Questions

05 January 2021 13:52

Statistics

1. **What is correlation and covariance in statistics?**
   Ans:
   Covariance and Correlation are two mathematical concepts; these two approaches are widely used in statistics. Both Correlation and Covariance establish the relationship and also measure the dependency between two random variables. Though the work is similar between these two in mathematical terms, they are different from each other.
   Correlation: Correlation is considered or described as the best technique for measuring and also for estimating the quantitative relationship between two variables. Correlation measures how strongly two variables are related.
   Covariance: In covariance two items vary together and it's a measure that indicates the extent to which two random variables change in cycle. It is a statistical term; it explains the systematic relation between a pair of random variables, wherein changes in one variable reciprocal by a corresponding change in another variable.

2. **What is Central Limit Theorem and where do we use it ?**

3. **encoding poison distribution ?**
   Ans:
   A Poisson distribution is a tool that helps to predict the probability of certain events from happening when you know how often the event has occurred. It gives us the probability of a given number of events happening in a fixed interval of time
   <https://www.statisticshowto.com/poisson-distribution/>
4. **model traffic at the gate ?**
   Ans:
5. **What is F - Test ?**

Machine Learning

Fundamentals

1. **What are some of the steps for data wrangling and data cleaning before applying machine learning algorithms?**
   Ans:
   There are many steps that can be taken when data wrangling and data cleaning. Some of the most common steps are listed below:
   1. **Data profiling**: Almost everyone starts off by getting an understanding of their dataset. More specifically, you can look at the shape of the dataset with .shape and a description of your numerical variables with .describe().
   2. **Data visualizations**: Sometimes, it's useful to visualize your data with histograms, boxplots, and scatterplots to better understand the relationships between variables and also to identify potential outliers.
   3. **Syntax error**: This includes making sure there's no white space, making sure letter casing is consistent, and checking for typos. You can check for typos by using .unique() or by using bar graphs.
   4. **Standardization or normalization**: Depending on the dataset your working with and the machine learning method you decide to use, it may be useful to standardize or normalize your data so that different scales of different variables don't negatively impact the performance of your model.
   5. **Handling null values**: There are a number of ways to handle null values including deleting rows with null values altogether, replacing null values with the mean/median/mode, replacing null values with a new category (eg. unknown), predicting the values, or using machine learning models that can deal with null values. Read more here.
   6. **Other things include**: removing irrelevant data, removing duplicates, and type conversion.
2. **Describe different regularization methods, such as L1 and L2 regularization ?**
   Ans:
   Both L1 and L2 regularization are methods used to reduce the overfitting of training data. Least Squares minimizes the sum of the squared residuals, which can result in low bias but high variance.
   L2 Regularization, also called ridge regression, minimizes the sum of the squared residuals plus lambda times the slope squared. This additional term is called the Ridge Regression Penalty. This increases the bias of the model, making the fit worse on the training data, but also decreases the variance.
   If you take the ridge regression penalty and replace it with the absolute value of the slope, then you get Lasso regression or L1 regularization.
   L2 is less robust but has a stable solution and always one solution. L1 is more robust but has an unstable solution and can possibly have multiple solutions
3. **How to define/select metrics?**
   Ans:
   There isn't a one-size-fits-all metric. The metric(s) chosen to evaluate a machine learning model depends on various factors:
   Is it a regression or classification task?
   What is the business objective? Eg. precision vs recall
   What is the distribution of the target variable?
   There are a number of metrics that can be used, including adjusted r-squared, MAE, MSE, accuracy, recall, precision, f1 score, and the list goes on.

4. **Why is dimension reduction important?**
   Ans:
   Dimensionality reduction is the process of reducing the number of features in a dataset. This is important mainly in the case when you want to reduce variance in your model (overfitting).
   Wikipedia states four advantages of dimensionality reduction (see here):
   It reduces the time and storage space required
   Removal of multi-collinearity improves the interpretation of the parameters of the machine learning model
   It becomes easier to visualize the data when reduced to very low dimensions such as 2D or 3D
   It avoids the curse of dimensionality

5. **What is bias-variance trade-off?**
   Ans:
   Bias: Bias is an error introduced in your model due to oversimplification of the machine learning algorithm. It can lead to underfitting. When you train your model at that time model makes simplified assumptions to make the target function easier to understand.
   Low bias machine learning algorithms — Decision Trees, k-NN and SVM High bias machine learning algorithms — Linear Regression, Logistic Regression
   Variance: Variance is error introduced in your model due to complex machine learning algorithm, your model learns noise also from the training data set and performs badly on test data set. It can lead to high sensitivity and overfitting.
   Normally, as you increase the complexity of your model, you will see a reduction in error due to lower bias in the model. However, this only happens until a particular point. As you continue to make your model more complex, you end up over-fitting your model and hence your model will start suffering from high variance.

