Week 2 Homework

Samuel Cuthbertson SAMUEL.CUTHBERTSON@COLORADO.EDU

1. Exercise 10.1

1.1. Part A

We want to classify $P(B|x)$ where x = (0110). Our data looks like this:

Over 60 (A)					Under 60 (B)		
x_1	1	1	1	0	x_1	0	1
x_2	0	0	1	0	x_2	1	1
x_3	0	0	1	0	x_3	1	1
x_4	0	1	1	1	x_4	0	0

Given Bayes formula and the maximum likelihood, the probability $P(B|x)$ is given by this formula:

$$P(B|x) = \frac{P(x|B)P(B)}{P(x)}$$

$$P(B|x) = \frac{P(x|B)P(B)}{P(x|B)P(B) + P(x|A)P(A)}$$

Using the Naive Bayes assumption:

$$P(B|x) = \frac{\frac{1}{2} * 1 * 1 * 1 * \frac{2}{6}}{\frac{1}{2} * 1 * 1 * 1 * \frac{2}{6} + \frac{1}{4} * \frac{1}{4} * \frac{1}{4} * \frac{1}{4} * \frac{4}{6}} = \frac{64}{65}$$

$$P(B|x) = \boxed{0.98461538}$$

1.2. Part B

Since we now have 100 classes instead of 2, the probability of $P(OverSixty|x)$ is in the same for as in part A, just as a sum of all the possible classes and associated probabilities:

$$P(OverSixty|x) = \frac{\sum_{age=60}^{100} P(x|age)P(age)}{\sum_{age=0}^{100} P(x|age)P(age)}$$

2. Exercise 10.5

2.1. Part 1

$$p(c = 1) = \frac{1}{2}$$

Since there are 2 classes, the probability of one class is obviously $\frac{1}{2}$.

$$p(x_i = 1|c = 1) = \frac{\sum_{n=1}^{D} [x_n = 1, c = 1]}{\sum_{i=1}^{D} i}$$

$$p(x_i = 1|c = 0) = \frac{\sum_{n=1}^{D} [x_n = 1, c = 0]}{\sum_{i=1}^{D} i}$$

2.2. Part 2

Given the trained model $p(x, c)$, we would use Bayes Rule to form the classifier $p(c|x)$. Since $p(x, c)$ is equivalent to $p(c) * p(x|c)$ we can rewrite Bayes Rule as:

$$p(c|x) = \frac{p(x|c)p(c)}{p(x)} = \frac{p(x, c)}{p(x)}$$

2.3. Part 3

If a new email contains a word that isn't in any of the training data, that *should* trigger a false negative due to the effects of 0's on Bayesian classifiers. A spammer could exploit this with uncommon words or uncommon misspellings of common words in order to trigger that false negative.

3. Dirichlet Distribution Questions

3.1. Question A

My best understanding of why we would study a Dirichlet distribution is that it enables us to understand (or simply to model) the form that our distributions from different data sets will take, and to understand the spread or concentration of data.
In more concrete terms, my intuition thinks of natural language processing of reddit posts. The individual distributions of words is going to vary from subreddit to subreddit, but perhaps a dirichlet distribution would enable us to see the similarities or differences from one subreddit's data to another.

3.2. Question B

My ranking of entropy (from highest to lowest) is

$$Dir(< 1, 1, 1 >)$$

$$Dir(< .1, .1, .1 >)$$

$$Dir(< 2, 2, 2 >)$$

This ranking is based solely on the definition of entropy as the average number of bits needed to encode the data, and on my mental understanding of Morse code and the way that highly concentrated distributions need less bits (on average) of storage. Basically, I sorted based on the uniformity of the distributions.