Machine Learning Interview Questions

- This series of practice questions will focus on your general knowledge of some machine learning algorithms.
- These will be mainly asking you about your understanding of different machine learning algorithms.

- The best way to approach these questions is to “whiteboard” them.
- Use a whiteboard or piece of paper to help convey your understanding of the solutions to the questions.

- We'll ask about the following algorithms
    - Linear Regression
    - Logistic Regression
    - Decision Trees and Random Forests
    - Naive Bayes
    - Support Vector Machines (SVM)
    - Model Evaluation and Training

Machine Learning Interview Question 1

- What are the main assumptions of a linear regression?
- What are the most common types of linear regression? (More formally put, what are the most common estimation techniques for linear regression)

Machine Learning Interview Question 2

- Describe the formula for logistic regression and how the algorithm is used for binary classification.

Machine Learning Interview Question 3

- How does a Decision Tree decide on its splits (what is the criteria for a split point)?

Machine Learning Interview Question 4

- What advantages does a decision tree model have?

Machine Learning Interview Question 5

- What is the difference between a random forest versus boosting tree algorithms?

Machine Learning Interview Question 6

- Given a data set of features **X** and labels **y** , what assumptions are made when using Naive Bayes methods?

Machine Learning Interview Question 7

- Describe how the support vector machine (SVM) algorithm works.

Machine Learning Interview Question 8

- What is overfitting and what causes it?
- What ways can you attempt to avoid overfitting?

Machine Learning Interview Question 9

- Describe the differences between accuracy, precision, and recall.

Machine Learning Interview Question 10

- What metrics can be used to evaluate a regression task?

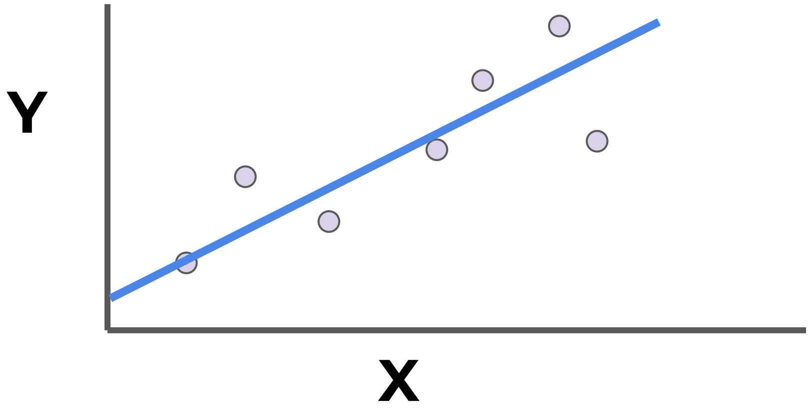
SOLUTIONS

SOLUTIONS ARE UP NEXT!

Solution to Machine Learning Question 1

- What are the main assumptions of a linear regression?
- What are the most common types of linear regression? (More formally put, what are the most common estimation techniques for linear regression)

- A linear regression models the relationship between the dependent variable y and the independent variable x



- Two main assumptions are:
  - The relationship between the dependent variable y and the explanatory variables X is linear
  - The residual errors from the regression fit are normally distributed.

- Most common types of linear regression
  - Ordinary Least Squares
  - Generalized Least Squares
  - Penalized Least Squares
    - L1 (LASSO)
    - L2 (Ridge)

Solution to Machine Learning Question 2

- Describe the formula for logistic regression and how the algorithm is used for binary classification.