   Bias-Variance trade-off: The goal of any supervised machine learning algorithm is to have low bias and low variance to achieve good prediction performance.

1. The k-nearest neighbor algorithm has low bias and high variance, but the trade-off can be changed by increasing the value of k which increases the number of neighbors that contribute to the prediction and in turn increases the bias of the model.
2. The support vector machine algorithm has low bias and high variance, but the trade-off can be changed by increasing the C parameter that influences the number of violations of the margin allowed in the training data which increases the bias but decreases the variance.

There is no escaping the relationship between bias and variance in machine learning. Increasing the bias will decrease the variance. Increasing the variance will decrease bias.

6. **Factors that affect the model/algorithm we choose ?**

Ans:

Based On Data / Type of Business Problem (regression, classification clustering etc.)
Accuracy of the model
Interpretability of the model
Complexity of the model
Scalability of the model
Time it takes to build, train, and test the model
Time it takes to make predictions using the model
If the model meets your business goals

7. **How to combat Overfitting and Underfitting?**

Ans:

To combat overfitting and underfitting, you can resample the data to estimate the model accuracy (k-fold cross-validation) and by having a validation dataset to evaluate the model.
Regularization. In Neural networks -> Dropouts etc.

8. Correlation coefficients between different kinds of variables ?

Ans: https://medium.com/@outside2SDs/an-overview-of-correlation-measures-between-categorical-and-continuous-variables-4c7f85610365#:~:text=Here's%20the%20problem%3A%20there%20are,categorical%20and%20categorical%2Dcontinuous%20variable

9. Problems with Multi collinearity?

Ans: https://towardsdatascience.com/multicollinearity-why-is-it-a-problem-398b010b77ac

10. **What is Curse of Dimensionality and how to Mitigate it ?**

Ans: 1) Data Sparsity 2) Distance Concentration

Mitigating Curse of Dimensionality

1) Feature selection Techniques
2) Low Variance Filter
3) High Correlation Filter
4) Multicollinearity
5) Feature Ranking
6) Adjusted R2

https://www.mygreatlearning.com/blog/understanding-curse-of-dimensionality/

11. **Normal distribution and its applications in a data science problem ?**

Ans :https://medium.com/analytics-vidhya/the-normal-distribution-for-data-scientists-6de041a01cb9

12. **When to use Label encoder and When to use One Hot Encoding ?**

Ans :

challenge :

label encoding uses alphabetical ordering , here is a very high probability that the model
captures the relationship of order of the variable.
When there are more no of categories for a feature it may increase number of dimensions if we perform one hot encoding then we may conider
Label encoding.
When Variables are ordinal then we may use label encoding.

13. **Challenges of One-Hot Encoding: Dummy Variable Trap**

1) One-Hot Encoding results in a Dummy Variable Trap as the outcome of one variable can easily be predicted with the help of the remaining variables.
2) The Dummy Variable Trap leads to the problem known as multicollinearity. Multicollinearity occurs where there is a dependency between the independent features. Multicollinearity is a serious issue in machine learning models like Linear Regression and Logistic Regression.

Scenario Based Questions

1. **How to deal with unbalanced binary classification?**

Ans:

There are a number of ways to handle unbalanced binary classification (assuming that you want to identify the minority class):

1. First, you want to reconsider the metrics that you'd use to evaluate your model. The accuracy of your model might not be the best metric to look at because and I'll use an example to explain why. Let's say 99 bank withdrawals were not fraudulent and 1 withdrawal was. If your model simply classified every instance as "not fraudulent", it would have an accuracy of 99%. Therefore, you may want to consider using metrics like precision and recall.
2. Another method to improve unbalanced binary classification is by increasing the cost of misclassifying the minority class. By increasing the penalty of such, the model should classify the minority class more accurately.
3. Lastly, you can improve the balance of classes by oversampling the minority class or by under sampling the majority class.
4. Using stratified sampling.

2. **Do you think 50 small decision trees are better than a large one? Why**

Ans:

Another way of asking this question is "Is a random forest a better model than a decision tree?" And the answer is yes because a random forest is an ensemble method that takes many weak decision trees to make a strong learner. Random forests are more accurate, more robust, and less prone to overfitting.

ML ALGOS

1. **How to Determine the Optimal K for K-Means?**

From <https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb>

Ans:

The basic idea behind k-means consists of defining k clusters such that total within-cluster variation (or error) is minimum.

A cluster center is the representative of its cluster. The squared distance between each point and its cluster center is the required variation. The aim of k-means clustering is to find these k clusters and their centers while reducing the total error.

