A Crash Course in Statistical Learning – How much Data do you need?

Quiz Questions

Makarand Madhavi

Q1) Question: Which of the following statements about test and train data in machine learning are correct?

A. The purpose of train data is to fit the model, while the purpose of test data is to evaluate the model's performance.
B. It is acceptable to use the same data for both training and testing a model.
C. Cross-validation is a technique used to evaluate a model's performance using multiple splits of the data.
D. Overfitting occurs when a model performs well on the train data but poorly on the test data.

Answer: A, C, and D are correct.

A. The purpose of train data is to fit the model, while the purpose of test data is to evaluate the model's performance.
This statement is correct. Train data is used to train and fit a machine learning model, while test data is used to evaluate the model's performance on new, unseen data.

B. It is acceptable to use the same data for both training and testing a model.
This statement is incorrect. Using the same data for training and testing a model can lead to overfitting, where the model memorizes the training data instead of learning the underlying patterns in the data.

C. Cross-validation is a technique used to evaluate a model's performance using multiple splits of the data.
This statement is correct. Cross-validation is a technique used to evaluate a model's performance by splitting the data into multiple train and test sets, and averaging the results across the different splits.

D. Overfitting occurs when a model performs well on the train data but poorly on the test data.
This statement is correct. Overfitting occurs when a model fits the training data too closely, and as a result, performs poorly on new, unseen data. This is a common problem in machine learning, and can be addressed by using techniques such as regularization and cross-validation.

Q2) Question: Which of the following statements about null deviance and residual deviance in machine learning are correct?

A. Null deviance measures the amount of variability in the response variable that cannot be explained by the model when only the intercept is used.

B. Residual deviance measures the amount of variability in the response variable that cannot be explained by the model when all the predictors are included.
C. Null deviance is always larger than residual deviance.
D. The ratio of null deviance to residual deviance can be used to assess the goodness-of-fit of a model.

Answer: A, B, and D are correct.

A. Null deviance measures the amount of variability in the response variable that cannot be explained by the model when only the intercept is used.
This statement is correct. Null deviance is a measure of the amount of variability in the response variable that cannot be explained by the model when only the intercept (i.e., no predictors) is used.

B. Residual deviance measures the amount of variability in the response variable that cannot be explained by the model when all the predictors are included.
This statement is correct. Residual deviance is a measure of the amount of variability in the response variable that cannot be explained by the model when all the predictors are included.

C. Null deviance is always larger than residual deviance.
This statement is incorrect. Null deviance is typically larger than residual deviance, but this is not always the case. If the model fits the data well, then the residual deviance may be larger than the null deviance.

D. The ratio of null deviance to residual deviance can be used to assess the goodness-of-fit of a model.
This statement is correct. The ratio of null deviance to residual deviance can be used to assess the goodness-of-fit of a model. A smaller ratio indicates a better fit of the model to the data. However, this measure should be used with caution, as it is sensitive to the sample size and number of predictors in the model.

Q3) Question: Which of the following statements about Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) in machine learning are correct?

A. MSE is the average of the squared differences between the predicted and actual values.
B. RMSE is the square root of the variance of the errors.
C. MSE and RMSE are both measures of the accuracy of a regression model.
D. RMSE is always greater than or equal to MSE.

Answer: A, C, and D are correct.

A. MSE is the average of the squared differences between the predicted and actual values.

This statement is correct. MSE is a measure of the average squared difference between the predicted and actual values, and is calculated by taking the average of the squared differences between the predicted and actual values.

B. RMSE is the square root of the variance of the errors.
This statement is incorrect. RMSE is the square root of the MSE, and is not directly related to the variance of the errors.

C. MSE and RMSE are both measures of the accuracy of a regression model.
This statement is correct. Both MSE and RMSE are measures of the accuracy of a regression model, with lower values indicating better accuracy.

D. RMSE is always greater than or equal to MSE.
This statement is correct. RMSE is the square root of MSE, so it is always greater than or equal to MSE.

Q4) Which of the following statements about Shapley values in machine learning are correct?

A. Shapley values are a way to estimate the contribution of each feature to the prediction of a model.
B. Shapley values can be calculated using a combinatorial approach.
C. Shapley values provide an approximation of the average marginal contribution of a feature across all possible feature subsets.
D. Shapley values are always positive and additive across features.

Answer: A, B, and C are correct.

A. Shapley values are a way to estimate the contribution of each feature to the prediction of a model.
This statement is correct. Shapley values are a technique for estimating the contribution of each feature to the prediction of a model, by considering the marginal contribution of each feature to the model prediction.

B. Shapley values can be calculated using a combinatorial approach.
This statement is correct. Shapley values can be calculated using a combinatorial approach, which involves calculating the contribution of each feature to the model prediction for all possible subsets of features.

