HACETTEPE UNIVERSITY

COMPUTER ENGINEERING DEPARTMENT

BBM 406 - FUNDAMENTALS OF MACHINE LEARNING - 2022 FALL

Written Assignment - 2

Due Date : December 11, 2022

Student names:	*Student Number:*
Numan KAFADAR	b21946242
Umut GÜNGÖR	b21946198

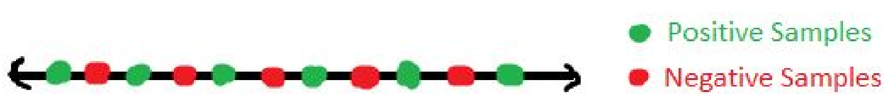
1 Question 1

Suppose we have training set with m=3 examples, which are (1,1), (2,2) and(3,3). Our hypothesis representation is $h(\theta = \theta_1 x)$ with parameter θ_1.The cost function is:
$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^i) - y^i)^2$
What is J(0)?

$$\begin{aligned} J(\theta) &= \frac{1}{2*3} \sum_{i=1}^{3} (h_\theta(x^i) - y^i)^2 \\ &= \frac{1}{2*3} \sum_{i=1}^{3} (0 - y^i)^2 \\ &= \frac{1}{6}[(0-1)^2 + (0-2)^2 + (0-3)^2] \\ &= \frac{7}{3} = 2.33 \end{aligned} \tag{1}$$

2 Question 2

Create a 1-Dimensional classication dataset in which the 1-Nearest Neighbors method always gives a leave-one out cross validation error value of 1 (In other words, the method can't guess correct class for any specific point in the dataset). State also a proper explanation about your reasoning.



As we can see in 1-Dimensional dataset above, in each iteration of leave-one-out cross validation, 1-KNN algorithm will always cause an wrong result since the closest class will always be the wrong class due to positioning. Every positive sample classified as negative since the closest class is always negative and nice versa.

3 Question 3

Assume that you have five students have registered to a class and the class have a midterm and the final exam. You have obtained a set of their marks on two exams, which is in the table below:

Student	Midterm Exam	Midterm exam(Squared)	Final Exam
$x^{(1)}$	87	7569	94
$x^{(2)}$	70	4900	72
$x^{(3)}$	92	8464	85
$x^{(4)}$	67	4489	76
$x^{(5)}$	45	2025	51

You plan to model which form's is $f_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$ for fitting the data above. The x_1 shows midterm exam score while x_2 shows square of the midterm score.Besides you plan to use feature scaling (using divide operation by the "max-min", or range of a feature) and mean normalization. What is the normalized value of the feature $x_2^{(2)}$?

$$MeanNormalization : x' = \frac{x_2^{(2)} - \overline{x_2}}{max(x_2) - min(x_2)}$$

$x_2^{(2)}$=4900
max(x_2)=8464
min(x_2)=2025
$\overline{x_2} = [7569 + 4900 + 8464 + 4489 + 2025]/5 = 5489.41$
$x_2^{(2)^{norm}} = \frac{4900-5489.41}{8464-2025}$ $x_2^{(2)^{norm}} = -0.092$

4 Question 4

Assume that you have three variables, which are A,B and C.

4.1 Suppose that you have the following informations P(C | A) = 0.7 and P(C | B) = 0.4. State that whether you can compute P(C | A,B) with the informations given previously or not. Besides show your solution if you can and explain the reason if you can not.

$P(C|A) = \frac{(P(C \cap A)}{P(A)} = 0.7$
$P(C|B) = \frac{(P(C \cap B)}{P(B)} = 0.4$
$P(C|A,B) = \frac{(P(C \cap A \cap B)}{P(A \cap B)} =?$

In this part , if C and A are independent , then P(C |A)=P(C) . Because when C and A are independent $P(C \cap A) = \frac{P(C)P(A)}{P(A)} = P(C)$. When we apply same logic for 2nd equation , we obtain P(C |B)=P(C). Therefore , in 1st equation , we get P(C)=0.7 and 2nd equation we get P(C)=0.4 .There is a contradiction. So ,both "C and A" and "C and B" cannot be independent. Hence we cannot compute P(C |A,B). Because we cannot calculate $P(C \cap A \cap B)$.

4.2 Suppose that besides two informations above,P(A) = 0.3 and P(B) = 0.5 informations are given. State that whether you can compute P(C | A,B) with the informations given previously or not. Besides show your solution if you can and explain the reason if you can not.

