered far away from the predicted outcome.

[Reference](https://medium.com/30-days-of-machine-learning/day-3-k-nearest-neighbors-and-bias-variance-tradeoff-75f84d515bdb)

Q82. Your machine learning system is using labeled examples to try to predict future data, compare that data to the predicted result, and then the model. What is the best description of this machine learning method?

- [ ] unsupervised learning
- [ ] semi-supervised learning
- [ ] supervised learning
- [ ] semi-reinforcement learning

[Reference](https://www.expert.ai/blog/machine-learning-definition/)

Q83. In the 1983 movie WarGames, the computer learns how to master the game of chess by playing against itself. What machine learning method was the computer using?

- [ ] binary learning
- [ ] supervised learning
- [ ] unsupervised learning
- [ ] reinforcement learning

[Reference](https://venturebeat.com/2016/03/16/heres-what-alphagos-historic-win-means-for-the-enterprise/)

Q84. You are working with your machine learning algorithm on something called class predictor probability. What algorithm are you most likely using?

- [ ] multiclass binary classification
- [ ] naïve Bayes
- [ ] unsupervised classification
- [ ] decision tree analysis

Explanation: You could use a naïve Bayes algorithm, to differentiate three classes of dog breeds — terrier, hound, and sport dogs. Each class has three predictors — hair length, height, and weight. The algorithm does something called class predictor probability.

[Reference](https://www.linkedin.com/pulse/machine-learning-algorithms-doug-rose/)

Q85. What is one of the most effective way to correct for underfitting your model to the data?

- [ ] Create training clusters
- [ ] Remove predictors
- [ ] Use reinforcement learning
- [ ] Add more predictors

Q86. Your data science team is often criticized for creating reports that are boring or too obvious. What could you do to help improve the team?

- [ ] Suggest that the team is probably underfitting the model to the data.
- [ ] Suggest that unsupervised learning will lead to more interesting results.
- [ ] Make sure that they are picking the correct machine learning algorithms.
- [ ] Encourage the team to ask more interesting questions.

Q87. What is the difference between unstructured and structured data?

- [ ] Unstructured data is always text.
- [ ] Unstructured data is much easier to store.
- [ ] Structured data has clearly defined data types.
- [ ] Structured data is much more popular.

Q88. You work for a startup that is trying to develop a software tool that will scan the internet for pictures of people using specific tools. The chief executive is very interested in using machine learning algorithms. What would you recommend as the best place to start?

- [ ] Using an unsupervised machine learning algorithm to cluster together all the photographs.
- [ ] Crate a data lake with an unsupervised machine learning algorithm.
- [ ] Use a combination of unsupervised and supervised machine learning to create machine-defined data clusters.
- [ ] Use supervised machine learning to classify photographs based on a predetermined training set.

Q89. In supervised machine learning, data scientist often have the challenge of balancing between underfitting or overfitting their data model. They often have to adjust the training set to make better predictions. What is this balance called?

- [ ] the under/over challenge
- [ ] balance between clustering classification
- [ ] bias-variance trade-off
- [ ] the multiclass training set challenge

Q90. What is conditional probability?

- [ ] the probability that doing one thing has an impact on another thing
- [ ] the probability that certain conditions are met
- [ ] the probability that, based on certain conditions, something will always be incorrect
- [ ] the probability of something being the correct answer

Q91. K-means clustering is what type of machine learning algorithm?

- [ ] reinforcement
- [ ] supervised
- [ ] unsupervised
- [ ] classification

Q92. What is ensemble modeling?

- [ ] when you create an ensemble of your training and test data set
- [ ] when you create an ensemble of different servers to run the algorithms
- [ ] when you find the one best algorithm for your ensemble
- [ ] when you use several ensembles of machine learning algorithms

Q93. What is the best definition for bias in your data model?

- [ ] Bias is when your predicted values are scattered.
- [ ] Bias is the gap between your predicted value and the outcome.
- [ ] Bias is when your data is wrong for different reasons.
- [ ] Bias is when your values are always off by the same percentage.

Q94. Which project might be best suited for supervised machine learning?

- [ ] data scrubbing
- [ ] predicting a risk score
- [ ] tax filing software
- [ ] spreadsheet consolidation

Q95. When is a decision tree most commonly used?

- [ ] with big data products
- [ ] for supervised machine learning binary classification challenges
- [ ] to find thd best data cluster
- [ ] to determine "Q" in Q-learning reinforcement learning

Q96. An organisation that owns dozens of shopping malls wants to create a machine learning product that will use facial recognition to identify customers. What is the main challenge of developing such a model?

