ML 4375 – Intro to Machine Learning – Summer 2018 – Mazidi – Exam 1 Review

General information:

* Test format: 50 questions, mostly multiple (multiple) choice, some short answer
* You will not have to produce R code but should recognize what R code does
  + 1 – Introduction to Machine Learning
  + 2 – Learning R
  + **3 – Linear Regression \***
  + **4 – Logistic Regression \*\***
  + **8 – kNN \*\*\***
  + **9 – Clustering \*\*\*\* hierarchal**

Clustering Notes

Unsupervised

Company might have lots of data

might not be labeled because it costs money to label

uses:

market segmentation, neighborhood analysis, clustering species

K-mean

Hierarchal

Combines observation, Denogram

Workflow

1.) Preprocess data (scaling)

2.) determine similarity

3.) cluster

4.) analyze

5.) repeat from 2 to 4, any step necessary

K means

1. Random assignment
2. Assign obs to closest centroid
3. Recalculate centroid
4. Repeat from 2 until convergence

* When would it be useless to scale the data? (when the units of measure are not the same, scaling doesn’t matter, ie feet to lbs) different units of measure
* Smaller cluster is better model

What is good clustering?

1. Homogenous
2. Features similar by a metric
   1. Or variance
3. Minimize sum of squares
4. Within ss
   1. Small ss = cluster is compact
5. Between ss
   1. How separated clusters are
   2. Ideally, clusters well separated
6. Play with number of obs in a cluster to adjust the clustering attributes

K means is example of Expectation-Maximization algorithim

1.) Expectation Step

2.) Maximization Step

Parameter – refers to the data

Hyperparater- refers to the model

k-Means

Pros

Cons

Have to find best k

Hierarchal

Procs

Don’t have to specify number of clusters

Uses dendogram

Can find a structure in the data that other algorithms might miss

Cons

Bogs down with lots of data

Greedy algorithim

Bottom up

many variations, the “bottom-up” version:

1. place each observation in its own cluster
2. calculate the distance between each cluster and every other cluster
3. combine the two closest clusters
4. repeat 2 and 3 until all clusters are merged into one big cluster

3 types of measurement:

1. single linkage – shortest distance between any points in the cluster; tends to create elongated clusters
2. complete linkage – longest distance; more compact but sensitive to outliers
3. average linkage – average distance

Understand the algorithms learned so far (linear regression, logistic regression, knn, clustering):

* a conceptual understanding of the algorithm
* familiarity with the math underpinning the algorithms
* loss and cost functions
  + this will be the short answer
  + be able to write a function in english
* tendencies towards bias or variance in different situations (for supervised learning)
  + Linear regression high bias low variance
* usage: classification, regression, both
* category: supervised or unsupervised learning
* which metrics are used for evaluation and why
* intepreting algorithm output
* interpreting summary(), anova(), residual plots

Demo in class

Which are quantitative? All variables except type. These variables are useful for Linear regression.

Which are qualitative? Type. This data is useful for Logistic regression.

What is a factor? R convention, internally stored as integer but they display as a label.

What do we seed? To ensure data is exactly reproducible during randomization.

What does sample do? Randomly extract data from a frame

How did we subset the data?

Why do we divide into train and test? Because some algorithms attempt to memorize the data. So having separate sets will allow use to evaluate the model while still using the same data

How do you use all predictors? Dot operator

How do you tell if a predictor is good? Low p value, the p value rejects the null hypothesis

What is the difference between r squared and rse? RSE is in units of y

Look at summary output by using summary. Quality = .007efixed.end + -1.4 … +.09

Understand the following machine learning terminology:

* Residuals
* Regression vs. classification
* Quantitative vs. qualitative (categorical) data
* Underfitting vs. overfitting
* Bias variance tradeoff
* loss function and cost function for different algorithms
* argmin, argmax
* metrics: mse, rmse, rse, R^2, accuracy
* causation vs. correlation
* confounding variable
* gradient descent and its alpha hyperparameter
* train and test sets; validation sets; cross validation
* instance == example == observation == row
* attribute == feature == predictor (not including target)
* factors in R and dummy variables
* curse of dimensionality
* Occam’s razor
* k-fold cross validation