$P(C|A) = \frac{(P(C \cap A)}{P(A)} = 0.7$ $P(C|B) = \frac{(P(C \cap B)}{P(B)} = 0.4$
$P(A) = 0.3$ $P(B) = 0.5$
$P(C \cap A) = 0.21$ $P(C \cap B) = 0.2$
$P(C|A,B) = \frac{(P(C \cap A \cap B)}{P(A \cap B)} =?$

Like in part 4.1, if we assume "C and A" and "C and B" are independent , there will be a contradiction. Because if C and A are independent P(C)=0.21/0.3=0.7 and if C and B are independent P(C) =0.2 /0.5=0.4. So, both "C and A" and "C and B" cannot be independent. Thus we cannot compute P(C |A,B). Because we cannot calculate $P(C \cap A \cap B)$.

4.3 Finally assume that you have only informations, which are P(C | A) = 0.2 and P(A) = 0.3 and P(B) = 1. State that whether you can compute P(C | A,B) with the informations given previously or not. Besides show your solution if you can and explain the reason if you can not.

P(C | A) = 0.2   P(A)=0.3   P(B) = 1

$P(C|A,B) = \frac{(P(C \cap A \cap B)}{P(A \cap B)} = ?$

In this part , if we assume that A,B and C are mutually independent, then we have :

P(C $\cap A) = P(C)P(A) = 0.2 \quad => P(C) * 0.3 = 0.2 \quad => P(C) = 0.67$

P(A $\cap B) = P(A)P(B) = 0.3 * 1 = 0.3$

P(C $\cap A \cap B) = P(A)P(B)P(C) = 0.67 * 0.3 * 1 = 0.2$

$P(C|A,B) = \frac{(P(C \cap A \cap B)}{P(A \cap B)} = \frac{0.2}{0.3} = 0.67$

Hence ,we can compute P(C |A,B), and its value is 0.67.

5 Question 5

Assume that you have a data consisting of $x_1, x_2, \ldots, x_m$ where each x_i represent a single real value, which means you have m instances in data and each instance has an single real-valued attribute. Assume also that the given data has random uniform distribution between w and w. You are expected to find the maximum likelihood estimate of w with respect to the given data.

5.1 Specify a likelihood function F(w).

$$F(w) = \prod_{i=1}^{m} P_w(x_i) \quad where \quad P_w(x_i) = \begin{cases} \frac{1}{2w} & -w <= x_i <= w \\ 0 \end{cases}$$

5.2 Specify the maximum likelihood estimate for w. Consider your answer based on the likelihood function you state.

5.3 Assume that this time you are given a labelled data (x_i, y_i),where y_i is 1 or 0. Remember that a generative classifier will try to model and P(y) and P(x|y). Define an example dataset the generative classifier utilizing the model you defined above for each P(x|y) could not perform well on.

5.4 Remember that a discriminative classifier will try to model P(y|x). State that whether you can classify the labelled data given in previously part using such a discriminative classifier or not. If your answer is answer is yes, then please also show that what your suggested classifier looks like.

6 Question 6

x	y	z	C
1	0	1	1
1	1	1	1
0	1	1	0
1	1	0	0
1	0	1	0
0	0	0	1
0	0	0	1
0	0	1	0

Consider that you are given the dataset in the table above consisting of boolean variables x ,y and z and a single boolean output variable C. Suppose that the Naive Bayes classifier is going to be used.

P(C = 0)=4/8 P(C = 1)=4/8
P(x=0 | C=0)=2/4 P(x=1 | C=0)=2/4
P(y=0 | C=0)=2/4 P(y=1 | C=0)=2/4
P(z=0 | C=0)=1/4 P(z=1 | C=0)=3/4
P(x=0 | C=1)=2/4 P(x=1 | C=1)=2/4
P(y=0 | C=1)=3/4 P(y=1 | C=1)=1/4
P(z=0 | C=1)=2/4 P(z=1 | C=1)=2/4

6.1 Specify the value of (C = 1|x = 1, y = 1, z = 0) . Show your solution step by step

$$P(C=1|x=1,y=1,z=0) = \frac{P(x{=}1,y{=}1,z{=}0|C{=}1)P(C{=}1)}{P(x{=}1,y{=}1,z{=}0)P(C{=}0)+P(x{=}1,y{=}1,z{=}0)P(C{=}1)}$$