- [ ] most machine learning models are not designed for video
- [ ] it might be unethical for the business to identify people without their consent
- [ ] it will be difficult to decide between supervised and unsupervised learning
- [ ] the image in the video would not be high quality enough to identify individuals

Q97. Which of the following machine learning algorithms is unsupervised?

- [ ] Random forest
- [ ] k-nearest neighbors
- [ ] Support-vector machines
- [ ] K- means

Explanation:
During training, k-means partitions observations into k clusters. During inference, it assigns a given data point to the nearest cluster by distance. k-means is unsupervised, because it doesn't require labeled data to be trained.

Q98. Averaging the output of multiple decision trees helps to::

- [ ] Increase variance
- [ ] Increase bias
- [ ] Decrease variance
- [ ] Decrease bias

Explanation: Averaging models leads to higher stability and a lower variance than individual models. Mathematically, remember that $\text{Var}(\bar{X})=\frac{\text{Var}(X)}{N}$

Q99. To optimize your objective function, you are performing full batch gradient descent using the entire training set (not stochastic gradient descent). Is it required to shuffle your training set?

- [ ] Yes. If you don't, the optimization will oscillate around the minimum at the end of training.
- [ ] Yes, in order to help the model generalize to the test dataset.
- [ ] No, it is not necessary because the dataset can already be considered shuffled from the data collection process.
- [ ] No, because each update passes through the entire dataset anyway and the order doesn't matter.

Explanation: At every iteration, full batch gradient descent uses the entire training set to compute a gradient. The order in which data is processed doesn't impact the gradient value.

At every iteration, full batch gradient descent uses the entire training set to compute a gradient. The order in which data is processed doesn't impact the gradient value.

Q100. You've received 1,000,000 images and have split it in 96%/2%/2% between train, dev and test sets. You've trained your model, and analyzed the results. After working further on the problem, you've decided to correct the incorrectly labeled data on the dev set.

Which of these statements do you agree with?

- [ ] You should also correct the incorrectly labeled data in the test set, so that the dev and test sets still come from the same distribution.
- [ ] You should correct incorrectly labeled data in the training set as well so as to avoid your training set now being even more different from your dev set.
- [ ] You should not correct the incorrectly labeled data in the test set, because the test set should reflect the data distribution of the real world.
- [ ] If you want to correct incorrectly labeled data, you should do it on all three sets (train/dev/test) in order to maintain similar distributions.

Explanation: It is important that your dev and test set have the closest possible distribution to "real" data.

Q101. You're working on a binary classification task, to classify if an image contains a cat ("1") or doesn't contain a cat ("0"). What loss would you choose to minimize in order to train a model?

- [ ] $L = y \log \hat{y} + (1-y) \log (1- \hat{y})$
- [ ] $L = - y \log \hat{y} - (1-y) \log (1- \hat{y})$
- [ ] $L = || y - \hat{y} ||_2^2$
- [ ] $L = || y - \hat{y} ||_2^2 + constant$

Explanation: You are trying to minimize the binary cross entropy loss over the training set..

Q102. You want to create a machine learning algorithm that finds the top 100 people who have shared photographs of themselves on social media. What is the best machine learning method to use?

- [ ] reinforcement learning
- [ ] binary classification
- [ ] K-nearest neighbor
- [ ] unsupervised learning

Q103. The famous data scientist Andrew Ng has been quoted as saying, "Applied machine learning is basically feature engineering." What is feature engineering?

- [ ] scraping new features from web data
- [ ] creating new variables by combining and modifying the original variables
- [ ] designing innovative new user features to add to software
- [ ] using deep learning to find features in the data

Q104. In the context of calculus, what is df/dx?

- [ ] the prediction function
- [ ] the derivative of f of x
- [ ] the derivative of x
- [ ] equivalent to f divided by x

Q105. What is a well-designed/well-fitted model?

- [ ] one that has been trained and tested with the same data
- [ ] one that has a high degree of accuracy and is able to accurately predict results
- [ ] one that has been trained with labeled training data
- [ ] one that has been trained with an exhaustive set of all conditions and permutations in the training data

Q106. **_**-based collaborative filtering occurs when a person is recommended an item similar to an item they have purchased.

- [ ] History
- [ ] Item
- [ ] Similarity
- [ ] Purchase

Q107. Fill in the blanks: Two multivariate imputer techniques are the **_** imputer and the **_** imputer.

