Chris Piech
CS109

Fun and Review
June 7th, 2017

# Extra Practice Problems

Here are five exam questions which were either on previous exams that I have written or were candidates for this year's exam. We will go over as many answers as we can in class. We probably won't get through all of them, and if so, please use the extra problems as practice.

1. Spotify notices that users do not listen so songs with equal probability. Instead the probability that the ith most popular song is listened to is distributed as a Zipf random variable. A Zipf random variable is discrete and has PMF:

$$P(X = i) = \frac{\frac{1}{i}}{\sum_{n=1}^{N} \frac{1}{n}}$$

It is parameterized by N, the total number of songs.

a. What is the probability that a user listens to **the 10th most** popular song out of Spotify's 30 million songs?

b. If 1 billion songs are listened to on a given day, what is the probability that the most popular song is listened to more than 100 million times that day? Use an approximation but leave your answer as a formula that could be used to solve for the probability.

2. (25 points) A colleague has collected samples of heights of corgis that live on two different islands. The colleague collects 10 samples from both islands and observes that the island A has a sample mean that is 3 cm greater than island B. The colleague wants to make a scientific claim that corgis on island A are significantly taller than corgis on island B. You are skeptical. It is possible that heights are identically distributed across both islands and that the observed difference in means was a result of chance and a small sample size (the null hypothesis).

   To calculate the probability of the null hypothesis, find the probability that two sets of 10 numbers (E, F) which are IID samples from *the same* Gaussian with mean 0 and variance 20 have a difference in means greater than or equal to the one your friend observed.

Figure 1: Two corgis.

   a. How do you calculate the sample mean of set of 10 numbers?

   b. What is the probability that the difference between Es sample mean and Fs sample mean is greater than or equal to 3? Give an analytic solution.

3. (25 points) Consider the following functions:

```
int Intersection() {
   int wait = 0;
   bool cars = randomBool(0.5); //equally likely to be true or false
   if(cars) {
       wait = randomInt(0, 2);  //equally likely to be 0, 1 or 2
   }
   return wait;
}

int BikeSimulation() {
   int time = 0;
   for(int i = 0; i < 4; i++){
       time += Intersection();
   }
   return time;
}
```

Let W = the value returned by `Intersection()`.

a. What is E[W]? Give a numeric answer.

b. What is Var(W)? Give a numeric answer.

c. Let A be an indicator variable that is 1 if “cars” is true. What is Cov(A, W)?

d. Let T = the return value of `BikeSimulation()`. What is E[T]? You can express your answers in terms of E[W] and Var(W).

e. What is Var(T)? You can express your answers in terms of E[W] and Var(W).

4. (30 points) In a particular domain, we are able to observe three real-valued input variables $X_1$, $X_2$, and $X_3$ and want to predict a single binary output variable Y (which can have values 0 or 1). We know the functional forms for the input variables are all uniform distributions, namely: $X_1 \sim \text{Uni}(a_1, b_1)$, $X_2 \sim \text{Uni}(a_2, b_2)$, and $X_3 \sim \text{Uni}(a_3, b_3)$, but we are not given the values of the parameters $a_1, b_1, a_2, b_2, a_3$ or $b_3$. We are, however, given the following data set of 8 training instances:

| $X_1$ | $X_2$ | $X_3$ | Y |
|---|---|---|---|
| 0.1 | 0.8 | 0.4 | 0 |
| 0.7 | 0.6 | 0.1 | 0 |
| 0.3 | 0.7 | 0.2 | 0 |
| 0.4 | 0.4 | 0.6 | 0 |
| 0.8 | 0.2 | 0.5 | 1 |
| 0.5 | 0.7 | 0.8 | 1 |
| 0.9 | 0.4 | 0.7 | 1 |
| 0.6 | 0.6 | 0.4 | 1 |

a. (10 points) Which of the following values for $a_1$, and $b_1$ maximize the likelihood of the observed values of $X_1$:

i. $a_1 = 0.1$ and $b_1 = 0.9$ # The parameters are the min and max observed
ii. $a_1 = 0.0$ and $b_1 = 1.0$ # The parameters are 0 and 1
iii. $a_1 = 0.3$ and $b_1 = 0.7$ # The parameters are the 25th and 75th percentile

b. (10 points) Use Maximum Likelihood Estimators to estimate the parameters $a_1$, $b_1$, $a_2$, $b_2$, $a_3$ and $b_3$ in the case where Y = 0 as well as the case Y = 1. (I.e., estimate the distribution $P(X_i \mid Y)$ for $i = 1, 2$, and 3). Note that the parameter values for $a_1$, $b_1$, $a_2$, $b_2$, $a_3$ and $b_3$ may be different when Y = 0 versus when Y = 1.

c. (10 points) You are given the following 3 testing instances, numbered 1, 2 and 3. (Note that the testing instances do not have output variable Y specified).

| | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| test instance 1 | 0.5 | 0.6 | 0.4 |
| test instance 2 | 0.7 | 0.7 | 0.7 |
| test instance 3 | 0.5 | 0.4 | 0.3 |

Using the Naive Bayes assumption and your probability estimates from part (a), predict the output variable Y for *each* instance (you should have 3 predictions). Show how you derived the prediction by showing the computations you made for each test instance.

5. (25 points) You have a newly discovered document that dates back to the 1600s and you want to identify whether or not it was written by Shakespeare. Before doing any analysis, your prior belief is that the document is equally likely to be authored by Shakespeare or not by Shakespeare.

   To assist in your analysis you have *k* documents written by Shakespeare $D_1, \dots, D_k$. You also have m documents written by other authors $F_1, \dots, F_m$. In your answers you may use a function contains(X, W) which returns 1 if the document X contains the world W.

   a. Write an expression for the probability that a document contains the word "eyeball" given that it was written by Shakespeare.

   b. The document contains n unique words $W_1, \dots, W_n$. Write an expression for the probability that the document was written by Shakespeare given that it contains those n words? Use the Naive Bayes assumption.