# Assignment 8
# Linear Regression, K-Means, KNN and Decision trees

## Applied Data Science

### Instructions

This assignment has three (3) parts.

First, you will answer questions about the algorithms introduced in this module. Second, you will use each algorithm to model the low birth rate data set. Third, you will match each algorithm to a scenario for which that algorithm is optimal. Some of the questions will be explicitly required to be answered with pen and paper. For other questions, you may use R or any programming language you wish. You are encouraged to use preexisting implementations.

Submit your assignment as a PDF file with your name and email in the header, single-spaced and font no larger than 12pt. For the questions required to be solved with pen and paper, take pictures of your answers and insert them to the PDF file. Assignment will be graded on the correctness of the answers and overall quality.

### Part 1

#### Linear Regression

1. Answer the question: How can we determine if a given model is overfitting or underfitting the data? (Hint: Explain with reference to bias and variance)

2. Apply linear regression to linear_regerssion_data.txt, plot the residuals against the predicted values. There are 5 columns of data in linear_regerssion_data.txt; the first column is the target, and the other 4 columns are 4 features. The residuals are equal to target data minus your predicted values.

#### K-Means

1. Cluster the following set of points in 2-Dimensional space into 2 groups using the K-Means algorithm:

$$(2, 5)$$
$$(1, 5)$$
$$(22, 55)$$
$$(42, 12)$$
$$(15, 16)$$

Use the given distance metric to calculate the distance between the centroid and the points:

$$d((x_1, y_1), (x_2, y_2)) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Your are required to finish this question using pen and paper. Please include all the details in every iteration of K-Means algorithm.

2. Does K-Means always converge to the Global minima? Why/why not?

3. Explain two drawbacks of using the K-Means clustering algorithm.

**K-Nearest Neighbors**

The given data contains pairs of points and their corresponding class.

$$(2, 5), 1$$
$$(1, 5), 1$$
$$(48, 35), 2$$
$$(42, 12), 2$$

Use KNN algorithm, determine the class of the points:

$$(15, 16)$$
$$(30, 40)$$
$$(0, 0)$$

Report results for k=1, 2, 3.

Your are required to finish this question using pen and paper. Please include all the details for different k.

**Decision Trees**

1. Define Entropy and Information Gain.

2. For the given data, calculate the entropy for splitting based on the features A, B and C Label gives the class of the input. (You DO NOT have to split based on Information Gain. It is enough to calculate the entropy for splitting the root node based on each of the 3 variables)

| A | B | C | Label |
|---|---|---|---|
| a | a | a | 1 |
| b | b | a | 2 |
| a | a | b | 1 |
| b | b | a | 2 |

Your are required to finish this question using pen and paper. Be sure to include how you calculate the entropy.

## Part 2

Use each of the algorithms introduced in this module – regression, K-means, K nearest neighbors, and decision trees – to model the low birth rate dataset lbr-train.csv. For each algorithm, be sure to select which attribute is most relevant to predict (LOW or BWT) and which are relevant as features. You may wish to refer back to your data exploration to assist in this. Apply each model to the low birth rate dataset lbr-test.csv and describe its performance. Compare the performance of each model and discuss.

## Part 3

For each question, determine which algorithm is best. Discuss why you chose the algorithm you did and why you did not select the others.

1. Given dog breed data (height, length, weight, ear length, fur length), can we classify new dogs?
2. Given an unlabeled dataset of information(height, length, weight, ear length, fur length) about dogs, can we estimate four breed distinctions?
3. Given some information about a dog (height, length, weight, ear length, fur length, age, time at shelter) can we predict the amount someone will pay to adopt the dog?
4. Given some information (breed, color, age category, size category, house-trained, good with kids) about a dog at a shelter, can we predict if that dog will be taken home or not?