CS-419m: Practice questions on Probabilistic Classifiers

1. An insurance company is trying to classify customers as "Risky" ($y = 1$) or "Not-risky" ($y = 2$) based on two attributes of a customer: x_1 denoting the number of "incidents" in the past ten years and x_2 denoting the type of the vehicle. Assume for each class y that the first attribute follows a Poisson distribution with parameter λ_y (Recall that for a Poisson distribution $P(x = k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$) and the second attribute follows a multinomial distribution with parameter θ_{yr} for $r \in (F, T, S)$. The table below denotes a sample training set.

	x_1	x_2	y
$\mathbf{x}^0$	0	F	2
$\mathbf{x}^1$	3	T	1
$\mathbf{x}^2$	1	T	2
$\mathbf{x}^3$	2	S	1
$\mathbf{x}^4$	2	F	2
$\mathbf{x}^5$	3	T	1
$\mathbf{x}^6$	4	S	1
$\mathbf{x}^7$	5	S	1

(a) Write the estimate of λ_y for each y using the training data above? [Show the derivation for one y] ..2

$\lambda_y^{MLE} = \texttt{argmax}_{\lambda_y} \sum_{i=1}^{N_y}(k_i log\lambda_y - \lambda_y) \Rightarrow \frac{1}{\lambda_y}\sum_{i=1}^{N_y} k_i - N_y = 0$

$\lambda_y = \frac{1}{N_y}\sum_{i=1}^{N_y} k_i$

$\lambda_1 = (3 + 2 + 3 + 4 + 5)/5$ and $\lambda_2 = (0 + 1 + 2)/3$

(b) Is the maximum likelihood objective concave in λ_y? Justify. ..3

$F = \sum_{i=1}^{N}(k_i log\lambda - \lambda + costant)$

$\nabla_y F = \frac{1}{\lambda}\sum_{i=1}^{N} k_i - N$

$\nabla_y^2 F = -\frac{1}{\lambda^2}\sum_{i=1}^{N} k_i \leq 0$ because # of incidents, $k_i \geq 0$

$\therefore$ F is concave in λ by Hessian test

(c) Write the estimate of all the θ_{yr}s for $r \in (F, T, S)$ and each y ..1

$\theta_{1F}^{MLE} = 0, \theta_{1T}^{MLE} = \frac{2}{5}, \theta_{1S}^{MLE} = \frac{3}{5},\ \theta_{2F}^{MLE} = \frac{2}{3}, \theta_{1T}^{MLE} = \frac{1}{3}, \theta_{1S}^{MLE} = 0$

(d) Assume the class labels follow a Bernoulli distribution with parameter p denoting the probability of class 1. What is the maximum likelihood estimate of p. ..1 $p^{MLE} = \frac{5}{8}$

2. Consider a binary classification problem ($y \in \{0, 1\}$) and two dimension data ($d = 2$) where both x_1 and x_2 are binary. Let $n_{00}^y, n_{10}^y, n_{01}^y, n_{11}^y$ denote the number of training instances in class y where $[x_1\ x_2]$ is [0 0], [1 0], [0 1], [1 1] respectively.

(a) Suppose we use a naive Bayes classifier where x_1 is assumed independent of x_2.

i. What is the maximum likelihood estimate of the two Bernoulli parameters in class y in terms of the counts $n_{00}^y, n_{10}^y, n_{01}^y, n_{11}^y$ ..1

$p^y(x_1 = a) = \frac{n_{a0}^y + n_{a1}^y}{n_{00}^y + n_{10}^y + n_{01}^y + n_{11}^y}$ $p^y(x_2 = a) = \frac{n_{0a}^y + n_{1a}^y}{n_{00}^y + n_{10}^y + n_{01}^y + n_{11}^y}$

ii. What is the maximum likelihood estimate of $\Pr(y)$ in terms of the counts $n_{00}^y, n_{10}^y, n_{01}^y, n_{11}^y$

$p(y = 0) = \frac{n_{00}^0 + n_{10}^0 + n_{01}^0 + n_{11}^0}{\sum_{a=\{0,1\}} n_{00}^a + n_{10}^a + n_{01}^a + n_{11}^a}$ ..1

iii. For what values of $n^y_{00}, n^y_{10}, n^y_{01}, n^y_{11}$ will the naive Bayes classifier incur the maximum percentage training error? Justify. Assume number of instances in both classes is the same. ..3 If $n^0_{00} = n^0_{11}, , n^0_{10} = n^0_{01} = 0$ and $n^1_{01} = n^1_{10}, , n^1_{00} = n^1_{11} = 0$, the estimated Bernoulli parameter will be 1/2 in both classes. The $\Pr(\mathbf{x}|y)$ value will also be the same. The error will be 50%.

(b) Now suppose we decide to use a model more powerful than naive Bayes. In each class, we model the probability that $\Pr([x_1\ x_2] = [a\ b]\ |y)$ with a parameter p^y_{ab}. Further we follow a LDA like parameter sharing to reduce the number of parameters by requiring that $p^0_{11} = p^1_{11} = p_{11}$. Clearly, $p^y_{00} + p^y_{10} + p^y_{01} + p_{11} = 1$. Write the log-likelihood function on the training data in terms of the counts n^y_{ab}, and parameters p^y_{ab}, $\Pr(y = 1), \Pr(y = 0)$. ..2

$\sum_i logP(y_i) * P(y_i|x) = \sum_i logP(y_i) + \sum_i logP(y_i|x)$
$= n^1 * logP(y = 1) + n^0 * logP(y = 0) + \sum_y (n^y_{00} * logp^y_{00} + n^y_{01} * logp^y_{01} + n^y_{10} * logp^y_{10}) + (n^0_{11} + n^1_{11}) * logP_{11}$

(c) Derive the maximum likelihood estimates of the $p^y_{00}, p^y_{10}, p^y_{01}$ parameters only in terms of p_{11} and the training counts. [Hint: Use Lagrangian multiplier to push the two constraints $p^y_{00} + p^y_{10} + p^y_{01} + p_{11} = 1$, one for each y to the objective.] ..3

We Maximize the objective created above (F^y_{ab}) using the lagrangian dual.
We thus have the objective as: Maximize $F^y_{ab} - \lambda_1 * (p^1_{00} + p^1_{01} + p^1_{10} + p^1_{11}) - \lambda_0 * (p^0_{00} + p^0_{01} + p^0_{10} + p^0_{11})$

$\frac{\partial F(P^y_{ab})}{\partial P^y_{ab}} = \frac{n^y_{ab}}{P^y_{ab}} - \lambda_y$

Using the constraints we get $\frac{n^y_{00}}{\lambda_y} + \frac{n^y_{01}}{\lambda_y} + \frac{n^y_{10}}{\lambda_y} + P_{11} = 1$ Thus $P^y_{ab} = \frac{n^y_{ab}}{(n^y - n^y_{11})} * (1 - P_{11})$

(d) Now derive the maximum likelihood estimate for p_{11}. ..1

Using the values obtained above in $F(P^y_{ab})$ and taking only the terms concerning P_{11} we get $log(1 - P_{11}) * \sum_y (n^y_{00} + n^y_{10} + n^y_{01}) + (n^0_{11} + n^1_{11}) * logP_{11}$

On differentiating this with P_{11} and solving for P_{11} we get $P_{11} = \frac{(n^0_{11} + n^1_{11})}{(n^0 + n^1)}$

..1