C. Shapley values provide an approximation of the average marginal contribution of a feature across all possible feature subsets.
This statement is correct. Shapley values provide an approximation of the average marginal contribution of a feature across all possible feature subsets, and can be used to rank the importance of features in a model.

D. Shapley values are always positive and additive across features.
This statement is incorrect. Shapley values can be positive or negative, depending on whether a feature has a positive or negative impact on the model prediction. Additionally, Shapley values are not always additive across features, as the contribution of each feature can be dependent on the presence or absence of other features in the model.

Q5) Which of the following statements about Gini importance in machine learning are correct?

A. Gini importance is a measure of the total reduction of the impurity of a node in a decision tree when a given feature is used for splitting.
B. Gini importance can only be used for decision trees and random forests.
C. Gini importance provides a ranking of feature importance in a model.
D. Gini importance can be used to identify interactions between features.

Answer: A, C, and D are correct.

A. Gini importance is a measure of the total reduction of the impurity of a node in a decision tree when a given feature is used for splitting.
This statement is correct. Gini importance is a measure of the total reduction of the impurity of a node in a decision tree when a given feature is used for splitting, and is calculated as the sum of the impurity decrease over all the nodes in which the feature is used for splitting.

B. Gini importance can only be used for decision trees and random forests.
This statement is incorrect. While Gini importance is commonly used for decision trees and random forests, it can also be applied to other machine learning models, such as gradient boosting and support vector machines.

C. Gini importance provides a ranking of feature importance in a model.
This statement is correct. Gini importance provides a ranking of feature importance in a model, with higher values indicating greater importance.

D. Gini importance can be used to identify interactions between features.
This statement is correct. Gini importance can be used to identify interactions between features, as it takes into account the impact of each feature on the impurity reduction at each node in the decision tree. This can reveal whether the impact of one feature is dependent on the presence or absence of another feature.

Q6) Question: Which of the following statements about the relationship between model complexity, computational resources, and amount of data in machine learning are correct?

A. Increasing the model complexity can improve its predictive power, but also increase the risk of overfitting.
B. Increasing the computational resources can always compensate for a smaller amount of data.

C. Increasing the amount of data can reduce the risk of overfitting and improve the generalization performance of a model.
D. Increasing the amount of data can always lead to an improvement in the model performance.

Answer: A and C are correct.

A. Increasing the model complexity can improve its predictive power, but also increase the risk of overfitting.
This statement is correct. Increasing the model complexity can allow it to capture more complex relationships in the data, leading to improved predictive power. However, increasing the model complexity can also increase the risk of overfitting, which occurs when the model becomes too complex and starts to fit the noise in the data instead of the underlying patterns.

B. Increasing the computational resources can always compensate for a smaller amount of data.
This statement is incorrect. While increasing the computational resources, such as using more powerful hardware or parallel processing, can improve the training time and allow for more complex models, it cannot compensate for a smaller amount of data. With a smaller amount of data, even the most powerful computational resources cannot extract more information that simply isn't present in the dataset.

C. Increasing the amount of data can reduce the risk of overfitting and improve the generalization performance of a model.
This statement is correct. Increasing the amount of data can provide more information to the model, making it less likely to overfit the data and improving its ability to generalize to new, unseen data.

D. Increasing the amount of data can always lead to an improvement in the model performance.
This statement is incorrect. While increasing the amount of data can often lead to improved model performance, there is a point of diminishing returns. After a certain point, adding more data may not result in significant improvement in model performance. Additionally, there may be cases where the quality of the additional data is low, or the data is not representative of the population, which can actually harm the model performance.

Q7) Question: Which of the following statements about the relationship between overfitting and the size of the data in machine learning are correct?

A. Overfitting occurs more frequently when the size of the data is small.
B. Overfitting occurs less frequently when the size of the data is large.
C. Increasing the size of the data can always prevent overfitting.
D. Increasing the size of the data can sometimes worsen overfitting.

Answer: A and D are correct.

A. Overfitting occurs more frequently when the size of the data is small.

This statement is correct. When the size of the data is small, the model has fewer examples to learn from, which can lead to overfitting. With a smaller amount of data, it is easier for the model to memorize the training data instead of learning the underlying patterns, resulting in poor generalization performance.

B. Overfitting occurs less frequently when the size of the data is large.
This statement is incorrect. While increasing the size of the data can reduce the risk of overfitting, it is not a guarantee. Overfitting can still occur with a large amount of data if the model is too complex and there is noise or irrelevant information in the data.

C. Increasing the size of the data can always prevent overfitting.
This statement is incorrect. While increasing the size of the data can reduce the risk of overfitting, it is not a guarantee. There may be cases where the quality of the additional data is low, or the data is not representative of the population, which can actually harm the model performance.

D. Increasing the size of the data can sometimes worsen overfitting.
This statement is correct. While increasing the size of the data can help reduce the risk of overfitting, it can also worsen it in certain cases. For example, if the additional data contains outliers or noise, it may actually make the model more prone to overfitting. Additionally, if the model is too complex, adding more data may not help and may actually make the problem worse.

