HW 03: Classification

due Thursday, Oct. 6 at 11:59pm

Exercise 1

When the number of features p is large, there tends to be a deterioration in the performance of KNN and other local approaches that perform prediction using only observations that are near the test observation for which a prediction must be made. Organizing/clustering/searching data often relies on detecting spaces where objects form groups with similar properties. When p is large, we will see that the objects (features) will seem very dissimilar. This phenomenon is known as the *curse of dimensionality*, and it ties into the fact that non-parametric approaches often perform poorly when p is large. We will now investigate this curse (which is common in many disciplines, e.g. genetics).

a) Suppose that we have a set of observations, each with measurements on $p = 1$ feature, X. We assume that X is uniformly (evenly) distributed on $[0, 1]$. Associated with each observation is a response value Y. Suppose that we wish to predict a test observation's response using only observations that are within 10% of the range of X closest to that test observation. For instance, in order to predict the response for a test observation with $X = 0.6$, we will use observations with associated X in the range $[0.55, 0.65]$. On average, what fraction of the available observations will we use to make the prediction?

b) Now suppose that we have a set of observations, each with measurements on $p = 2$ features, X_1 and X_2. We assume that (X_1, X_2) are uniformly distributed on $[0, 1] \times [0, 1]$. We wish to predict a test observation's response using only observations that are within 10% of the range of X_1 and within 10% of the range of X_2 closest to that test observation. For instance, in order to predict the response for a test observation with $X_1 = 0.6$ and $X_2 = 0.35$, we will use observations in the range $[0.55, 0.65]$ for X_1 and in the range $[0.3, 0.4]$ for X_2. On average, what fraction of the available observations will we use to make the prediction?

c) Now suppose that we have a set of observations on $p = 100$ features. Again the observations are uniformly distributed on each feature, and again each feature ranges in value from 0 to 1. We wish to predict a test observation's response using observations within the 10% of each feature's range that is closest to that test observation. What fraction of the available observations will we use to make the prediction?

d) Using your answers to parts (a)–(c), argue that a drawback of KNN when p is large is that there are very few training observations "near" any given test observation.

e) Now suppose that we wish to make a prediction for a test observation by creating a p-dimensional hypercube centered around the test observation that contains, on average, 10% of the training observations. For p = 1,2, and 100, what is the length of each side of the hypercube? Comment on your answer.

Note: A hypercube is a generalization of a cube to an arbitrary number of dimensions. When $p = 1$, a hypercube is simply a line segment, when $p = 2$ it is a square, and when $p = 100$ it is a 100-dimensional cube.

Exercise 2

Suppose that you wish to classify an observation $X \in \mathbb{R}$ into `toxic` and `nontoxic`. You fit a logistic regression model and find that

$$\widehat{\Pr}(Y = \text{toxic}|X = x) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0+\hat{\beta}_1 x}}$$

I also fit a logistic regression model to the same data, but I use the *softmax* formulation and find that

$$\widehat{\Pr}(Y = \text{toxic}|X = x) = \frac{e^{\hat{\alpha}_{0,\text{toxic}}+\hat{\alpha}_{1,\text{toxic}} x}}{e^{\hat{\alpha}_{0,\text{toxic}}+\hat{\alpha}_{1,\text{toxic}} x} + e^{\hat{\alpha}_{0,\text{nontoxic}}+\hat{\alpha}_{1,\text{nontoxic}} x}}$$

a) In your model, what are the log-odds of `toxic` vs `nontoxic`?

b) In my model, what are the log-odds of `toxic` vs `nontoxic`?

c) Suppose that in your model, $\hat{\beta}_0 = -2$ and $\hat{\beta}_1 = 3$. What are the coefficient estimates in my model? Be as specific as possible.

d) Now suppose that we fit the same two models on a different data set. This time, I estimate the coefficients $\hat{\alpha}_{0,\text{toxic}} = -1.5$, $\hat{\alpha}_{1,\text{toxic}} = 2$, $\hat{\alpha}_{0,\text{nontoxic}} = 1$, and $\hat{\alpha}_{1,\text{nontoxic}} = -1.25$. What are the coefficient estimates in your model?

e) Finally, suppose you apply both models from (d) to a data set with 2000 test observations. What fraction of the time do you expect the predicted class labels from your model to agree with those from my model? Explain your answer.

f) In the binary response case, would you prefer the usual logistic regression with the logit (also known as sigmoid) coding, or the softmax coding? Why?

Exercise 3

Recall that the Naive Bayes classifier assumes that the p features are independent, i.e. $f_k(x) = f_{k1}(x_1) \times f_{k2}(x_2) \times \cdots \times f_{kp}(x_p)$.

Assume in this problem that we have $p = 2$ features, X_1 and X_2, and the response has $K = 3$ classes. Further, assume that X_1 and X_2 are both qualitative. Where X_1 has three levels, and X_2 has two levels. An estimate for each of the f_{kj} is simply the proportion of training observations for the j-th predictor corresponding to each class k. Suppose we have $n = 200$ training observations, and observe the following data.

Note: the following is "aggregated" or summary data. That is, you are provided with the total counts within each class and predictor combination, as opposed to the individual, raw observations. For example, 22 observations had a $X_1 =$ `brown` and $Y =$ class 1. Similarly, 8 observation had $X_2 =$ `tree` and $Y =$ class 2.

```
##        Y
## X1       1  2  3
##   brown 22 16 33
##   red   11  6 23
##   white 29 23 37
```

```
##         Y
## X2        1  2  3
##   ground 48 37 71
##   tree   14  8 22
```

a) What are the number of observations falling into each class of Y? Call them n_1, n_2, n_3.

b) Using your answer from (a), what are $f_{kj}(x_j)$ for $k = 1, 2, 3$ and $j = 1, 2$?

c) Assume that $\hat{\pi}_1 = 0.3$, $\hat{\pi}_2 = 0.3$, and $\hat{\pi}_3 = 0.4$. Suppose that we wish to classify a new observation $x^* = (\text{red}, \text{ground})'$. Under the Naive Bayes classifier, calculate $\Pr(Y = k|X = x^*)$ for each k. Based on your answer, what would you classify this new observation as?

d) Now let's say that our new observation is $x^* = (\text{green}, \text{ground})'$. For this new observation, what would $\Pr(Y = k|X = x^*)$ be for each $k = 1, 2, 3$, and why? This is known as the *zero-frequency* problem.

Submission

Upload your assignment as a PDF file to Canvas. Please show all work!