CS464
Introduction to Machine Learning

Homework 1
08.11.2018

Sabit Gökberk Karaca
21401862

Question 1.1

$$P(Loss \mid M_1) = \frac{675}{1000} \qquad P(Loss \mid M_2) = \frac{796}{1000}$$

$$P(M_2 \mid Loss) = \frac{P(Loss \mid M_2)\, P(M_2)}{P(Loss \mid M_2)\, P(M_2) + P(Loss \mid M_1)\, P(M_1)}$$

$$= \frac{\frac{796}{1000}\frac{85}{100}}{\frac{796}{1000}\frac{85}{100} + \frac{675}{100}\frac{15}{100}} = 0.870$$

Question 1.2

$$My\ losing\ probability\ on\ Machine\ 1 = 0.60$$
$$My\ friend's\ losing\ probability\ on\ Machine\ 1 = 0.75$$
$$On\ machine\ 1,\ my\ friend\ is\ more\ likely\ to\ lose$$

$$My\ losing\ probability\ on\ Machine\ 2 = 0.796$$
$$My\ friend's\ losing\ probability\ on\ Machine\ 2 = 0.800$$
$$On\ machine\ 2,\ my\ friend\ is\ more\ likely\ to\ lose$$

Question 1.3

$$My\ total\ wins = 252$$
$$My\ total\ loses = 888$$
$$win\ rate = 0.221$$

$$Friend's\ total\ wins = 43$$
$$Friend's\ total\ loses = 147$$
$$win\ rate = 0.226$$

$$My\ friend\ is\ more\ likely\ to\ win\ in\ total$$

Question 1.4

$$Me\ (Loss)\ \rightarrow\ Friend\ (Win)\ \rightarrow\ Me\ (Win)\ \rightarrow\ Friend\ (Loss)$$
$$P(Game\ starts\ with\ me)\ =\ \frac{1}{2}$$

$$\frac{60.25.40.75}{100.100.100.100}\ =\ 0.045$$

$$Friend\ (Loss)\ \rightarrow\ Me\ (Win)\ \rightarrow\ Friend\ (Win)\ \rightarrow\ Me\ (Loss)$$
$$P(Game\ starts\ with\ friend)\ =\ \frac{1}{2}$$

$$\frac{75.40.25.60}{100.100.100.100}\ =\ 0.045$$

$$P(Given\ order\ of\ loses\ and\ wins\ occur)\ =\ \frac{1}{2}*\frac{45}{1000}\ +\ \frac{1}{2}*\frac{45}{1000}\ =\ 0.045$$

Question 2.1

$$C\ =\ (b\ is\ not\ equal\ to\ 1\ or\ 6)\ and\ (r\ is\ not\ equal\ to\ 1\ or\ 2)$$
$$P(b = 5,\ r\ = 5\,|\,C)\ =\ P(b\ =\ 5\,|\,C)\ P(r\ =\ 5\,|\,C)$$
$$=\ \frac{1}{4}*\frac{1}{4}\ =\ \frac{1}{16}$$

Question 2.2

$$D\ = multiplication\ of\ the\ outcomes,\ b*r\ is\ an\ odd\ number$$
$$= b\ is\ odd\ and\ r\ is\ odd$$

$$b\ =\ \{1, 3, 5\},\ r\ =\ \{1, 3, 5\}$$
$$P(b = 5,\ c = 5\,|\,D)\ =\ P(b = 5\,|\,D)\ P(r\ =\ 5\,|\,D)$$
$$=\ \frac{1}{3}*\frac{1}{3}\ =\ \frac{1}{9}$$

Question 2.3

$$Given\ C,\ set\ of\ possible\ outcomes\ of\ b\ =\ \{2,3,4,5\}$$
$$Given\ D,\ set\ of\ possible\ outcomes\ of\ b\ =\ \{1,3,5\}$$

$$Given\ C,\ set\ of\ possible\ outcomes\ of\ r\ =\ \{3,4,5,6\}$$
$$Given\ D,\ set\ of\ possible\ outcomes\ of\ r\ =\ \{1,3,5\}$$