Q8) Question: Which of the following statements about evaluating feature importance in machine learning are correct?

A. Permutation feature importance is a model-agnostic method that involves randomly permuting the values of a feature and measuring the decrease in model performance.
B. Tree-based models such as random forests and gradient boosting can provide feature importance scores based on the reduction in impurity or loss when splitting on a feature.
C. Principal component analysis (PCA) can be used to rank features based on their contribution to the variance in the data.
D. Correlation coefficients between each feature and the target variable can be used to measure feature importance.

Answer: A, B, and C are correct.

A. Permutation feature importance is a model-agnostic method that involves randomly permuting the values of a feature and measuring the decrease in model performance.
This statement is correct. Permutation feature importance is a model-agnostic method that involves randomly permuting the values of a feature and measuring the decrease in model performance. The decrease in performance after permuting a feature is used as an indication of the importance of that feature in the model.

B. Tree-based models such as random forests and gradient boosting can provide feature importance scores based on the reduction in impurity or loss when splitting on a feature.
This statement is correct. Tree-based models such as random forests and gradient boosting can provide feature importance scores based on the reduction in impurity or loss when splitting on a feature. Features that result in the greatest reduction in impurity or loss when used in a split are considered to be more important.

C. Principal component analysis (PCA) can be used to rank features based on their contribution to the variance in the data.
This statement is correct. Principal component analysis (PCA) is a method that can be used to identify the most important features by ranking them based on their contribution to the variance in the data. Features that contribute more to the variance are considered to be more important.

D. Correlation coefficients between each feature and the target variable can be used to measure feature importance.
This statement is incorrect. Correlation coefficients can be used to measure the linear relationship between two variables, but they may not capture all aspects of the relationship between features and the target variable. Correlation may also be confounded by other variables, and may not capture non-linear relationships. Other methods such as permutation feature importance and tree-based feature importance are more reliable methods for measuring feature importance.

Q9) Which of the following statements about 70/30 or 80/20 test/train split in machine learning are correct?

A. The purpose of test data is to evaluate the model's performance, while the purpose of train data is to fit the model.
B. The test data should always be larger than the train data.
C. A larger test set can give more reliable estimates of a model's performance, but a smaller test set can give a more accurate estimate of the model's performance on new data.
D. The 70/30 or 80/20 split is a fixed rule that should always be followed.

Answer: A and C are correct.

A. The purpose of test data is to evaluate the model's performance, while the purpose of train data is to fit the model.
This statement is correct. The purpose of train data is to fit the machine learning model, while the purpose of test data is to evaluate the model's performance on new, unseen data.

B. The test data should always be larger than the train data.
This statement is incorrect. There is no fixed rule on the size of the train or test data, as it depends on the size and complexity of the dataset, as well as the specific problem being addressed.

C. A larger test set can give more reliable estimates of a model's performance, but a smaller test set can give a more accurate estimate of the model's performance on new data.
This statement is correct. A larger test set can give more reliable estimates of a model's performance, as it reduces the variance in the evaluation metric, but a smaller test set can give a more accurate estimate of the model's performance on new data, as it better reflects the size and distribution of new data that the model will encounter in the real world.

D. The 70/30 or 80/20 split is a fixed rule that should always be followed.
This statement is incorrect. The choice of the train/test split ratio is not a fixed rule, and depends on the specific problem being addressed, the size and complexity of the dataset, and the available computing resources. The 70/30 or 80/20 split is a common starting point, but it may be necessary to adjust the split ratio depending on the specific problem and dataset.

Q10) Question: Which of the following statements about big data in machine learning is/are true?

A. Big data refers to datasets that are too large and complex to be processed by traditional data processing methods.
B. The more data a machine learning model has, the better its performance will be.
C. Big data has made it easier to avoid overfitting in machine learning models.
D. One of the challenges of big data in machine learning is data quality.

Answer: A, B, and D.

A. Big data refers to datasets that are too large and complex to be processed by traditional data processing methods.
This statement is true. Big data typically involves datasets that are too large and complex to be processed using traditional data processing methods, such as spreadsheets or databases.

B. The more data a machine learning model has, the better its performance will be.
This statement is generally true. In machine learning, having more data to train a model can lead to better performance, especially when working with complex models such as deep learning neural networks.

C. Big data has made it easier to avoid overfitting in machine learning models.
This statement is false. While having more data can help to prevent overfitting, big data itself does not inherently make it easier to avoid overfitting. In fact, working with big data can actually make it more challenging to identify and address overfitting.

D. One of the challenges of big data in machine learning is data quality.
This statement is true. One of the challenges of working with big data in machine learning is ensuring that the data is of sufficient quality to train accurate models. This may involve dealing with missing data, data errors, and data inconsistencies.