Since Naive Bayes assumes that events are independent, this equation will become;

$$= \frac{P(x{=}1|C{=}1)P(y{=}1|C{=}1)P(z{=}0|C{=}1)P(C{=}1)}{P(x{=}1|C{=}0)P(y{=}1|C{=}0)P(z{=}0|C{=}0)P(C{=}0)+P(x{=}1|C{=}1)P(y{=}1|C{=}1)P(z{=}0|C{=}1)P(C{=}1)}$$

$$= \frac{\frac{2}{4}*\frac{1}{4}*\frac{2}{4}*\frac{4}{8}}{\frac{2}{4}*\frac{2}{4}*\frac{1}{4}*\frac{4}{8}+\frac{2}{4}*\frac{1}{4}*\frac{2}{4}*\frac{4}{8}} = 0.5$$

6.2 Specify the value of (C = 0|x = 1, y = 1) . Show your solution step by step

$$P(C = 0|x = 1, y = 1) = \frac{P(x{=}1,y{=}1|C{=}0)P(C{=}0)}{P(x{=}1,y{=}1)P(C{=}0)+P(x{=}1,y{=}1)P(C{=}1)}$$

Since Naive Bayes assumes that events are independent, this equation will become;

$$= \frac{P(x{=}1|C{=}0)P(y{=}1|C{=}0)P(C{=}0)}{P(x{=}1|C{=}0)P(y{=}1|C{=}0)P(C{=}0)+P(x{=}1|C{=}1)P(y{=}1|C{=}1)P(C{=}1)}$$

$$= \frac{\frac{2}{4}*\frac{2}{4}*\frac{4}{8}}{\frac{2}{4}*\frac{2}{4}*\frac{4}{8}+\frac{2}{4}*\frac{1}{4}*\frac{4}{8}} = 0.66$$

Now suppose that the Joint Bayes classifier (In which you can't make assumption of independence of the x, y and z variables and consequently multiplying of conditional probilities with respect to these variables) is used for the options below:

6.3 Specify the value of (C = 1|x = 1, y = 1, z = 0) . Show your solution step by step

$$P(C = 1|x = 1, y = 1, z = 0) = \frac{P(x{=}1,y{=}1,z{=}0|C{=}1)P(C{=}1)}{P(x{=}1,y{=}1,z{=}0)P(C{=}0)+P(x{=}1,y{=}1,z{=}0)P(C{=}1)}$$

Since events are dependent in Joint Bayes classifier, the result will be;

$$= \frac{0*\frac{4}{8}}{\frac{1}{4}*\frac{4}{8}+0*\frac{4}{8}} = 0$$

6.4 Specify the value of (C = 0|x = 1, y = 1) . Show your solution step by step

$$P(C = 0|x = 1, y = 1) = \frac{P(x{=}1,y{=}1|C{=}0)P(C{=}0)}{P(x{=}1,y{=}1)P(C{=}0)+P(x{=}1,y{=}1)P(C{=}1)}$$

Since events are dependent in Joint Bayes classifier, the result will be;

$$= \frac{\frac{1}{4}*\frac{4}{8}}{\frac{1}{4}*\frac{4}{8}+\frac{1}{4}*\frac{4}{8}} = 0.5$$

7 Question 7

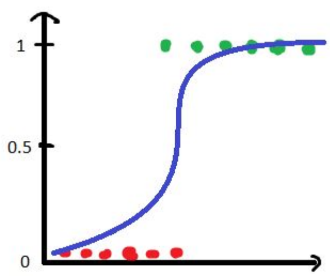
What happens if we change the hypothesis function of the Logistic Regression algorithm as follows? Can we use this function for the classification task? Explain your reasoning.

$h_\theta(x) = \sigma(\theta^T x) = \frac{1}{1+e^{\theta^T x}}$ **?**

Normally, hypothesis function of the logistic regression is

$h_\theta(x) = \sigma(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$

and it looks like below on the dataset.

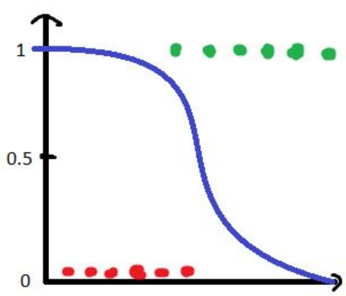


As we can see on the left model, this hypothesis function correctly classifies the given data. It's accuracy is high and it is a very valid model.

But if we change the hypothesis to

$h_\theta(x) = \sigma(\theta^T x) = \frac{1}{1+e^{\theta^T x}}$

then it will look like below on the dataset.



In this model , the hypothesis function does not fit the dataset .This will cause that all the instances will be classified wrongly since positive samples classified as negative and nice versa. It's accuracy is extremely low and it is a very bad model.