k Nearest Neighbor Question 0

1. **What is the difference between regression and classification?**
    a. Classification trees are used when the dataset must be divided into classes that belong to the response variable. These classes are usually “Yes” or “no” (two mutually exclusive classes). This is a predictive model that approximates a mapping function from input variables to identify discrete output variables).
    b. Regression trees are used when there is a continuous response variable (like if the response variable is something like the price of an object or the time). The main goal here is to estimate a mapping function based on input and output variables (where these are continuous).
2. **What is a confusion table? What does it help us understand about a model’s performance?**
    a. A confusion matrix is a performance measurement for machine learning classification problems where the output can be two or more classes. There are actual values on top and predicted values on the side. In a 2 x 2 matrix, there is true positive where you predicted positive and it’s true, true negative where you predicted it’s negative and it’s true, and then false positive and false negative.
    b. A confusion table can show you how many predictions are correct and incorrect per class which can help pinpoint which classes are getting confused
3. **What does the SSE quantify about a particular model?**
    a. The SSE is the sum of squared errors. This is a measure of accuracy of a simple linear regression model, which is calculated by summing the squared differences between the observed values and predicted values of the dependent variable. It essentially is telling us how good the model fits the data.
    b. You want a lower SSE (this means that the model fits the data well compared to a high SSE).
4. **What are overfitting and underfitting?**
    a. Underfitting is when the model can’t determine a meaningful relationship between the input and output data. This usually happens when there aren’t a lot of data points and the model hasn’t been trained for the right time. These models will give inaccurate results for the training data and test set.
    b. Overfit models give accurate results for training sets, but not for the test set. There is less bias but more increased variance. This happens when it learns too much with the training data and it fails to generalize.
5. **Why does splitting the data into training and testing sets, and choosing k by evaluating accuracy or SSE on the test set, improve model performance?**
    a. It is important to do this so that you can validate and evaluate machine learning models effectively. You can look at overall predictive accuracy by testing it on independent sets. This will help avoid overfitting to make sure it doesn’t generalize too much and fit the training data too much.
6. **With classification, we can report a class label as a prediction or probability distribution over class labels. Please explain the strengths and weaknesses of each approach.**
    a.