MSA Exit Exam (Summer 2020)

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?

Supervised learning takes place when using a labeled data set. The labels of the rows correspond to the target variable for machine learning purposes and the remainder of the data corresponds to features, or variables used to predict the labels. Unsupervised learning occurs when labeled data is either unavailable or not appropriate for the question at hand. In this case, the data is analyzed for similarities and patterns occurring within the data. The most common form of this is cluster analysis where the algorithm determines which data points are the most similar to each other with similarity measures being defined by the model builder.

Regression analysis is used to make predictions for a continuous target variable such as price or sales, where as classification is used to make predictions for discrete target variables. It does so by assigning probabilities to a given observation corresponding to the probability that that observation falls into a particular class.

1. 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.

Overfitting is when a model experiences decreased test prediction accuracy because it has learned too much from the noise in the training data set. It occurs when a model is fit too closely to the training data. We can detect overfitting by splitting our data into a training and test set. We then train the model only on the training set and determine the accuracy of the model on both sets. If the training accuracy is substantially higher than the testing accuracy there is likely overfitting occurring.

Three ways to deal with overfitting include using k-fold cross validation to train the model, using regularization techniques such as lasso or ridge regression, and lastly performing ensemble learning such as using a random forest or bagging or bootstrapping the model.

1. 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?

The primary measures used to compare classification models are classification accuracy, precision, recall, and F1 Score. For regression models it is typical to use mean squared error (or root mean squared error), mean absolute error, and the R squared statistic.

1. Please use your own language to describe the OLS (ordinary least square) linear regression algorithm.

The OLS regression algorithm aims to minimize the sum of squares of the residuals in order to estimate the parameters for the model. It then uses the logistic function to assign a probability of class membership.

1. What’s the purpose of using the logistic function in the logistic regression algorithm?

The logistic function maps the real number *z*, which is generated by the linear combination of the model parameters and features, to the range 0 to 1. This mapping can be interpreted as the probability that an observation belongs to class 1.

1. 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.

The training data set is used for training the model. This will be the largest of the three sets as you want as much data to train your model as possible. This data is fed into the model at the early stages and the parameters are learned from this data.

The validation set is used on a set of trained models for fine tuning hyperparameters and for selecting the best model.

The test set is used for determining the final accuracy or performance of the selected and fine tuned model.

1. 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?

The goal of the SVM algorithm is to fit a hyperplane to the data that maximizes the distance between an observation near the decision boundary and the decision boundary itself. This distance is called the margin and therefore the goal is to maximize the margin, thus a maximum margin classifier. Kernel functions are useful because they allow the model represent complex nonlinear decision boundaries by projecting the data into higher dimensional spaces. Hard margin SVM forces the model to maximize the margin for all points and does not allow consideration for noise or outliers. This can lead to overfitting when two classes are not completely linearly separated or have noise. Soft margin SVM on the other hand allows for tolerance within the model for a certain number of misclassifications. The threshold for this tolerance is set as a hyperparameter.

1. 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.

In a random forest the term random refers to the random selection of features used at each split in each of the trees in the forest. The term forest refers to the fact that many (often 100 or more) shallow trees are trained with random selections of features. To obtain final classifications, the model takes the average of the classifications of all of the trees in the forest. Random forests are a bagging algorithm.

A bagging algorithm is one which averages the classifications of a collection of classifiers whereas a boosting algorithm consists of weighted voting of a collection of classifiers. This occurs by training the individual classifiers sequentially with each one benefiting from what is learned by the previous. The main similarity between bagging and boosting is that both use a collection of smaller, simple models and aggregate the classifications to determine final classifications.

1. Use your own language to describe the model-based collaborative filtering algorithm.

The model-based collaborative filtering algorithm uses a machine learning method to find embedded features. The algorithm begins by initializing small values for the user feature vectors (theta) and the item feature vectors (X). Next we use some form of gradient descent to minimize the objective function. In this step, we use the values of theta to update the Xs and then the new Xs to update the values of theta until the objective function is minimized. Finally, the dot product of user with parameter theta and a movie with learned (embedded) features X predicts the rating given by that user to the given movie.

