CS446: Machine Learning **Spring 2017**

Problem Set 6

Kaizhao Liang(kl2) *Handed In: April 17, 2017*

1. a. $f_{TH(4,9)} = sign(w^T \cdot x + \theta)$, where $w = [1,1,1,1,1,1,1,1,1]$ and $\theta = -3.9$

   b. $f_{TH(4,9)} = sign(P(v_j = 1, x) - P(v_j = 0, x)) = sign(\frac{1-e^{-\sum_i w_i x_i + b}}{1+e^{-\sum_i w_i x_i + b}})$, where:
   $w_i = \sum_i log(P(x_i = 1|v_j = 1)) - log(P(x_i = 0|v_j = 1)) - log(P(x_i = 1|v_j = 0)) + log(P(x_i = 0|v_j = 0))$
   $b = log(P(v_j = 0)) - log(P(v_j = 1)) + \sum_i log(P(x_i = 0|v_j = 1)) - \sum_i log(P(x_i = 0|v_j = 0))$
   $P(v_j = 0) = \frac{C_9^0 + C_9^1 + C_9^2 + C_9^3}{2^9} = \frac{65}{256}$
   $P(v_j = 1) = 1 - P(v_j = 0) = \frac{191}{256}$
   $P(x_i = 0|v_j = 1) = \frac{C_8^4 + C_8^5 + C_8^6 + C_8^7 + C_8^8}{2^9 - (C_9^0 + C_9^1 + C_9^2 + C_9^3)} = \frac{163}{382}$
   $P(x_i = 1|v_j = 1) = 1 - P(x_i = 0|v_j = 1) = \frac{219}{382}$
   $P(x_i = 0|v_j = 0) = \frac{C_8^0 + C_8^1 + C_8^2}{C_9^0 + C_9^1 + C_9^2 + C_9^3} = \frac{37}{130}$
   $P(x_i = 1|v_j = 0) = 1 - P(x_i = 0|v_j = 0) = \frac{93}{130}$
   $f_{TH(4,9)} = sign(P(v_j = 1, x) - P(v_j = 0, x)) = sign(\frac{191}{256} \cdot (\frac{163}{382})^k \cdot (\frac{219}{382})^{9-k} - \frac{65}{256} \cdot (\frac{37}{130})^k \cdot (\frac{93}{130})^{9-k})$, where k is the number of $x_i = 0$ in one sample.

   c. Rewrite the final hypothesis in b:
   $f_{TH(4,9)} = sign(log(\frac{P(v_j=1,x)}{P(v_j=0,x)})) = sign(w^T x + b)$, where $w = [log(\frac{8103}{15159})]^9, b = log(\frac{191}{65}) + 9log(\frac{28740}{35526}) + 9log(\frac{15159}{8103})$
   So the two decision boundaries are different, because of the bias. After normalizing the weight vector to all 1's, $b \approx -7.68$, which is less than the bias in the a..

   d. The naive Bayes assumption is not satisfied by the $f_{TH}(4,9)$, since the value of each dimension is not conditionally independent to the others, given the label.

2. a. Prior probabilities:

$Pr(Y = A) = \frac{3}{7}$	$Pr(Y = B) = \frac{4}{7}$
$\lambda_1^A = 2$	$\lambda_1^B = 4$
$\lambda_2^A = 5$	$\lambda_2^B = 3$

   b. $Pr(X_2 = 2, X_1 = 3|Y = A) = Pr(X_1 = 2|Y = A) \cdot Pr(X_2 = 3|Y = A) \cdot Pr(Y = A)$
   $Pr(X_1 = 2|Y = A) = \frac{e^{-2} \cdot 2^2}{2!} = 2 \cdot e^{-2}$
   $Pr(X_2 = 3|Y = A) = \frac{e^{-5} \cdot 5^3}{3!} = \frac{125}{6} \cdot e^{-5}$
   $Pr(Y = A) = \frac{3}{7}$
   $Pr(X_2 = 2, X_1 = 3|Y = B) = Pr(X_1 = 2|Y = B) \cdot Pr(X_2 = 3|Y = B) \cdot Pr(Y = B)$
   $Pr(X_1 = 2|Y = B) = \frac{e^{-4} \cdot 4^2}{2!} = 8e^{-4}$

$Pr(X_2 = 3|Y = B) = \frac{e^{-3} \cdot 3^3}{3!} = \frac{9}{2} e^{-3}$
$Pr(Y = B) = \frac{4}{7}$
$\frac{Pr(X_1=2, X_2=3|Y=A)}{Pr(X_1=2, X_2=3|Y=B)} = \frac{125}{144}$

c. $Pred(X_1, X_2) = sign(log(\frac{Pr(X_1,X_2|Y=A)}{Pr(X_1,X_2|Y=B)}))$
$\frac{Pr(X_1,X_2|A)}{Pr(X_1,X_2|B)} = \frac{\frac{e^{-(\lambda_1^A+\lambda_2^A)} \cdot (\lambda_1^A)^{X_1} (\lambda_2^A)^{X_2}}{X_1! \cdot X_2!} \cdot Pr(Y=A)}{\frac{e^{-(\lambda_1^B+\lambda_2^B)} \cdot (\lambda_1^B)^{X_1} (\lambda_2^B)^{X_2}}{X_1! \cdot X_2!} \cdot Pr(Y=B)}$
$Pred(X_1, X_2) = sign(\lambda_1^B + \lambda_2^B - \lambda_1^A - \lambda_2^A + X_1 log(\lambda_1^A) + X_2 log(\lambda_2^A) - X_1 log(\lambda_1^B) - X_2 log(\lambda_2^B) + log(Pr(Y = A)) - log(Pr(Y = B))$
When $Pred(X_1, X_2) = 1$,Y=A. Otherwise, Y=B.

