

## Problem Set 6

*Handed Out: November 6<sup>th</sup>, 2014**Due: November 18<sup>th</sup>, 2014*

- Feel free to talk to other members of the class in doing the homework. I am more concerned that you learn how to solve the problem than that you demonstrate that you solved it entirely on your own. You should, however, write down your solution yourself. Please try to keep the solution brief and clear.
- Please use Piazza first if you have questions about the homework. Also feel free to send us e-mails and come to office hours.
- Please, no handwritten solutions. You will submit your solution manuscript as a single pdf file.
- The homework is due at 11:59 PM on the due date. We will be using Compass for collecting the homework assignments. Please submit your solution manuscript as a pdf file via Compass (<http://compass2g.illinois.edu>). Please do NOT hand in a hard copy of your write-up. Contact the TAs if you are having technical difficulties in submitting the assignment.
- No code is needed for any of these problems. You can do the calculations however you please. You need to turn in only the report. Please name your report as `<NetID>-hw6.pdf`.

**1. [Naïve Bayes and Learning Threshold Functions - 25 points]**

Consider the Boolean function  $f_{TH(3,7)}$ . This is a threshold function defined on the 7 dimensional Boolean cube as follows: given an instance  $x$ ,  $f_{TH(3,7)}(x) = 1$  if and only if 3 or more of  $x$ 's components are 1.

- (a) **[5 points]** Show that  $f_{TH(3,7)}$  has a linear decision surface over the 7 dimensional Boolean cube.
- (b) **[10 points]** Assume that you are given data sampled according to the uniform distribution over the Boolean cube  $\{0,1\}^7$  and labeled according to  $f_{TH(3,7)}$ . Use naïve Bayes to learn a hypothesis that predicts these labels. What is the hypothesis generated by the naïve Bayes algorithm? (You may assume that you have seen all the data required to get accurate estimates of the probabilities).
- (c) **[5 points]** Show that the final hypothesis in (b) does not represent this function.
- (d) **[5 points]** Are the naïve Bayes assumptions satisfied by  $f_{TH(3,7)}$ ? Justify.

**2. [Naïve Bayes over Multinomial Distribution - 35 points]**

In this question, we will look into training a naïve Bayes classifier with a model that uses a multinomial distribution to represent documents. Assume that all the documents are written in a language which has only three words  $a$ ,  $b$ , and  $c$ . All the documents have exactly  $n$  words (each word can be either  $a$ ,  $b$ , or  $c$ ). We are given a labeled document collection  $\{D_1, D_2, \dots, D_m\}$ . The label  $y_i$  of document  $D_i$  is 1 or 0, indicating whether  $D_i$  is “good” or “bad”.

This model uses the multinominal distribution in the following way: Given the  $i^{th}$  document  $D_i$ , we denote by  $a_i$  (respectively,  $b_i$ ,  $c_i$ ) the number of times that word  $a$

(respectively,  $b$ ,  $c$ ) appears in  $D_i$ . Therefore,  $a_i + b_i + c_i = |D_i| = n$ . We define

$$\Pr(D_i|y = 1) = \frac{n!}{a_i!b_i!c_i!} \alpha_1^{a_i} \beta_1^{b_i} \gamma_1^{c_i}$$

where  $\alpha_1$  (respectively,  $\beta_1$ ,  $\gamma_1$ ) is the probability that word  $a$  (respectively,  $b$ ,  $c$ ) appears in a “good” document. Therefore,  $\alpha_1 + \beta_1 + \gamma_1 = 1$ . Similarly,

$$\Pr(D_i|y = 0) = \frac{n!}{a_i!b_i!c_i!} \alpha_0^{a_i} \beta_0^{b_i} \gamma_0^{c_i}$$

where  $\alpha_0$  (respectively,  $\beta_0$ ,  $\gamma_0$ ) is the probability that word  $a$  (respectively,  $b$ ,  $c$ ) appears in a “bad” document. Therefore,  $\alpha_0 + \beta_0 + \gamma_0 = 1$ .

- (a) Write down the expression for the log likelihood of the document  $D_i$ ,  $\log \Pr(D_i, y_i)$ .
- (b) Derive the expression for the maximum likelihood estimates for parameters  $\alpha_1$ ,  $\beta_1$ ,  $\gamma_1$ ,  $\alpha_0$ ,  $\beta_0$ , and  $\gamma_0$ .

**Submission note:** You need not show the derivation of all six parameters separately. Some parameters are symmetric to others, and so, once you derive the expression for one, you can directly write down the expression for others.

**Grading note:** **10 points** for the derivation of one of the parameters, **5 points** each for the remaining five parameter expressions.

### 3. [Multivariate Poisson naïve Bayes - 30 points]

In this question, we assume that the output and input variables are related via a Poisson distribution<sup>1</sup>. You have two non-negative integer-valued numbers  $X_1$  and  $X_2$  as input-features and your class label  $Y$  can take two values,  $A$  and  $B$ . Conditioned on  $Y$ ,  $X_i$  for  $i = 1, 2$  follows a Poisson distribution with parameter  $\lambda$  specific to the class label of  $Y$ . That is

$$\Pr[X_i = x|Y = A] = \frac{e^{-\lambda_i^A} (\lambda_i^A)^x}{x!} \quad \text{and} \quad \Pr[X_i = x|Y = B] = \frac{e^{-\lambda_i^B} (\lambda_i^B)^x}{x!} \quad \text{for } i = 1, 2$$

Assume that the given data in Table 1 is generated by a Poisson naïve Bayes model. You will use this data to develop a naïve Bayes predictor over Poisson distribution.

- (a) **[10 points]** Compute the prior probabilities and parameter values i.e. fill out Table 2.
- (b) **[10 points]** Based on the parameter values from Table 2, compute

$$\frac{\Pr(X_1=2, X_2=3 | Y=A)}{\Pr(X_1=2, X_2=3 | Y=B)}$$

---

<sup>1</sup>[http://en.wikipedia.org/wiki/Poisson\\_distribution](http://en.wikipedia.org/wiki/Poisson_distribution)

$X_1$	$X_2$	$Y$
0	3	$A$
4	8	$A$
2	4	$A$
6	2	$B$
3	5	$B$
2	1	$B$
5	4	$B$

Table 1: Dataset for Poisson naïve Bayes

$\Pr(Y = A) =$	$\Pr(Y = B) =$
$\lambda_1^A =$	$\lambda_1^B =$
$\lambda_2^A =$	$\lambda_2^B =$

Table 2: Parameters for Poisson naïve Bayes

- (c) [**10 points**] (5 points each for two questions) Derive an algebraic expression for the Poisson naïve Bayes predictor for  $Y$  in terms of the parameters estimated from the data.

Use the parameters estimated from the data given in Table 1 to create a Poisson naïve Bayes classifier. What will the classifier predict as the value of  $Y$ , given the data point:  $X_1=2, X_2=3$ ?

4. [**Coin Toss - 10 points**]

In class, we discussed a scheme to generate a series of coin tosses as follows: For each element in the series, first a coin is tossed. If it comes up as a **T**, it is shown to the user. On the other hand, if the coin toss comes up as an **H**, then the coin is tossed the second time, and the outcome of this toss is shown to the user.

Assume that the probability of a coin toss coming up as an **H** is  $p$  (and hence, the probability of it coming up as a **T** is  $1 - p$ ). Suppose you see a sequence **TTHTHHTHTT** generated based on the scheme given above. What is the most likely value of  $p$ ? (Assume a Bernoulli model to compute probability of the coin toss sequence.)