1. What does betweenness centrality in a social network mean? What does closeness centrality mean?

Betweenness centrality is measured based upon the likelihood that a node is to occur on a randomly chosen shortest path between two randomly chosen nodes. A node with high betweenness centrality is very likely to occur on a randomly chosen shortest path between two randomly chosen nodes. Closeness centrality, on the other hand is measured based upon the distance between a node and all other nodes. A node with high closeness centrality has a lower total distance to all other nodes.

1. What does regularization mean? Why do we use regularization?

Regularization is the general term for several different methods used with regression that shrinks the estimates of the coefficients towards zero. This reduces the overall variance in the model and prevents the model from learning patterns that are too complex. Regularization is used to help prevent overfitting.

1. Please use your own language to describe the PageRank algorithm

The PageRank algorithm recursively defines the PageRank score of a webpage and is used to determine the importance of a page. Google uses this algorithm to rank search results. The algorithm determines the PageRank score of a page *i* by summing the page rank score of all pages pointing to page *i* with a link and dividing that number by the total number out-links from page *i* to other pages.

1. Compare neural networks with logistic regression. How is neural network similar to logistic regression? How are they different?

Linear regression and neural networks are similar in that both models consist of computing linear combinations of parameters (theta) and features (X\_i) with a bias of b. However, whereas a logistic regression model computes one linear combination and applies an activation function to output a probability of class membership, a neural network is built of many nodes with each node corresponding to a linear combination of parameters with features. Each node can be thought of as an individual logistic regression model where the features correspond to the outputs of the node that comes before it in the model. Depending on the structure of the NNET the data passes all the way through the model (all of the nodes) an activation function is used to calculate class membership probabilities and the loss is calculated. Backpropagation is then used to update the parameters of each node and this process is iterated to minimize the loss. While logistic regression similarly minimizes the loss calculated by a loss function the process of updating the parameters is much simpler than in a neural network due to the complex dependencies of the linear combinations in a neural network.

1. 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?

The entropy of a node is a measure of the randomness of the classes in the node. It is calculated using the probability of randomly selecting each class in a node. If there are two classes in the data and one class has very high probability of being randomly selected within one node and the other has very low probability of being selected than the node will have low entropy. This is intuitive because low entropy then means that a node consists of primarily one class and the ultimate goal of a decision tree is to separate the classes into their own nodes. On the other hand, if the probability of randomly selecting each class from a node is approximately the same, then the node has high entropy and is not particularly helpful in classification and therefore will not be a leaf node.

Information gain is a measure of the change in entropy between a parent node and the weighted average of each child node. Information gain is a summary that tells us whether splitting a tree on a particular feature will increase or decrease the overall entropy of the tree and by how much. It is used for deciding which features to split the tree on at each level.

1. Describe the advantages and disadvantages (at least 2) of K-means clustering.

One big disadvantage of K-means clustering is that it depends on random initialization. This means that you can get different outcomes from the model with different seeds. Another disadvantage is the requirement to specify the number of clusters you are looking for at the outset as this may not always be available in a real-world setting. Some of the biggest advantages to K-means clustering are that it is relatively easy to understand and implement, it runs quickly, and is easily scalable to large datasets.

1. Compare SAS, R, and Python.

R, SAS, and Python, are all languages developed to aid with statistical modeling and machine learning. R and Python are both open source while SAS is a commercial product and thus is typically only found in business settings. R is the open source counterpart to SAS as the focus of both languages is statistical analysis. Python was originally developed as an open source scripting language but now supports libraries that allow complex machine learning processes. SAS is a procedural language, R is a functional language that has some capabilities for object oriented programming and Python is an object oriented language.

1. Supposed you are working a text classification project, describe the steps involved in data preprocessing.

The steps of pre-processing for text classification are tokenization, which is a process of breaking texts into words, phrases, symbols, or other meaningful elements called tokens (this can be done using several different types of tokenizers), converting all letters to lower case, removing numbers, removing stopwords, which are common words that do not contribute to the subject of a document (these include but are not limited to prepositions, conjunctions and similar words), removing punctuations, stemming, which reduces words to their stems so that words run, running, and runs would all be considered the same, and vectorization.

1. Please explain what TFIDF means.

TFIDF is short for term frequency inverse document frequency and is a method of weighting terms in a bag of words representation of a text document. TFIDF gives greater weight to words that occur frequently in a particular document or few, but reduces weight if that word occurs frequently in many documents. This way common words that do not differ between documents are less important than words that help distinguish documents by being important to a few documents but not many.

