DSBA 6156
Assignment #2
Modules 4-7

1. A data analyst building a k-nearest neighbor model for a continuous prediction problem is considering appropriate values to use for k.

    a. Instead of experimenting with k values, the analyst thinks it might be best to just set k to the total number of instances in the training set. Do you think that the analyst is likely to get good results using this value for k? Why / why not?

    If you specified a number of k, you have a chance of losing sensitivity between variables as the locality is crucial when looking for specific nuances. Overgeneralization would also occur making any prediction be nearly the same rather to every new instance. It is also inefficient to use the whole dataset as it would be expensive to do it for every prediction.

    b. If the analyst was using a distance weighted KNN rather than a simple KNN for the predictions, would this have made the analyst's idea any more useful? Why / why not?

    Improved weighting causes the closer neighbors have a greater influence on the prediction than farther one creating a form of differentiation between instances based on their distance. However overgeneralization still persists the prediction would still be biased towards the central tendency of the target variable across the entire dataset, reducing the model's ability to capture local patterns thus still having inefficiency when calculating for all distances.

2. The table below lists a dataset that was used to create a nearest neighbor model that predicts whether it will be a good day to go surfing.

ID	Wave Size (ft)	Wave Period (secs)	Wind Speed (mph)	Good Surf
1	6	15	5	yes
2	1	6	9	no
3	7	10	4	yes
4	7	12	3	yes
5	2	2	10	no
6	10	2	20	no

Assuming that the model uses Euclidean distance to find the nearest neighbor, what prediction will the model return for each of the following query instances?

ID	Wave Size (ft)	Wave Period (secs)	Wind Speed (mph)	Good Surf
Q1	8	15	2	?
Q2	8	2	18	?
Q3	6	11	4	?

Use Excel to limit manual calcs in the solution:

ID	Wave Size (ft)	Wave Period (secs)	Wind Speed (mph)	Good Surf	Euc. Dist. to Q1	Euc. Dist. to Q2	Euc. Dist. to Q3
1	6	15	5	yes			
2	1	6	9	no			
3	7	10	4	yes			
4	7	12	3	yes			
5	2	2	10	no			
6	10	2	20	no			

ID	Wave Size	Wave Period	Wind Speed	Good Surf	Euclidean Distance to Q1	Euclidean Distance to Q2	Euclidean Distance to Q3
1	6	15	5	Yes	3.605551275	18.49324201	4.123105626
2	1	6	9	no	13.37908816	12.08304597	8.660254038
3	7	10	4	Yes	5.477225575	16.15549442	1.414213562
4	7	12	3	Yes	3.31662479	18.05547009	1.732050808
5	2	2	10	no	16.40121947	10	11.53256259
6	10	2	20	no	22.29349681	2.828427125	18.78829423

QID	Wave Size	Wave Period	Wind Speed	Good Surf
Q1	8	15	2	Yes
Q2	8	2	18	No
Q3	6	11	4	Yes

3. Linear Regression: A model is being built to predict the amount of oxygen that an astronaut consumes when performing five minutes of intense physical work. The descriptive features for the model will be the age of the astronaut and their average heart rate throughout the work. The regression model is:

$$\boldsymbol{OXYCON} = \boldsymbol{w}[\mathbf{0}] + \boldsymbol{w}[\mathbf{1}] \times \boldsymbol{AGE} + \boldsymbol{w}[\mathbf{2}] \times \boldsymbol{HEARTRATE}$$

The table that follows shows a historical dataset that has been collected for this task.

ID	OXYCON	AGE	HEARTRATE
1	37.99	41	138
2	47.34	42	153
3	44.38	37	151
4	28.17	46	133
5	27.07	48	126
6	37.85	44	145
7	44.72	43	158

8	36.42	46	143
9	31.21	37	138
10	54.85	38	158
11	39.84	43	143
12	30.83	43	138

Reference slides 40 - 47 in 7A lecture material, to solve (use Excel / other tool so you don’t have to do it by hand):

a. Assuming that the current weights in a multivariate linear regression model are w[0] = -59.50, w[1] = -0.15, and w[2] = 0.60, make a prediction for each training instance using this model.

ID	OxyCon	Age	HeartRate	Prediction
1	37.99	41	138	17.15
2	47.34	42	153	26
3	44.38	37	151	25.55
4	28.17	46	133	13.4
5	27.07	48	126	8.9
6	37.85	44	145	20.9
7	44.72	43	158	28.85
8	36.42	46	143	19.4
9	31.21	37	138	17.75
10	54.85	38	158	29.6
11	39.84	43	143	19.85
12	30.83	43	138	16.85

b. Extra Credit: Calculate the sum of squared errors for predictions generated in Part(a).

Error	Squared Error
20.84	434.3056
21.34	455.3956
18.83	354.5689
14.77	218.1529
18.17	330.1489
16.95	287.3025
15.87	251.8569
17.02	289.6804
13.46	181.1716
25.25	637.5625
19.99	399.6001
13.98	195.4404
Sum of Squared Errors:	4035.1863

c. Extra Credit: Assuming a learning rate of 0.000002, calculate the weights at the next iteration of the gradient descent algorithm.