d. The classifier constructed:
$Pred(X_1, X_2) = sign(-X_1 log(2) + X_2(log(5) - log(3)) + log(\frac{3}{4}))$
When $X_1 = 2, X_2 = 3$, $Pred(2, 3) = -1$. So, Y=B.

3. a. The order of the words.

b. $log(Pr(D_i, y_i)) = log(Pr(D_i|y_i) \cdot Pr(y_i)) = log(Pr(D_i|y_i)) + log(Pr(y_i))$
$log(Pr(D_i, y_i)) = y_i log(Pr(D_i|y_i = 1) \cdot Pr(y_i = 1)) + (1 - y_i) log(Pr(D_i|y_i = 0) \cdot Pr(y_i = 0))$
$log(Pr(D_i, y_i)) = y_i log(\theta) + (1-y_i) log(1-\theta) + y_i a_i log\alpha_1 + y_i b_i log(\beta_1) + y_i c_i log(\gamma_1) + (1 - y_i) a_i log(\alpha_0) + (1 - y_i) b_i log(\beta_0) + (1 - y_i) a_i log(\gamma_0) + log(\frac{n!}{a_i! b_i! c_i!})$

c. $log(PD) = \sum_i log(Pr(D_i, y_i))$
To find the optimization of the $\alpha_0, \beta_0, \gamma_0$ for $log(PD)$, construct function:
$L(\alpha_0, \beta_0, \gamma_0, \lambda) = f(\alpha_0, \beta_0, \gamma_0) - \lambda \cdot (g(\alpha_0, \beta_0, \gamma_0) - 1)$
where,
$f(\alpha_0, \beta_0, \gamma_0) = \sum_i log(Pr(D_i, y_i))$
$g(\alpha_0, \beta_0, \gamma_0) = \alpha_0 + \beta_0 + \gamma_0 = 1$
Solve the equations below:
$\sum_i (1 - y_i) a_i = \lambda \cdot \alpha_0$
$\sum_i (1 - y_i) b_i = \lambda \cdot \beta_0$
$\sum_i (1 - y_i) c_i = \lambda \cdot \gamma_0$
$\alpha_0 + \beta_0 + \gamma_0 = 1$
We have:
$\alpha_0 = \frac{\sum_i (1-y_i) a_i}{\lambda}$
$\beta_0 = \frac{\sum_i (1-y_i) b_i}{\lambda}$
$\gamma_0 = \frac{\sum_i (1-y_i) c_i}{\lambda}$
$\alpha_0 + \beta_0 + \gamma_0 = 1$
$\lambda = \sum_i (1 - y_i) a_i + \sum_i (1 - y_i) b_i + \sum_i (1 - y_i) c_i$
Then,
$\alpha_0 = \frac{\sum_i (1-y_i) a_i}{\sum_i (1-y_i) a_i + \sum_i (1-y_i) b_i + \sum_i (1-y_i) c_i}$
$\beta_0 = \frac{\sum_i (1-y_i) b_i}{\sum_i (1-y_i) a_i + \sum_i (1-y_i) b_i + \sum_i (1-y_i) c_i}$
$\gamma_0 = \frac{\sum_i (1-y_i) c_i}{\sum_i (1-y_i) a_i + \sum_i (1-y_i) b_i + \sum_i (1-y_i) c_i}$

By symmetry:
$\alpha_1 = \frac{\sum_i y_i a_i}{\sum_i y_i a_i + \sum_i y_i b_i + \sum_i y_i c_i}$
$\beta_1 = \frac{\sum_i y_i b_i}{\sum_i y_i a_i + \sum_i y_i b_i + \sum_i y_i c_i}$
$\gamma_1 = \frac{\sum_i y_i c_i}{\sum_i y_i a_i + \sum_i y_i b_i + \sum_i y_i c_i}$

4.

$$p_{max} = argmax_p P(h|D) = argmax_p(\frac{P(D|h) \cdot P(h)}{P(D)})$$

$$p_{max} = argmax_p(P(D|h)) \text{ Assuming that } P(h_i) = P(h_j) \ \forall \ h_i, h_j \in H$$
$$P(D|h) = (p^2)^4(\tfrac{1-p}{5})^6 \textstyle\sum_{i=0}^{6} C_6^i \cdot p^i$$
$$P(D|h) = \tfrac{(1-p^2)^6 \cdot p^8}{5^6}$$

To maximize $P(D|h)$, $\frac{\partial P(D|h)}{\partial p} = 0$
So $8 \cdot p^7 \cdot (1-p^2)^6 - 12 \cdot p^9 \cdot (1-p^2)^5 = 0$
$p^2 = \frac{2}{5}$
$p = \sqrt{\frac{2}{5}}$