**DS/ML Course Review**

**DESCRIPTIVE STATISTICS**

Mean

Mode

Median

Percentile

Quartile

IQR

Imputation (beware if data is skewed, Median may be a better choice)

Variance

Standard Deviation

Range

Standardization (describe it, what’s the formula)

Why do we standardize features?

Once standardized what will the new feature values look like? (the range)

Normalization (describe it, what’s the formula)

Once normalized what will the new feature values look like? (the range)

**DATA SCIENCE**

What’s the data science process?

What is cleaning data?

What is transforming data?

What is one-hot encoding?

How can we determine the most important features? (Random Forest, observing each feature’s correlation with the target (if the target is continuous) or by visualizing the relationships).

What is feature engineering?

After our data is cleaned, what do we do next?

What is in X and what is in y?

What do we do after we’ve separated the data into X and y variables? (split the data into train/test sets)

Why do we split the data into Train and Test sets?

What’s a validation set? (the hold-out fold ink-fold cross validation)

**MACHINE LEARNING**

What is machine learning?

What does it mean to model the data?

What is unsupervised learning?

What is clustering?

What is k-means clustering?

What is supervised learning?

Describe some of the algorithms:

Regression -

Linear Regression

Classifiers -

Logistic Regression – linear classifier, decision boundary, binary problems, gives probabilities

How does it determine a specific classification? (using the threshold, 50% is the default)

Support Vector Machine (maximal margin classifier, kernels, non-linear classifier)

Decision Tree (non-linear classifier)

KNN (non-linear classifier)

What’s a Perceptron? (a single neuron in a neural network)

Neural Network (What’s another name for a neural network?: multi-layer perceptron)

How do we decide which algorithm is? (K-fold Cross Validation)

Once selected, how can we improve/refine the model? (tune the hyperparameters manually or using Grid Search.)

What are hyperparameters? (the parameters that we can modify in an ML algorithm)

Describe how model performance is determined. (using a Confusion Matrix is one way.)

What is a confusion matrix (provides performance metrics regarding a binary target.)

What is precision?

What is recall?

What is an F1 score? (geometric mean of precision and recall)

What do we do after we have selected and refined a model and we’re happy with its performance? (Re-train the model on the FULL training set first, and then test the model on the Test set.)

If we do well with the Test set, what do we do before using our model in production? (Train the model one more time using ALL of the data.)

**Whiteboarding**

Draw an SVM: what are the hyperparameters

Draw a Decision Tree (What are a Random Forest’s hyperparameters)

Draw KNN: What are the hyperparameters (how is distance calculated)

Draw NN: What are the hyperparameters