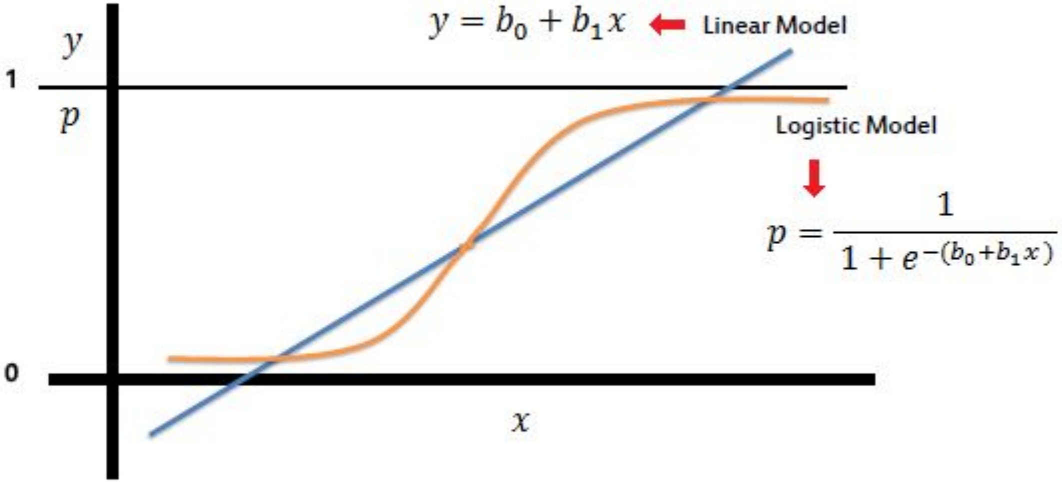
- Formula for Logistic Regression

$$\delta(t) = \frac{1}{1 + e^{-t}}$$

$$t = \beta_0 + \beta_1 x$$

$$f(x) = \delta(t) = \frac{1}{1 + e^{\beta_0 + \beta_1 x}}$$

- We can then use the result of f(x) as the probability of the data point being in the positive class.



Solution to Machine Learning Question 3

- How does a Decision Tree decide on its splits (what is the criteria for a split point)?

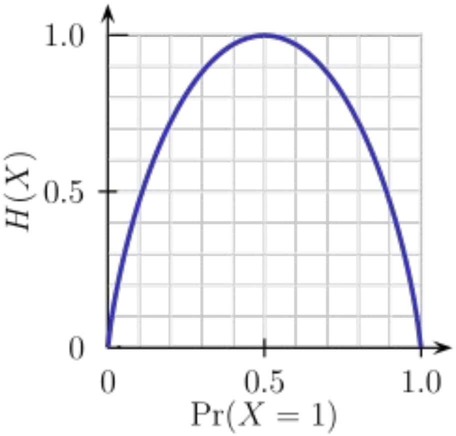
- A decision tree can use the Information Gain to decide on the splitting criteria.
- Let’s give a brief overview of how this works.

- The decision tree is built in a top-down fashion, but the question is how do you choose which attribute to split at each node?
- The answer is find the feature that best splits the target class into the purest possible children nodes.

- This measure of purity is called the information.
- It represents the expected amount of information that would be needed to specify whether a new instance should be classified 0 or 1, given the example that reached the node.

Data Science Career Guide

- Entropy on the other hand is a measure of impurity (the opposite). It is defined for a binary class with values a/b as:
    - **- p(a)*log(p(a)) - p(b)*log(p(b))**



- Now by comparing the entropy before and after the split, we obtain a measure of information gain, or how much information we gained by doing the split using that particular feature:
- **Information_Gain = Entropy_before - Entropy_after**

Solution to Machine Learning Question 4

- What advantages does a decision tree model have?

- Advantages of Decision Trees
  - Very easy to interpret and understand
  - Works on both continuous and categorical features
  - No normalization or scaling necessary
  - Prediction algorithm runs very fast

Solution to Machine Learning Question 5

- What is the difference between a random forest versus boosting tree algorithms?

- Boosting Trees
  - Reassign weights to samples based on the results of previous iterations of classifications.
  - Harder to classify points get weighted more.

- Boosting Trees
  - Iterative algorithm where each execution is based on the previous results.

- Random Forest
    - RF applies bootstrap aggregation to train many different trees.
    - This creates an ensemble of different individual decision trees

- Random Forest
  - In random forest algorithm, Instead of using information gain or gini index for calculating the root node, the process of finding the root node and splitting the feature nodes will happen randomly.

Solution to Machine Learning Question 6

- Given a data set of features **X** and labels **y** , what assumptions are made when using Naive Bayes methods?