1. Please use your own language to describe PCA for variable reduction.

PCA is a process that projects higher dimensional data into lower dimensions. Unlike methods for feature selection, the new features produced by PCA are completely new and are based on the Eigen vectors of the SVD of the matrix of the original features. Because of this, the new PCA features are completely de-correlated, or orthogonal, and each successive principle component is in the direction of the maximal remaining variance. By this method a majority of the variance can be accounted for with relatively few features simplifying a model.

1. 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.

An example of a scenario where an OLTP system should be used is a library that needs a system to record which books have been checked out by which library card holders. The system would also need to be able to be queried in real time which books are not available on the shelf because they have been checked out, and then which books are again available because they have been checked back in. Because the primary use for this system will be transactional, because live updates are important, and because any querying needed will be primarily transaction retrieval an OLTP system would be best.

An example of a scenario where a data warehouse would be best would be a food company that wishes to understand how its products are performing based upon scanner data from national grocery chains. In this case, the food company would not be as interested in the specifics of each transaction but rather would need to be able to answer questions such as how many dollars in sales did brand x do during time y at store z. Because the food company is interested in summarized data that can easily answer complex business questions a data warehouse would be more appropriate for this case because it would provide the ability to easily answer these types of questions without the need for complex queries.

The main differences between an OLTP system and a data warehouse are that an OLTP system mostly performs update operations whereas a data warehouse mostly performs read operations. An OLTP system tends to perform many small transactions whereas a data warehouse processes long and complex queries. An OLTP system can manage MB-TB worth of data whereas a data warehouse can handle data storage in the range of GB-PB. OLTP systems store and present raw transactional data whereas data warehouses store summarized data. OLTP systems are typically used by clerical users or data entry types, whereas data warehouses are more useful for answering the questions posed by decision makers. Lastly, OLTP data is live and is always up to date nearly to the present moment, whereas a data warehouse will be updated on a schedule and thus will only be as current as the most recent update.

1. 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?

The main difference between a star schema and a snowflake schema is that the star schema is derived from the snowflake schema by denormalizing all of the tables belonging to a dimension which would otherwise be normalized in a snowflake schema. For a snowflake schema the each classification level gets its own table, but for a star schema the entire classification is expressed in just one table.

The major components of a star schema are the fact table in the center of the schema which contains the fact and measures. In addition to the fact table there is a table for each dimension and each dimension table is related to the fact table.

1. List the functions (at least 5) performed by OLAP. How is Slice operation different from dice operation?

OLAP functions include pivoting, rotation, roll-up, drill down, drill across, slice, and dice. The slice operation selects a single dimension from a data cube and returns a sub-cube, whereas the dice operation does the same but for two or more dimensions and typically for a select set of values of that dimension.

1. What is a Namenode in a Hadoop system? What is a Datanode?

In a Hadoop system the Namenode stores an image of the namespace and manages the HDFS metadata. It is also sometimes called the master node. On the other hand, the datanode is sometimes called the slave node and is the node where the actual data is stored. The data node is agnostic of the HDFS structure and stores each block of data in local files as directed by the Namenode.

1. What are the different phases of map-reduce? Briefly describe each of them.

The first phase of map reduce is the map phase. In this stage, the mapper takes a key/value pair as an input and returns an output of zero or more key value pairs. The mapper does this by mapping the input key value pairs into intermediary key value pairs that will be used by the next two steps. Each separate map task usually occurs on the datanode where the data is stored.

The second stage is the shuffle and sort stage. This shuffles and sorts the intermediate data from the mappers to help facilitate the most efficient use of resources for reducing.

The last stage is the reduce stage. This takes the shuffled and sorted output of the map stage and performs whatever function is required on the data to reduce it to the final output.

1. How is Apache Spark different from Hadoop Map-reduce?

One of the main differences between Apache Spark uses RAM and sometimes hard disc for storage whereas MapReduce only uses a hard disc. Additionally, instead of map and reduce operations, Spark’s operations are transformation and action. MapReduce’s execution is limited to only batch executions, whereas Spark can be used for batch executions, streaming, interactive, and iterative executions. Lastly, MapReduce is written in Java and whereas Spark has APIs for Scala, Java, Python, and R.