MSA Exit Exam (Spring 2019)

Jeremiah Lowhorn

You are not allowed to directly copy answers from online resources. I will use Turnin to check plagiarism.

Short Answer Questions:

1. How is supervised learning different from unsupervised learning? How is regression different from classification?
   1. Supervised learning essentially means that you have a target variable that you are interested in either estimating a mean, median, probabilities, or classification whereas unsupervised learning has no target variable and the goal is to draw insights in the data, possibly creating clusters or principal components. Examples of supervised learning would include linear regression, logistic regression, or recursive partitioning methods (CART, random forest, GBM) amongst others. Unsupervised methods could include principal component analysis, k-means, and hierarchical clustering where the goal is to group the variables together instead of predict a target variable.
   2. Assuming you mean linear regression (least squares), which predicts a continuous variable estimating the mean, a classifier can estimate the probability of a binary variable (logistic regression), or use voting or recursive partitioning to determine the outcome variable that is typically binary (1-0, T-F, Y-N). Logistic regression has a similar form to linear regression of a continuous variable, but uses a sigmoid function instead of a linear function to predict the target, thus the algorithm predicts the log-odds of the binary target variable. Other methods of linear regression, such as LAD or quantile regression can estimate the median or the distributions theoretical quantiles respectively.
2. What does overfitting mean? Why does overfitting occur? How can we detect and deal with overfitting? Please name three approaches to dealing with over-fitting.
   1. Overfitting means that you have trained a model to fit your training set perfectly but the model may not generalize well to new observations. There are many methods to reduce overfitting which include regularization, ensembling, removing features, and sometimes training with more data. Regularization penalizes the variables that do not contribute to the model’s prediction via a lambda value via an L1 or L2 penalty. Ensembling combines multiple models together, typically through bagging, which can smooth out the predictions amongst all models. Removing features make models simpler so that they can generalize on new observations better. Finally, sometimes by adding data you can more accurately capture the signal of the data better which will reduce overfitting and help the model generalize better on new observations. Further, using a training, validation, and test set based on random sampling (without replacement) froma series of observations may be the most common example of a method to mitigate overfitting. Not asked for, but an important thing to consider for detecting this problem is cross-validation which will help the researcher to determine what the expected accuracy of the model is and help tune the hyperparameters.
3. What are the measures that can be used to compare different classification models? What are the measures that can used to compare different regression models?
   1. I typically focus on the area under the curve, but you can also look at the confusion matrix (tpr & fpr), precision & recall, and the model and class accuracies. For regression I typically use mean absolute error, but you can also use mean squared error which is what least squares is optimized to. For time series problems I always use the mean absolute percentage error or MAPE. For regression, AIC and BIC can also be used to assess the measure of fit with the complexity of the mode.
4. Please use your own language to describe the OLS (ordinary least square) linear regression algorithm.
   1. To me it is a linear algebra problem unless optimized with an LP or gradient descent. The method is solved by taking the inverse of the transpose of matrix A multiplied by itself. That product is then multiplied by the transpose of A to b, where b is a vector of the response variable and A is a matrix of observations. This method is however sensitive to outliers as indicated in its name (“least squares”), because we are minimizing the sum of squared residuals. A more robust method, least absolute deviation can be used to overcome this with the cost of a more difficult solution, typically an LP. At conclusion the idea is to minimize the sum of squared errors which can be accomplished by a variety of optimization techniques.
5. What’s the purpose of using the logistic function in the logistic regression algorithm?
   1. The purpose, as mentioned previously, is to calculate the log-odds of the response. The logistic, or sigmoid function, allows for the modeling of a non-linear probability distribution. Linear regression would not work in this case because the solution is not linear (0-1).
6. When working on a data mining project, we often split a dataset into three datasets: training, validation and test. Please describe the uses of each dataset.
   1. The training set is used to train the model and can be split into a variety of ratios depending on the size of the data, but is typically 70% of the original data set. The validation set is used to validate your model and tune the hyperparameters. It is easier for me to think of these two in a k-fold cross validation sense. The validation set typically comprises of 20% of the original data. The test set is the final data set that is used to test the tuned model. This set typically uses 10% of the original data. The difference between the training accuracy and validation accuracy can indicate overfitting if the training accuracy is significantly higher than validation.
7. Why is SVM a maximum margin classifier? Why do we often need to use kernel functions when we fit an SVM classifier? How is soft-margin SVM different from hard margin SVM?
   1. The SVM algorithm transforms the feature space so that the two classes are divided by a clear gap. The kernel function is the function that transforms the feature space to make the binary classes separable. Soft margin SVMs essentially mean the algorithm allows errors to be made when making the hyperplane to divide the classes. The benefit of using a soft margin SVM is that you will not overfit to outliers in the data thus your model will be better at generalizing to new observations. A smoothing parameter is tuned in a soft margin SVM to prevent overfitting to outliers. A hard margin SVM creates a hyperplane that tries to perfectly classify every example correctly, thus is less robust in the presence of outliers.
