COGS109
Homework 1
Gustav Sto. Tomas
A15358078

1. a) **Predictive modeling**: Given a dataset consisting of emails that are already classified as either ‘spam’ or ‘not-spam’ (e.g. the Enron email dataset), we could train a model to classify a new dataset of unseen emails depending on the similarity of words contained therein to those entailed in the training set. To do so we could use bag-of-words models and *n*-grams.

b) **Inference**: This concerns situations in which we want to visualize the correlation and potential causality between two variables and draw certain conclusions therefrom. E.g. we could see what the correlation is between temperature on certain days of the year and ice creams sold at a given vendor. (If we instead looked at the amount of ice creams sold on a given day, we could actually use this to predict whether it was a sunny day or not on that day and thus predict that in the future we will sell more ice cream on sunny days than not).

c) **Clustering**: For a dataset consisting of recorded natural speech we are trying to label the data points as belonging to different emotional states (e.g. happiness, sadness, fear, anger, surprise, disgust, and neutral, *or* simply as many or few distinct emotions as we can detect). By clustering the input data we may be able to recognize patterns and group together data points that are represented by vectors that appear similar according to some dimension.

2.1 a) For an extremely large set of observations, we would prefer a flexible model, as the risk of overfitting is lesser on larger datasets.

b) For a smaller set of observations and larger set of predictors, we would prefer an inflexible model. This is due to the risk of overfitting our model on the smaller amount f observations and their assumed larger variance.

c) If the relationship between p and n is highly non-linear, an inflexible model, e.g. *linear* regression, will not detect the *non-linearity* as accurately as a flexible one. We would therefore prefer a flexible method.

d) A flexible method would potentially overfit the model on the basis of high variance of error terms and we would not be able to use it efficiently on the test set. We would therefore prefer an inflexibe method.

2.7 a)
Yn1 = sqrt((0-0)^2+(3-0)^2+(0-0)^2); Red = 3
Yn2 = sqrt((2-0)^2+(0-0)^2+(0-0)^2); Red = 2
Yn3 = sqrt((0-0)^2+(1-0)^2+(3-0)^2); Red = 3.1623
Yn4 = sqrt((0-0)^2+(1-0)^2+(2-0)^2); Green = 2.2361
Yn5 = sqrt((-1-0)^2+(0-0)^2+(1-0)^2); Green = 1.4142
Yn6 = sqrt((1-0)^2+(1-0)^2+(1-0)^2); Red = 1.7321

b) Y = Green with K = 1, because we only consider the qualitative response of nearest neighbour (k=1), which is Obs 5 (1.4142 being the value nearest 0).

c) Y = Red with K = 3, because the majority of the k=3 nearest neighbours are classified as Red (Obs 6 and Obs 2), giving a p(>5).

d) A lower amount of kNNs would mean a more flexible model with low bias but high variance, which would correspond to the non-linearity of the Bayes decision boundary. However, it is possible that such a low amount of kNNs would also overfit to unwanted data points. A smaller number of kNNs can often give better training results but not yield better test results. In this particular problem, we can observe that k=1 gives us the classification “Green,” but with larger numbers of k and less flexibility the bias increases toward Red.