- [ ] supervised, unsupervised
- [ ] iterative, KNN
- [ ] similarity, regressive
- [ ] normalized, scaled

Q108. You are working on a regression model using the Keras library. What method on the Model class do you use to train the model?

- [ ] predict

- [ ] compile
- [ ] fit
- [ ] get_weights

Q109. What is the goal of regularization in the K nearest neighbors algorithm?

- [ ] normalizing the data points so they can be compared with each other
- [ ] using a straight line model to make predictions based on training data
- [ ] finding the slope of the line that represents the model
- [ ] making the decision boundaries more regular

Q110. If there is no trend between two variables x and y, we say that there is a **_** connection between x and y.

- [ ] linear
- [ ] exponential
- [ ] non-random
- [ ] random

Q111. If you are thinking about using machine learning algorithms, the best thing you can do today is to ensure you have quality **_**.

- [ ] data
- [ ] processors
- [ ] networking
- [ ] statistical techniques

Explanation: "Ensuring you have good data quality prior to running machine learning algorithms is a crucial step within the overall data science and machine learning workflow." [Source](https://towardsdatascience.com/data-quality-considerations-for-machine-learning-models-dcbe9cab34cb)

Q112. Your organization's chief diversity officer is concerned that your engineering department lacks racial and gender diversity. You are asked to create a supervised machine learning system to help sort through hundreds of thousands of new employment applications. The human resources department insists on using internal hiring data. What are some of the dangers that you might run into?

- [ ] There will be too much data for your artificial neural network to process efficiently.
- [ ] Machine learning systems cannot define diversity, so there is no way to use one to improve hiring.
- [ ] Machine learning systems cannot be used with this type of data.
- [ ] If the system uses internal data, then it may amplify any existing bias in hiring.

Explanation: "If an AI is trained on a biased data set, it will naturally make biased decisions which can give calamitous results."
[Source](https://www.linkedin.com/pulse/reducing-bias-ai-diversity-inclusion-shashwat-siddhant/)

Q113. In 2013, Google´s DeepMind project created a machine learning algorithm that could play an old-style Atari video game, Pong. The algorithm taught the machine how to play by creating a series of rewards. Each time the machine successfully returned the ball, the machine got a reward; each time the opponent missed the ball, the machine got a reward. How would you describe this type of machine learning algorithm?

- [ ] big data machine learning.
- [ ] Good Old-Fashioned Artificial Intelligence (GOFAI).
- [ ] reinforcement learning.
- [ ] supervised learning.

Explanation: Reinforcement learning is the branch of machine learning where the algorithm interacts with the environment and gets rewards or penalizations
[Source](https://www.deepmind.com/publications/playing-atari-with-deep-reinforcement-learning)

Q114. An organization that owns dozens of shopping malls want to create a machine learning product that will use facial recognition to identify customers. What is one of the main challenges with developing such a product?

- [ ] The images in the video would not be high-quality enough to identify individuals.
- [ ] It would be difficult to decide between supervised and unsupervised learning.
- [ ] It might be unethical for the business to identify people without their consent.
- [ ] Most machine learning algorithms are not designed for video.

Explanation: there are many ethical questions about consent and privacy in machine learning algorithms [Source](https://www.nature.com/articles/d41586-020-03187-3)

answer key
114 c
113 c
112 d
111 a

110	d
109	d
108	c
107	b
106	b
105	b
104	b
103	b
102	b
101	b
100	a
99	d
98	c
97	d
96	b
95	b
94	b
93	b
92	d
91	c
90	a
89	c
88	d
87	c
86	d
85	d
84	b
83	d
82	c
81	a
80	d
79	c
78	a
77	c
76	a
75	c
74	d
73	c
72	b

71	d
70	b
69	a
68	b
67	d
66	a
65	a
64	c
63	d
62	c
61	a
60	c
59	c
58	c
57	c
56	c
55	c
54	d
53	c
52	d
51	c
50	b
49	d
48	c
47	d
46	d
45	d
44	a
43	a
42	c
41	c
40	a
39	d
38	b
37	b
36	b
35	a
34	a
33	b

32	d
31	b
30	c
29	a
28	a
27	c
26	d
25	a
24	a
23	d
22	a
21	a
20	d
19	d
18	c
17	b
16	b
15	b
14	a
13	c
12	a
11	c
10	a
9	d
8	a
7	d
6	d
5	a
4	d
3	d
2	d
1	b