1. Discuss the advantages and drawbacks of using a small value versus a large value for k.

|  |  |
| --- | --- |
| Small Value for k | |
| **Advantage** | **Disadvantage** |
| Ability to find pockets of homogenous behavior. | A small value of k means that noise will have a higher influence on the result. |
|  | Result becomes more localized. |
|  | high variance or over fitting problem |
|  |  |

|  |  |
| --- | --- |
| Large Value for k | |
| **Advantage** | **Disadvantage** |
| Classifier becomes less localized and smoother. | A large value make it computationally expensive and kinda defeats the basic philosophy behind KNN |
|  | Less able to find pckets of homogenous behavior in the data. |
|  | Higher bias |
|  |  |

1. The following dataset has two numeric predictor variables and one categorical target variable.
   1. Calculate the Euclidian distances for each possible row pair.

Suppose record #1 (t1, t2, …, tp) and record #2 (u1, u2, …, up)

Euclidean Distance =

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Row # | X1 | X2 | Response |  | Rows | Euclidian Distances |  | k=1 | k=3 | Naïve k=? |
| 1 | 3 | 2 | 1 |  | 1&2 | 1 |  | 0 | 1 | 1 |
| 2 | 4 | 2 | 0 |  | 1&3 | 4.123 |  | 1 | 1 | 1 |
| 3 | 4 | 6 | 1 |  | 1&4 | 2.236 |  | 1 | 1 | 1 |
| 4 | 2 | 4 | 1 |  | 2&3 | 4 |  | 1 | 1 | 1 |
|  |  |  |  |  | 2&4 | 2.8284 |  |  |  |  |
|  |  |  |  |  | 3&4 | 2.8284 |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |

* 1. Predict the Response for k=1, k=3, and the Naïve model for each row.

|  |  |  |
| --- | --- | --- |
| k=1 | k=3 | Naïve k=? |
| 0 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |

* 1. Find the mean squared error (MSE) and % Error for k=1, k=3, and the Naïve model. Note the MSE and % Error are the same measure calculated differently.

K=1

|  |  |  |
| --- | --- | --- |
|  | Predicted | |
| Actual | 0 | 1 |
| 0 | 0 | 1 |
| 1 | 1 | 2 |

Error rate=2/4 = 50 %

MSE = ((1-1)^2 + (1-0)^2 + (1-1)^2 + (1-1)^2) / N = 2/4 = 50%

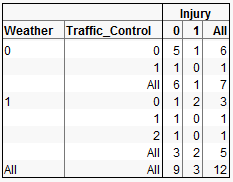
K =3

|  |  |  |
| --- | --- | --- |
|  | Predicted | |
| Actual | 0 | 1 |
| 0 | 0 | 1 |
| 1 | 0 | 3 |

Error rate : ¼ = 25 %

MSE = ((1-1)^2 + (1-0)^2 + (1-1)^2 + (1-1)^2) / N = 1/4 = 25%

1. Using the shown data summary, where we are predicting if an injury occurred (yes=1), if weather was an issue (1=yes), and the traffic control device (0=none, 1=signal, 2=other (sign, officer, etc.)) at the accident location, determine the following probabilities and note the predicted class for each combination (assume cutoff of .5).



* 1. Prob(Injury=1|Weather=1, Traf=0)

2/3=.66

* 1. Prob(Injury=1|Weather=1, Traf=2)

0/1=0

* 1. Prob(Injury=0|Weather=0, Traf=0)

5/6 = .83

1. Using the UniversitiesNoMissing.xlsx dataset. Try to predict if the university is public or private using the following variables. Partition the data and use 60% for training:

# appl. rec’d

# appl. Accepted

In-state tuition

Out-of-state tuition

Stud/fac ratio

* Use the k-nearest neighbor algorithm using K from 4 to 6 and compare the Error rates for the training and test data. Which is (are) the best K(s) to use?

|  |  |
| --- | --- |
| **4NN** | |
|  | **Error Rate** |
| Training | .94% |
| Testing | .35% |

|  |  |
| --- | --- |
| **5NN** | |
|  | **Error Rate** |
| Training | 1.17% |
| Testing | .35% |

|  |  |
| --- | --- |
| **6NN** | |
|  | **Error Rate** |
| Training | 1.17% |
| Testing | .35% |

We can see that error rate for testing data is same for all the models but error rate in training data is minimum for 4NN so 4NN seems to better option.

However, given the fact that even number of neighbors can create additional issues in terms of resolving ties so 5NN can also be considered as there is not significant difference in training error rate of 5NN and 4NN.

* Use KNN and have it automatically select k. What did it choose? What was the error rate?

It chooses 3 and error rate comes down to 0.

|  |  |
| --- | --- |
| **3NN** | |
|  | **Error Rate** |
| Training | .94% |
| Testing | 0 |

* Use the Naïve Bayes algorithm. What is the error rate for the training and test data?

|  |  |
| --- | --- |
| **Naïve Bayes** | |
|  | **Error Rate** |
| Training | 3.51% |
| Testing | 4.18% |

* Use the Naïve Rule. What is the error rate for the training and test data?

|  |  |
| --- | --- |
| **Naïve Rule** | |
|  | **Error Rate** |
| Training | 42.3% |
| Testing | 28.9% |