Module 6

Statistical Modeling and Visualization

▼ What should you consider when collecting data?

- Source: How was the data collected? Was the data altered or corrected in any way? Has the data been collected from the original source, or from a secondary source?
- Accuracy: The data should accurately represent what it was intended to represent. How was it measured? Was care taken to account for known variables when being collected? Is there any way to confirm the accuracy of the data?
- Data accessibility: Is the data easily accessible to your model? Can the data be delivered as a local file, or is a modern distributed data processing and storage framework like Hadoop or Spark useful in your use-case?
- Data privacy and security: Is it traceable to any individuals? In the case of medical data, does it conform to privacy laws such as HIPAA? Is the data safeguarded from unauthorized access at all points of the project?
- Data richness and consistency: Are all aspects of the data relevant to the use case? Is the data rich enough? Are all important variables included?
- Data currency and relevance: Is the data current? Is the data still relevant to the problem at the current time?
- Data granularity: Does the data provide a fine enough level of detail for the model to work out the essence of the problem?

▼ What is Naive Bayes classification?

- It relies on the maximum a posterior probability (MAP) Bayesian decision rule.
- As with any model-based classifier, the model attempts to probabilistically determine the probability of features present in the class, using this information to determine the class of the object.
- When a Bayes classifier is provided a sample to classify, it then finds all other samples whose predictor variables, predicts the class they all belong to, and uses these to decide which class the new sample belongs to. If it cannot decide, the model fails it is classification.

▼ What is the event model of the classifier?

the assumptions made about the distribution of features

▼ What does Gaussian or “normal” distribution represent?

- real-valued random variables whose distributions are not known
- features are expressed in decimal form

▼ What is Bernoulli Variables Binomial Distribution?

- A binomial distribution represents a simple 0 or 1, true or false feature that is either present or not present.
- The binomial distribution is the sum of Bernoulli variables that are independent and identically distributed.

▼ What is Multinomial Distribution?

- Unlike binomial distribution, which assigns 0 or 1 dependent on the presence of a feature, multinomial distribution will count the number of times a feature is present and utilize that number.
- Multinomial works well with discrete features.

▼ What is Random Forest classification?

- A type of ensemble learning method, meaning more than one algorithm is used
- A very useful technique that is more versatile, has higher accuracy, and can utilize less data than Naive Bayes.

- It is also usable for both classification and regression problems.
- uses bagging

▼ Steps of Random Forest?

- The first step in Random Forest is pre-processing the data, allowing us to extrapolate.
- Then you generate a new dataset from the original, essentially hand-picking the subset of variables that you need to solve your problem.
- The data is then split and run through the decision trees.
  - This sub-sampling is an essential component of the Random Forest method.
  - The default number of variables to consider as candidates at each split point should be the square root of the total number of candidate inputs.

▼ Why use ensemble learning?

helps prevent overfitting and enables the model to work with less data

▼ What is bagging?

- the process of extrapolating data by generating new datasets based on the old data
- while this results in repeating variables, it improves the distribution by averaging low-bias and high-variance predictors

▼ What is Tensorboard?

- A graphical dashboard program developed by Tensorflow
- Despite being used heavily with Tensorflow, Tensorboard is also an extremely effective tool for any type of statistical model, just by plugging in the graphs.
- With Tensorboard, you can visualize vectors or collections of data as histograms.

▼ What is Jupyter Notebooks?

a tool for sharing and running live code in a web browser