8. What does the term “random” in the random forest algorithm mean? What does “forest” mean? Is random forest a bagging or a boosting algorithm? Name the differences and similarities between bagging and boosting.
   1. The random forest algorithm is a bagging method that draws random samples of features and observations. This ensemble method builds decision trees and averages their predictions for the final ensemble. The random in the algorithm is due to the random sampling and the forest is because many trees are grown, typically 500 – 1000 based upon hardware limitations and time. Boosting is a method that takes a weak learner and improves the predictions of the tree overtime with gradient descent. Both boosting and bagging use weak learners but bagging averages the trees whereas boosting uses optimization to improve the classification of the trees over time. Boosting tries to reduce the error in predictions and hence reduces bias. Bagging on the other hand reduces variance because the model tries to reduce the complexity of the model.
9. Use your own language to describe the model-based collaborative filtering algorithm.
   1. Model-based collaborative filtering is a simplistic method that assumes that people like things similar to highly rated things they have reviewed in the past or that someone with similar interests may have liked. This approach starts with a sparse utility matrix of users, and a category of items that multiple users have rated. Model based collaborative filtering is named as such because it uses an algorithm to compute user preferences. There are different types of algorithms that can be uses such as K-Nearest Neighbors, matrix factorization, neural networks, or singular value decomposition.
10. What does betweenness centrality in a social network mean? What does closeness centrality mean?
    1. In a social network, betweenness centrality measures the number of times a node in the network graph lies on the shortest path between other nodes on the graph. Essentially it is a measure that shows us the bridges between nodes in a graph. On the other hand, closeness centrality scores each node on the graph based upon how close it is to all of the other nodes in the network. This measure will give us the shortest path between all nodes.
11. What does regularization mean? Why do we use regularization?
    1. Regularization assigns a penalty to each variable of the dependent data that, in the case of regression, shrinks the coefficient when it does not contribute to the prediction of the response. There are two major types, L1 (lasso) and L2 (ridge) that is both tuned by a parameter lambda. Regularization is used to prevent overfitting in the model and helps the model to generalize better on new observations.
12. Please use your own language to describe the PageRank algorithm
    1. PageRank works under the assumption that the more important websites are likely to have more links on their page to other websites. It then ranks the pages based upon the number of quality links on the webpage. The score of a page is the summation of all of the web pages links that point to the web page in question. Thus the more prestigious the links are on the page, the more prestigious the web page in question is.
13. Compare neural networks with logistic regression. How is neural network similar to logistic regression? How are they different?
    1. Logistic regression and neural networks, in a classification setting, both try to predict the probability of an event happening. Logistic regression will produce a linear boundary while a neural network does not have the assumption that the probability distribution follows a sigmoid function. A neural network can adapt to any probability distribution thus is more robust than logistic regression in situations where the sigmoid function cannot map to the input space. A single layer neural network is essentially equivalent to a logistic regression because the weights are not affected in the additional layers of the network and reinforcement learning does not take place.
14. In the context of decision tree learning, what does the measure “entropy” mean? (Briefly describe it. You don’t need to copy the formula). What does information gain mean? Why do we often need to prune a decision tree?
    1. Entropy essentially decides how a decision tree splits the data. It calculates the similarity of the partition and if the entropy is zero, then entropy is zero, if the sample is equally divided then it has entropy of one. Information gain is based upon a decrease in entropy after the data is partitioned. The decision tree algorithm tries to maximize information gain so that the tree splits the data with features that have the highest information gain first.
15. Describe the advantages and disadvantages (at least 2) of K-means clustering.
    1. K-means is computationally inefficient because the algorithm has to calculate a distance to the centroid on every single point in the data which can get very difficult with highly dimensional data. Furthermore, k-means is sensitive to data that is on different scales, so if the data is not scaled properly it can produce inaccurate results. Finally, with regard to disadvantages, it is difficult to know the number of optimal centroids so to add further computation time; you typically have to perform cross validation to select the optimum number of centroids. K-means does have advantages, namely it’s easy to implement and understand and makes no assumption about the data, other than it is on the same scale.
16. Compare SAS, R, and Python.
    1. SAS is commercial statistical software that is paid meaning you cannot utilize the software unless you have a subscription. For some companies this can be a burden, but for some like that the algorithms are tested and vetted by SAS professionals, not an open-source community. R is typically at the cutting edge of statistical and data visualization, but suffers from being single threaded in nature. R is in-memory statistical software which means that you can run into problems with larger datasets if you do not have sufficient memory to store the data. Python is a multi-purpose language who has strengths in machine learning and deep learning. Python is not bound to the same memory limitations as R and takes advantage of multi CPU architecture much easier. Python and R are both open-source software which means that you use the functions at your own risk and there is not a professional support group that can help you identify and solve problems with the software like SAS has.