Question 3.1

$$X \sim Poisson(\lambda) \qquad P(X = x \mid \lambda)\ =\ \frac{\lambda^x e^{-\lambda}}{x!}$$

$$Likelihood\ function\ can\ be\ written\ as$$

$$\begin{aligned} L(\lambda\ ;\ X)\ &=\ \prod_{i=1}^{n} P(X = x_i \mid \lambda) \\ &=\ \left(e^{-\lambda n}\right) \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!} \end{aligned}$$

$$In\ order\ to\ find\ maximum\ likelihood\ estimator,$$
$$take\ the\ derivative\ and\ find\ the\ value\ that\ makes\ the\ equation\ equal\ to\ 0.$$

$$\begin{aligned} ln(L(\lambda\ ;\ X))\ &=\ ln\left(e^{-\lambda n}\lambda^{\sum_{i=1}^{n} x_i}\right)\ -\ ln\left(\prod_{i=1}^{n} x_i!\right) \\ \frac{d}{d\lambda} ln(L(\lambda\ ;\ X))\ &=\ \frac{d}{d\lambda} ln\left(e^{-\lambda n}\right)\ +\ \frac{d}{d\lambda} ln\left(\lambda^{\sum_{i=1}^{n} x_i}\right)\ +\ 0 \\ 0\ &=\ \frac{d}{d\lambda}(-\lambda n)\ +\ \frac{d}{d\lambda}\left(ln(\lambda)\sum_{i=1}^{n} x_i\right) \\ 0\ &=\ -n\ +\ \frac{d}{d\lambda} ln(\lambda).\ \sum_{i=1}^{n} x_i + ln(\lambda).\ \frac{d}{d\lambda}\left(\sum_{i=1}^{n}\right) \\ 0\ &=\ -n\ +\ \frac{\sum_{i=1}^{n} x_i}{\lambda} \\ \hat{\lambda}\ &=\ \frac{\sum_{i=1}^{n} x_i}{n} \end{aligned}$$

Question 3.2

$$\widehat{\lambda_{MAP}} = \underset{\lambda}{argmax}(L(\lambda; X). P(\lambda))$$

$$We\ can\ find\ the\ \lambda\ value\ that\ maximizes\ the\ function,$$
$$first\ we\ need\ to\ calculate\ the\ function$$

$$L(\lambda; X). P(\lambda) = \left(e^{-\lambda n}\right) \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!} .\ Pareto(\lambda|k, 1)$$
$$= \left(e^{-\lambda n}\right) \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!} .\ k\lambda^{-(k+1)}$$

$$Take\ the\ logarithm\ to\ convert\ production\ to\ summation$$
$$ln(L(\lambda; X). P(\lambda)) = ln\left(e^{-\lambda n}\lambda^{\sum_{i=1}^{n} x_i}\right) - ln\left(\prod_{i=1}^{n} x_i!\right) + ln\left(k\lambda^{-(k+1)}\right)$$

$$Take\ the\ derivative\ to\ find\ the\ maximum\ point$$
$$\frac{d}{d\lambda}ln(L(\lambda; X). P(\lambda)) = \frac{d}{d\lambda}ln\left(e^{-\lambda n}\lambda^{\sum_{i=1}^{n} x_i}\right) - \frac{d}{d\lambda}ln\left(\prod_{i=1}^{n} x_i!\right) + \frac{d}{d\lambda}ln\left(k\lambda^{-(k+1)}\right)$$
$$= \frac{d}{d\lambda}ln\left(e^{-\lambda n}\right) + \frac{d}{d\lambda}ln\left(\lambda^{\sum_{i=1}^{n} x_i}\right) - 0 + \frac{d}{d\lambda}(ln(k) - (k+1)ln(\lambda))$$
$$= -n + \frac{\sum_{i=1}^{n} x_i}{\lambda} - \frac{k+1}{\lambda}$$

$$\widehat{\lambda_{MAP}} = \frac{\sum_{i=1}^{n} x_i}{n} - \frac{k+1}{n}$$

$$Additionally,\ if\ we\ can\ find\ an\ interval\ for\ k\ such\ that\ \lambda \geqslant 1\ holds.$$
$$\widehat{\lambda_{MAP}} \geqslant 1$$
$$\frac{\sum_{i=1}^{n} x_i}{n} - \frac{k+1}{n} \geqslant 1$$
$$\sum_{i=1}^{n} x_i - k - 1 \geqslant n$$
$$k \leqslant \sum_{i=1}^{n} x_i - n - 1$$

Question 3.3

$$\widehat{\lambda_{MAP}} = \underset{\lambda}{argmax}(L(\lambda; X).P(\lambda))$$
$$where\ P(\lambda) \sim U(a, b),\ b > a$$

$$L(\lambda; X).P(\lambda) = \left(e^{-\lambda n}\right) \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!} .\ U(a, b)$$

$$Take\ the\ logarithm\ to\ convert\ production\ to\ summation$$
$$ln(L(\lambda; X).P(\lambda)) = ln\left(e^{-\lambda n}\lambda^{\sum_{i=1}^{n} x_i}\right) - ln\left(\prod_{i=1}^{n} x_i!\right) + ln(U(a, b))$$

