CS446: Machine Learning **Fall 2015**

Problem Set 6

Handed Out: November 10^{th}*,* 2015 *Due: November* 19^{th}*,* 2015

- 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 $f_{TH(4,9)}$. This is a threshold function defined on the 9 dimensional Boolean cube as follows: given an instance x, $f_{TH(4,9)}(x) = 1$ if and only if 4 or more of x's components are 1.

   (a) [**5 points**] Show that $f_{TH(4,9)}$ has a linear decision surface over the 9 dimensional Boolean cube.

   (b) [**10 points**] Assume that you are given data sampled according to the uniform distribution over the Boolean cube $\{0,1\}^9$ and labeled according to $f_{TH(4,9)}$. 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 $f_{TH(4,9)}$? Justify.

2. [**Multivariate Poisson naïve Bayes - 30 points**] In this question, we consider the problem of classifying piazza posts (Y) into two categories: student posts (A), and instructor posts (B). For every post, we have two attributes: number of words (X_1), and number of mathematical symbols (X_2). We assume that each attribute (X_i, $i = 1, 2$) is related to a post category (A/B) via a Poisson distribution[1] with a particular mean (λ_i^A/λ_i^B). That is

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

[1] 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

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 the Poisson distribution.

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

Table 2: Parameters for Poisson naïve Bayes

(a) **[10 points]** Compute the prior probabilities and parameter values, i.e., fill out Table 2. [Hint: Use MLE to compute the λ's]

(b) **[10 points]** Based on the parameter values from Table 2, compute

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

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

(d) **[5 points]** 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: $X_1=2, X_2=3$?

3. **[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 $\{D_1, D_2, \ldots, D_m\}$. 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 α_1 (respectively, β_1, γ_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 α_0 (respectively, β_0, γ_0) is the probability that word a (respectively, b, c) appears in a "`bad`" document. Therefore, $\alpha_0 + \beta_0 + \gamma_0 = 1$.

(a) **[2 points]** What information do we lose when we represent documents using the aforementioned model?

(b) **[5 points]** Write down the expression for the log likelihood of the document D_i, $\log \Pr(D_i, y_i)$. Assume that the prior probability, $Pr(y_i = 1)$ is θ.

(c) **[28 points]** Derive the expression for the maximum likelihood estimates for parameters α_1, β_1, γ_1, α_0, β_0, and γ_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: 8 points for the derivation of one of the parameters, **4 points** each for the remaining five parameter expressions.

4. **[Dice Roll - 10 points]**

Consider a scheme to generate a series of numbers as follows: For each element in the series, first a dice is rolled. If it comes up as one of the numbers from 1 to 5, it is shown to the user. On the other hand, if the dice roll comes up as 6, then the dice is rolled for the second time, and the outcome of this roll is shown to the user.

Assume that the probability of a dice roll coming up as 6 is p. Also assume if a dice roll doesn't come up as 6, then the remaining numbers are equally likely. Suppose you see a sequence `3463661622` generated based on the scheme given above. What is the most likely value of p for this given sequence?