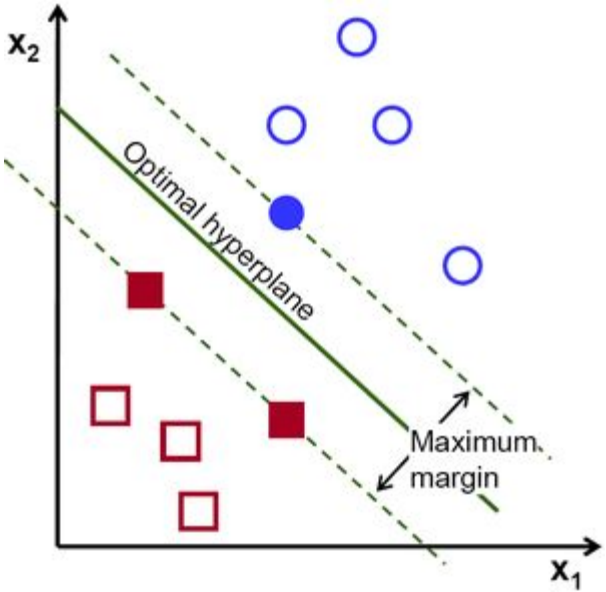
- Naive Bayes algorithm assumes that the features of **X** are conditionally independent of each other for the given **Y**.

- The idea that each feature is independent of each other may not always be true, but we assume it to be true to apply Naive Bayes. This "naive" assumption is where the namesake comes from.

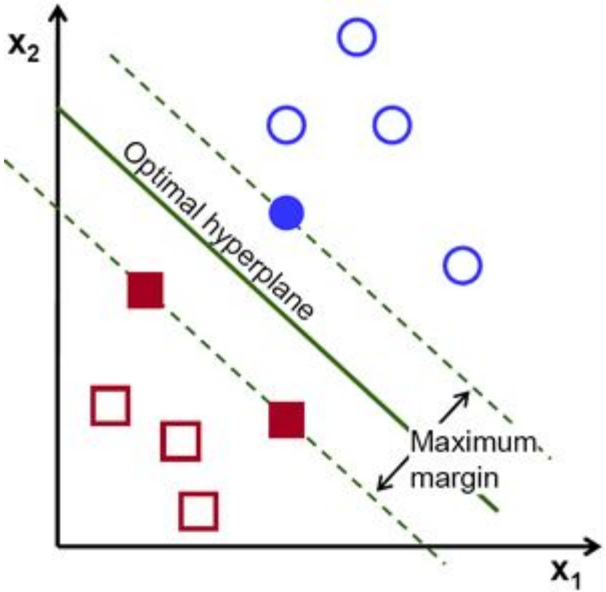
Solution to Machine Learning Question 7

- Describe how the support vector machine (SVM) algorithm works.

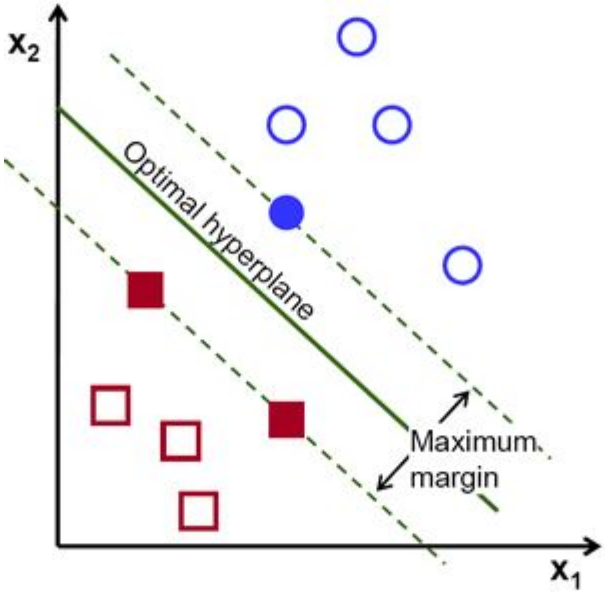
- SVM attempt to find a hyperplane that separates classes by maximizing the margin



- The filled in points in this diagram are the support vectors, against the decision hyperplane



- Here we show linear classification, but SVMs can perform nonlinear classification



