Homework 6-0

Please answer the following questions based on your reading in Chapter 4 of the “Doing Data Science” text, pages 93-112. Brief, but complete answers are best (10 points each, 40 points total).

1. In the example at the beginning of the chapter, both linear regression and KNN are not deemed suitable for the spam classification of email task. Give one reason why each modeling technique is not suitable.

Linear regression: The problem with using this technique is the target would have to be binary (0 for not spam and 1 for spam), but with a linear regression one would get a number instead. This also isn’t ideal since regression is aimed at modeling a continuous output and not a binary one.

KNN: With this technique one would use features corresponding to words and would have to define the value of the features as 0 or 1, depending on whether the word is present or not. Next one would have to define the emails according to when two emails are “near” each other based on which words they both contain. The problem is since there are 10,000 emails and 100,000 words the space one would have to work with contains too many dimensions. The main problem would be even the nearest neighbors would be really far away. This is called “the curse of dimensionality” and would KNN a poor algorithm to use in this case.

1. On the top of page 101, the author claims the Bayes rule classifier for single words is “overfitting” the data. Briefly explain why this is an example of overfitting.

This an example of overfitting because not all emails containing a specific word are spam. If important emails weren’t meant to be spam, but contained that word it would be marked as spam. It would be overfit by marking unnecessary emails as spam.

1. Explain why the naïve Bayes classifier is called naive.

This is called naïve since this classifier makes assumptions that may or may not turn out to be correct. It assumes that all attributes are independent of each other and is called naïve because of the assumption.

1. Explain in your own words the terms “prior” and “posterior” probability.

The prior probability is a probability that is assigned before any relevant evidence is examined. A posterior probability is a hypothesis that is true from making calculations on relevant evidence or observations.