CPSC 392– Introduction to Data Science

Exam I Review

* Exploratory Data Analysis
  + Why is it important before fitting any models?
  + What are the steps of conducting an EDA?
  + What are you looking for as a conclusion to EDA?
  + Be able to figure out which type of plot to use given the conditions
* Supervised Machine Learning
  + Use labels in the data to train and test a model
* Gradient Descent
  + It’s formula and how it works (explain using the formula)
  + How it is different for Linear and Logistic Regression models
* Linear Regression
  + What are the assumptions for the data?
  + What is the Ordinary Least Squares method?
  + What is the cost function for Linear Regression?
    - You don’t need to know how to compute the derivate of the cost function, but you should know the formula and what it is computing
  + What kind of data can Linear Regression be applied on successfully?
  + Linear Regression through gradient descent
    - Why is it done? What is it helping us do?
    - Steps of gradient descent for Linear Regression
  + R-squared value and other measures of Linear Regression
  + Multiple Linear Regression
* Logistic Regression
  + What are the assumptions for the data?
  + What the Logistic Regression does (looking at probabilities, not distances)
  + What is the log likelihood function? How is it used
    - Why is the log taken in the first place?
    - Again, you don’t need to know how find its derivative, but how it is applied for gradient ascent
  + Logistic Regression for binary and multiclass cases
* Validation
  + What is validation?
    - The simple kind (80/20 split)
  + What is k-fold cross validation?
    - How is it better than simple validation?
    - When will it be not viable to do cross validation?
* Confusion Matrix
  + What information it gives you
  + How you can compute the 3 different measures (recall, accuracy, precision)
    - Be able to understand what they mean for a model
* I can also give you a scenario where you are given a data set and you need to figure out which supervised machine learning technique to apply, or comment on if it is viable to use either one given a special condition.