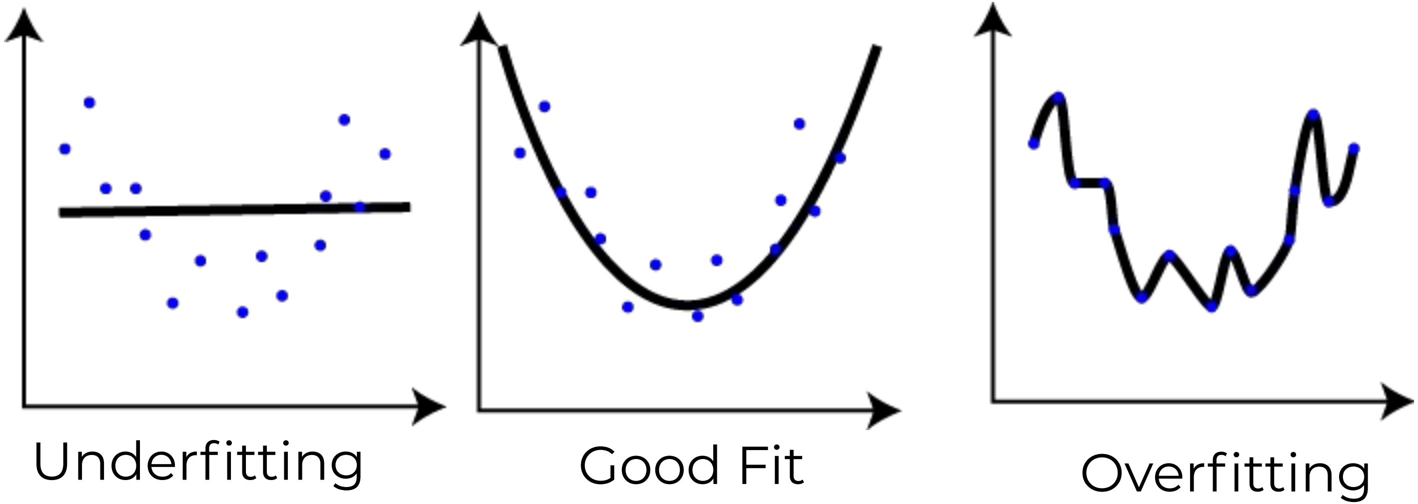
- SVMs can employ the kernel trick which can map linear non-separable inputs into a higher dimension where they become more easily separable.

Solution to Machine Learning Question 8

- What is overfitting and what causes it?
- What ways can you attempt to avoid overfitting?

- Overfitting is when the ML model does not generalize well to data it has not seen before. It “overfits” to the training data, usually indicating it has too much complexity in regards to the training data size.

Data Science Career Guide



- Many ways to attempt to fix overfitting
    - Increase Training Data Size
    - Regularization
    - Early Stopping
    - K-Fold Cross Validation

Solution to Machine Learning Question 9

Data Science Career Guide

- Describe the differences between accuracy, precision, and recall.

- Let's imagine the results of a model

	Predicted class	
Actual Class	Class = Yes	Class = No
Class = Yes	True Positive	False Negative
Class = No	False Positive	True Negative

- Accuracy is simply a ratio of correctly predicted observation to the total observations. Accuracy is a useful measure but only when you have symmetric datasets where values of false positive and false negatives are almost same. Therefore, we need to also consider Precision and Recall.
- **Accuracy = (TP+TN) / (TP+FP+FN+TN)**

	Predicted class		
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

Data Science Career Guide

- Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all points that labeled as positive, how many were actually positive?
- **Precision = TP / (TP+FP)**

	Predicted class	
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

Data Science Career Guide

- Recall (Sensitivity) is the ratio of correctly predicted positive observations to the all observations in actual class.
- **Recall = TP / (TP+FN)**

		Predicted class	
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

Solution to Machine Learning Question 10

- What metrics can be used to evaluate a regression task?

Mean Absolute Error (MAE) is the mean of the absolute value of the errors:

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Mean Squared Error (MSE) is the mean of the squared errors:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Root Mean Squared Error (RMSE) is the square root of the mean of the squared errors:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- MAE is the easiest to understand, because it's the average error.
- MSE is more popular than MAE, because MSE "punishes" larger errors, which tends to be useful in the real world.
- RMSE is even more popular than MSE, because RMSE is interpretable in the "y" units.