ID	OxyCon	Age	HeartRate	Prediction	Error	Squared Error	(D,W[0])	(D,W[1])	(D,W[2])
1	37.99	41	138	17.15	20.84	434.3056	20.84	854.44	2875.92
2	47.34	42	153	26	21.34	455.3956	21.34	896.28	3265.02
3	44.38	37	151	25.55	18.83	354.5689	18.83	696.71	2843.33
4	28.17	46	133	13.4	14.77	218.1529	14.77	679.42	1964.41
5	27.07	48	126	8.9	18.17	330.1489	18.17	872.16	2289.42
6	37.85	44	145	20.9	16.95	287.3025	16.95	745.8	2457.75
7	44.72	43	158	28.85	15.87	251.8569	15.87	682.41	2507.46
8	36.42	46	143	19.4	17.02	289.6804	17.02	782.92	2433.86
9	31.21	37	138	17.75	13.46	181.1716	13.46	498.02	1857.48
10	54.85	38	158	29.6	25.25	637.5625	25.25	959.5	3989.5
11	39.84	43	143	19.85	19.99	399.6001	19.99	859.57	2858.57
12	30.83	43	138	16.85	13.98	195.4404	13.98	601.14	1929.24
					Sum of Squared Errors:	4035.1863	216.47	9128.37	31271.96
		Learning Rate:	0.000002						
						New Weights:	-59.49956706	-0.13174326	0.66254392

4. Naive-Bayes: The table below lists a dataset of books and whether or not they were purchased by an individual (i.e., the feature PURCHASED is the target feature in this domain).

ID	SECONDHAND	GENRE	COST	PURCHASED
1	false	romance	expensive	true
2	false	science	cheap	false
3	true	romance	cheap	true
4	false	science	cheap	true
5	false	science	expensive	false
6	true	romance	reasonable	false
7	true	literature	cheap	false
8	false	romance	reasonable	false
9	frue	science	cheap	false
10	true	literature	reasonable	true

   a. Assuming conditional independence between features given the target feature value, calculate the probability (rounded to four places of decimal) of each outcome (PURCHASED=true, and PURCHASED=false) for the following book:

      SECONDHAND=*false,* GENRE=*literature,* COST=*expensive*

      *P(True) = .5 * .25 * .25 *.4 = .0125*
      *P(False) = .5 * .1667 * .1667 * .6 = .0083*
      a = .0125 + .0083 = .0208

      P(True) = .0125/.0208 = .6010
      P(False) = ..0083/.0208 = .3990

   b. What prediction would a naive Bayes classifier return for the above book?\

      It returns outcome with the maximum probability with its prediction. For each instance the outcome PURCHASED=true is the MAP prediction and will be the outcome returned by a naive Bayesmodel

5. What happens to the bias (underfit) variance (overfit) tradeoff as we increase K in a KNN model? Is it more likely to underfit or overfit the training data?

   The more we increase k, were likely to underfit as the model becomes more stable and

less sensitive to the noise in the training data. When K is lsmaller, your more likely to overfit the making the model more complex.

6. Which model is more sensitive to feature scaling: KNN or Decision Tree? Why?

   KNN is more sensitive to feature scaling as it uses different measures to determine the nearest neighbor, if one feature has a large range of values then it can oversaturate in the calculation making values weigh towards that one more than others

7. You are analyzing a KNN and a Decision Tree for runtime performance. Which would likely be faster once deployed for inference? Explain in terms of ‘eager’ vs ‘lazy’ learners.

   Eager learners generalize from the training data before receiving data for prediction, meaning they are ready to make predictions instantly. Decision Trees are an example of eager learners.

   Lazy Learners do not build the model in advance, they wait to receive data before they start anything to start generalizing. K-Nearest Neighbors (KNN) is an example of a lazy learner because it doesn't learn a discriminative function from the training data but memorizes the dataset instead.

   Eager Learners are typivallyu faster because they already have the model created during trainfng then apply the model to make predictions.

8. Linear regression models can be adjusted to model non-linear relationships (between target and descriptive features) — True / False?

   True

   You can use polynomial terms, interaction terms, and logs to fit a much wider rance while the model itself stays linear, this is call polynomial regression.

9. Regardless of learning rate, gradient descent will always end up at the global minimum for linear regression models (learning rate just changes how quickly it will find the minimum) — True / False?

   True

   Gradient descent is guaranteed to converge to the global minimum, given enough time and assuming the learning rate is not too high.

   If the learning rate is too large, the algorithm might overshoot the minimum and fail to converge, or it might diverge entirely. On the other hand, if the learning rate is too small, convergence will be very slow, which can be computationally inefficient. This is so the convex function can converge on a specific point.

   learning rate needs to be chosen carefully to ensure convergence within a reasonable timeframe and to avoid divergence

10. Regularization techniques such as Ridge / Lasso Regression can be used to improve the performance of a model that is exhibiting poor training accuracy (<20%) — True / False?

    Ridge and Lasso work to combat overfitting when the model you run has a high training accuracy but low test accuracy. If the model is showing low accuracy in the train (<20%), then this is a sign of underfitting making it not likely to improve the performance. Regularization is better when the model is too noisy/complex.

11. Outliers have little to no effect on the coefficients estimated by linear regression models — True / False?

    False, outliers can have a significant impact on the model coefficients as it can cause a large amount of skewing by falling far from the central data points. If there are outliers, the regression line will try to take account for it and pull the line towards them.