Quite an elegant algorithm. But there is a catch. How do you decide the number of clusters?

two methods that can be useful to find this mysterious k in k-Means are:

The Elbow Method

The Silhouette Method

The Elbow Method

This is probably the most well-known method for determining the optimal number of clusters. It is also a bit naive in its approach.

Calculate the Within-Cluster-Sum of Squared Errors (WSS) for different values of k, and choose the k for which WSS becomes first starts to diminish. In the plot of WSS-versus-k, this is visible as an elbow.

Within-Cluster-Sum of Squared Errors sounds a bit complex. Let's break it down:

The Squared Error for each point is the square of the distance of the point from its representation i.e. its predicted cluster center.

The WSS score is the sum of these Squared Errors for all the points.

Any distance metric like the Euclidean Distance or the Manhattan Distance can be used.

The Silhouette Method

The silhouette value measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation).

Source: Wikipedia

The range of the Silhouette value is between +1 and -1. A high value is desirable and indicates that the point is placed in the correct cluster. If many points have a negative Silhouette value, it may indicate that we have created too many or too few clusters.

2. **Random Forest Vs Gradient Boost ?**

   Ans: https://www.datasciencecentral.com/profiles/blogs/decision-tree-vs-random-forest-vs-boosted-trees-explained

Difficult

1. **Why Can't I use Linear Regression to classify a problem instead of logistic regression ?**

Ans: the predicted value is continuous, not probabilistic.

sensitive to imbalance data when using linear regression for classification.

https://towardsdatascience.com/why-linear-regression-is-not-suitable-for-binary-classification-c64457be8e28

Deep Learning

Fundamentals

1. **Purpose of Activation Function?**

   Ans:

   The ***Activation function*** is used to introduce non-linearity into the neural network helping it to learn more complex function. Without which the neural network would be only able to learn linear function which is a linear combination of its input data. An activation function is a function in an artificial neuron that delivers an output based on inputs.

2. What is vanishing gradients?

Ans:

While training an RNN, your slope can become either too small; this makes the training difficult. When the slope is too small, the problem is known as a Vanishing Gradient. It leads to long training times, poor performance, and low accuracy.

3 . What Is the Difference Between Epoch, Batch, and Iteration in Deep Learning?

Ans:

- Epoch – Represents one iteration over the entire dataset (everything put into the training model).
- Batch – Refers to when we cannot pass the entire dataset into the neural network at once, so we divide the dataset into several batches.
- Iteration – if we have 10,000 images as data and a batch size of 200. then an epoch should run 50 iterations (10,000 divided by 50).

4. What Is Dropout and Batch Normalization?

Ans:

Dropout is a technique of dropping out hidden and visible units of a network randomly to prevent overfitting of data (typically dropping 20 per cent of the nodes). It doubles the number of iterations needed to converge the network.

Batch normalization is the technique to improve the performance and stability of neural networks by normalizing the inputs in every layer so that they have mean output activation of zero and standard deviation of one.

Difficult:

1) How a very two part linear relu introduces non-linearity into neural networks ?

   https://www.quora.com/Is-a-single-layered-ReLu-network-still-a-universal-approximator/answer/Conner-Davis-2

NLP

1. **What is TF/IDF vectorization?**

   Ans:

   TF–IDF is short for term frequency-inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining.

   The TF–IDF value increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general.

2. **Stemming or lemmatization?**

   Ans: After going through the entire tutorial, you may be asking yourself when should I use Stemming and when should I use Lemmatization? The answer itself is in whatever you have learned from this tutorial. You have seen the following points:

   - Stemming and Lemmatization both generate the root form of the inflected words. The difference is that stem might not be an actual word whereas, lemma is an actual language word.
   - Stemming follows an algorithm with steps to perform on the words which makes it faster. Whereas, in lemmatization, you used WordNet corpus and a corpus for stop words as well to produce lemma which makes it slower than stemming. You also had to define a parts-of-speech to obtain the correct lemma.
   - So when to use what! The above points show that if speed is focused then stemming should be used since lemmatizers scan a corpus which consumed time and processing. It depends on the application you are working on that decides if stemmers should be used or lemmatizers. If you are building a language application in which language is important you should use lemmatization as it uses a corpus to match root forms.

CNN - IMAGE

1. **What Are the Different Layers on CNN?**

   Ans:

   1. Convolutional Layer – the layer that performs a convolutional operation, creating several smaller picture windows to go over the data.
   2. ReLU Layer – it brings non-linearity to the network and converts all the negative pixels to zero. The output is a rectified feature map.
   3. Pooling Layer – pooling is a down-sampling operation that reduces the dimensionality of the feature map.
   4. Fully Connected Layer – this layer recognizes and classifies the objects in the image.

2. **When should we use types of pooling?**