$$Take\ the\ derivative\ to\ find\ the\ maximum\ point$$
$$\frac{d}{d\lambda}ln(L(\lambda; X).P(\lambda)) = \frac{d}{d\lambda}ln\left(e^{-\lambda n}\lambda^{\sum_{i=1}^{n} x_i}\right) - \frac{d}{d\lambda}ln\left(\prod_{i=1}^{n} x_i!\right)$$
$$= -n + \frac{\sum_{i=1}^{n} x_i}{\lambda}$$
$$\widehat{\lambda_{MAP}} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$Since\ \widehat{\lambda_{MLE}} = \frac{\sum_{i=1}^{n} x_i}{n}\ by\ its\ definition,\ we\ prove\ that$$
$$\widehat{\lambda_{MAP}} = \frac{\sum_{i=1}^{n} x_i}{n} = \widehat{\lambda_{MLE}}$$

Question 4.1

Since the denominator is the same for both classes, it does not affect the classification result. Therefore, we can safely ignore the denominator.

Question 4.2

There are 800 *space emails and* 800 *medical emails in the dataset. Therefore, the training dataset is balanced.*

Question 4.3

We need to estimate $2V + 1$ *parameters since the classification is binary.*
In that case, $V = 26507$
$\Longrightarrow 2V + 1 = 53015$ *parameters are estimated*

Question 4.4

My classifier has predicted almost all of the test instances to be medical. This happened because almost all of the likelihood values were $-inf$ and we made the assumption that in the case of tie, the test instance is predicted to be medical.

Using MLE is not a good idea in that case because we do not benefit from the "prior distribution" knowledge while making the predictions.

The test accuracy using MLE was 0.515

Question 4.5

When I used MAP with add – one smoothing technique instead of MLE in my classifier, the test accuracy is increased to 0.9675

Question 4.6

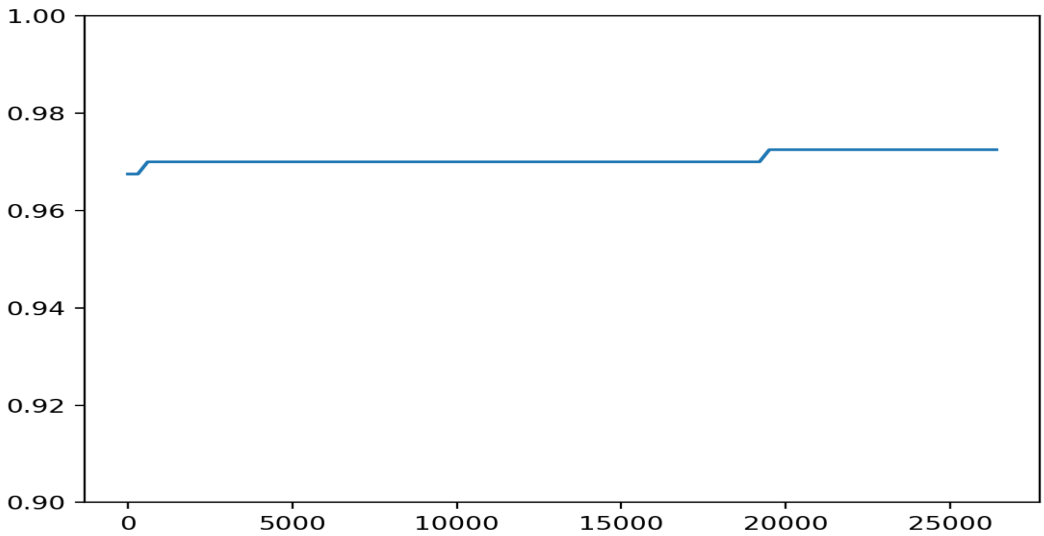
$$(25773,\ 0.22199134618517713)$$
$$(11999,\ 0.09428856018679171)$$
$$(13288,\ 0.0914173909185807)$$
$$(2848,\ 0.07988931992002964)$$
$$(14702,\ 0.07819961707710793)$$
$$(15990,\ 0.07680707708333484)$$
$$(614,\ 0.07473599079912649)$$
$$(3300,\ 0.07117925969237002)$$
$$(5749,\ 0.06856614476265091)$$
$$(12807,\ 0.0678314507329838)$$

Question 4.7

When we sort the features according to their mutual information scores and remove them in ascending order, we actually remove the irrelevant features, which increases the accuracy.

After some point, we start to remove excessive number of features, therefore accuracy should start to decrease.

In my experiment, I chose step size to be 300 *since the execution was taking too much time. Because of relatively high step size, only the increase is observed on accuracy graph.*



1.00
0.98
0.96
0.94
0.92
0.90
0
5000
10000
15000
20000
25000