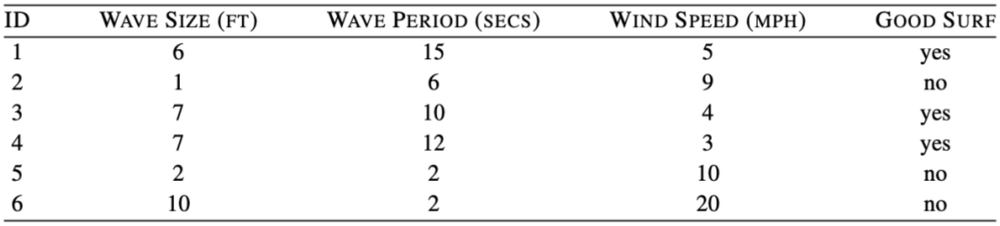
DSBA 6156
Assignment #2
Modules 4-7

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?

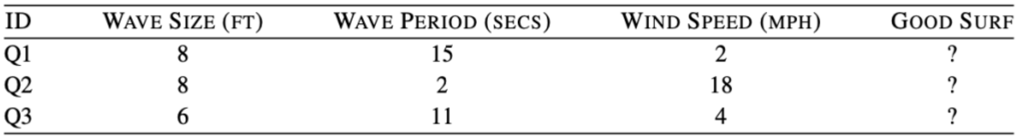
    b. If the analyst was using a distance weighted KNN rather than a simple KNN for the predictions, would this have made the analyst's idea any more useful? Why / why not?

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.



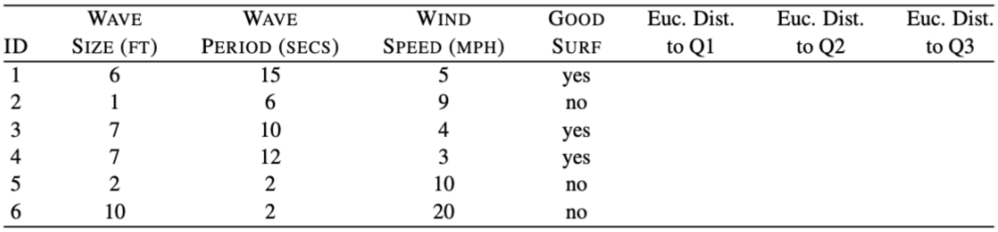
ID	Wave Size (ft)	Wave Period (secs)	Wind Speed (mph)	Good Surf
1	6	15	5	yes
2	1	6	9	no
3	7	10	4	yes
4	7	12	3	yes
5	2	2	10	no
6	10	2	20	no

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			

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

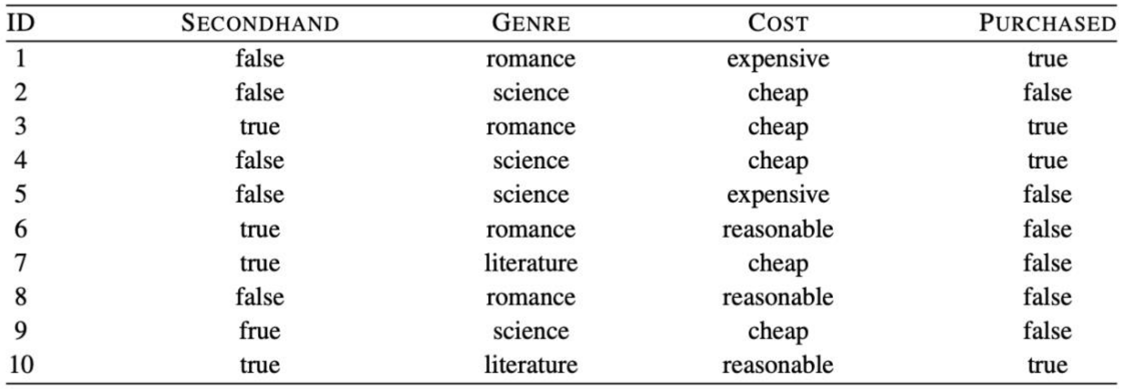
Reference slides 40 - 47 in 7A lecture material, to solve (use Excel / other tool so you don’t have to do it by hand):

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.

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

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

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*

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

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?

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

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.

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

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?

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?

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