17. Supposed you are working a text classification project, describe the steps involved in data preprocessing.
    1. Preprocessing involves tokenization, converting all characters to lowercase, removing non alphabetic characters, removing stop words, and stemming. Tokenization is the process of parsing text into words, n-grams, or other elements. Cleaning the text typically comes before tokenization and involves cleaning the text to make it useful for the task at hand; stop words include words such as ‘and’, ‘or’, and ‘but’ which typically do not provide any useful information. Stemming involves finding the root of words for example finding the root of a word with ‘ing’ or something that is plural.
18. Please explain what TFIDF means.
    1. TF-IDF is the inverse document frequency. It essentially assigns a weight to words that appear most frequently in the corpus being analyzed. The TF-IDF value increases as the number of times a word appears in a document and is offset by the number of documents that contain the word in the corpus. It is a method for finding the relative importance of a word in the context of the whole text. Those that appear less in the corpus as a whole will have a higher weight and be more important.
19. Please use your own language to describe PCA for variable reduction.
    1. Principal component analysis works by grouping variables together into primary components that contribute to the explanation of the variance of the feature space. This technique is useful with large feature spaces because it can find the groups of variables that matter the most in the feature space thus reducing the complexity of any subsequent models without using regularization.
20. Please describe a scenario where an OLTP system should be used and a scenario where a data warehouse should be used. Please name at least 4 differences between OLTP systems and data warehouses.
    1. Examples of OLTP systems include online data stores like online banking systems or text messaging data stores. Data warehouses can contain the historical data from an OLTP system and are generally engineered in star or snowflake schemas.
    2. OLTP systems are designed for highly transactional tasks and high concurrency. Data warehouses on the other hand typically operate with fewer transactions and lower concurrency. OLTP systems are primarily designed for modifying the data: update, delete, and insert statements. Data warehouses are typically used to query large amounts of data. OLTP systems typically have a very high response time whereas data warehouses are a bit slower as they store much more data. Finally, OLTP systems are designed for real time applications whereas data warehouses are designed for storage and analysis.
21. We often use star or snow-flake schemas in data warehouse design. What are the differences between a star schema and a snow-flake schema? What are the major components of a star schema?
    1. A star schema is characterized by a simple table structure that resembles a star. There is typically one fact table and multiple dimension tables that can be joined to the fact table with primary keys. Each dimension in this schema has only one dimension table. Furthermore, star schemas are not normalized, meaning that there are no lookup tables for additional attributes outside of the dimension table. Snowflake schemas build on the start schema and allow additional tables for normalization. This means that you can have lookup tables outside of the dimension tables. These lookup tables provide additional meta data than can describe the IDs within the dimension tables.
22. List the functions (at least 5) performed by OLAP. How is Slice operation different from dice operation?
    1. Functions of OLAP include: roll up, pivot, drill-down, slice, and dice. Roll up essentially means to aggregate the data whereas the drill-down is essentially the opposite. Pivot allows you to rotate the axes of the data to provide a different representation of the data. Slice selects one dimension which can create a new cube. Dice is different in that you select two or more dimensions to create a new cube.
23. What is a Namenode in a Hadoop system? What is a Datanode?
    1. The Namenode is essentially the centerpiece of HDFS or the master node. The name node only stores the metadata in HDFS which contains the directory of all the files in the system and tracks the files across the cluster. The data is not stored in the Namenode, but instead is stored in the Datanodes. The Datanodes primary function in HFDS is to store all of the data. The Namenode and Datanode are in constant communication. The Namenode will create copies of the data on multiple Datanodes so that if one Datanode goes down, the data is not lost entirely. Datanodes typically will have more hard disk space than Namenodes because they store the actual data.
24. What are the different phases of map-reduce? Briefly describe each of them.
    1. Map-reduce, as it is named, has two phases: map and reduce. The map phase is the initial phase in the job that maps a function to the data on the different Datanodes. In the reduce phase the outputs of the map phase are aggregated and combined into a single output to summarize the dataset with the aggregation designated in the map phase.
25. How is Apache Spark different from Hadoop Map-reduce?
    1. The major difference is that Spark is in memory while map-reduce has to read and write to disk thus the speed of Spark can be up to 100 times faster than map-reduce. This speed can also be of consequence because only so much data can be loaded onto the nodes of the HDFS cluster memory whereas there is far more disk space to perform a map-reduce job. Spark can be used for ETL processing very efficiently as long as the job does not exceed the memory limitations on the cluster. Spark can also be used to do more efficient joins, machine learning, graph processing, and iterative processing at much faster rates than map-reduce.