Module 3— Homework

Sadia Nasrin Tisha

11th September 2021

1 2.1

1. **2.17:**

   (a) Given that, the transformation from x to x^* is of square root. Mathematically,
   $x^* = \sqrt{x}$
   $(x^*)^2 = (\sqrt{x})^2$
   $(x*)^2 = x$
   That means an interval of (a, b) in x^* can be transformed into x by taking square of each term of the interval value. Therefore, the corresponding interval (A, B) in x will be (a^2, b^2).

   (b) Given that, the transformation from x to x^* is of square root. Mathematically,
   $x^* = \sqrt{x}$
   $(x^*)^2 = (\sqrt{x})^2$
   $(x*)^2 = x$
   It says that x^* and y are linearly related. Therefore, we can use linear equation to equate them.
   $y = m.x^* + c$
   $\Rightarrow y = m\sqrt{x} + c$
   $\Rightarrow y = m(x)^{1/2} + c$; which relates y to x.
   On the other hand, we can say that, in the interval(A, B), $y = x^2$ relates y to x.

2. **2.25:**

   (a) Given that, S is a set of points in Euclidean space, where the distance of each point in S to a point x.
   Here, If z is an arbitrary point of S, then the triangle inequality,
   $d(x, y) \leq d(x, z) + d(y, z)$.
   and $d(y, z) \geq d(x, y) - d(x, z)$.
   Another application of the triangle inequality,
   $d(x, z) \leq d(x, y) + d(y, z)$; where we need to show that, $d(y, z) \geq d(x, z) - d(x, y)$.
   If the lower bound of $d(y, z)$ obtained from one of these triangle inequalities is greater than ϵ, then $d(y, z)$ does not need to be calculated.
   Also, if the upper bound of $d(y, z)$ obtained from the triangle inequality $d(y, z) \leq d(y, x) + d(x, z)$ is less than or equal to ϵ, then $d(y, z)$ does not need to be calculated.
   So, in the triangle inequality we donot need to calculate $d(y, z)$.

   (b) If the distance between x and y go farther, then more calculations are needed for distance calculation. And if $x = y$, then no calculation needed for distance calculation between x and y.

   (c) Here,
   S is a set of points in Euclidean space, where the distance of each point in S to a point x, So, x

and y be the two points of S.
Let, x' and y' the points in S' that are closest to the two points x and y respectively. If x and y are as far from x' and y' as possible and as far or close to each other as possible, Then we can say that,
If $d(x', y') + 2\epsilon \leq \beta$, then $d(x, y) \leq \beta$.
And, If $d(x', y') - 2\epsilon \geq \beta$, then $d(x, y) \geq \beta$.

3. **3.11:**

    (a) Here, the dataset has 50 positives and 50 negatives. Let represent this dataset as $D = \{x_i\}_{i=1}^{100}$, where x_i is the ith data instance. In the leave-one-out method, the test error on every instance x_i is computed by applying a classification model trained on all data instances in D excluding x_i. Here, the cross validation is calculated based on mean squared error (MSE).
    If x_i is positive, then D_{-i} will be containing 49 positives and 50 negatives). The majority inducer classifier will thus assign x_i to the majority class, which is negative, and thus there will be an error.
    On the other hand, if x_i negative, then D_{-i} will contain 49 negatives and 50 positives, and the majority inducer will incorrectly assign x_i to be positive.
    That is how, the majority inducer would make an error on every data instance using leave-one-out. For this reason there is an error rate of 1.

    (b) According to 2-fold stratified cross-validation, If we divide the data set D into two folds where both folds have equal positives and negatives values, the majority inducer trained over any of the two folds will have same value. That is why, test instances are assigned to the default positive class. So the default class will be correct half of times so, 0.5 will be the error rate of majority inducer using 2-fold stratified cross-validation.

    (c) On this dataset, cross-validation provides a more reliable estimate of the generalization error rate of the majority inducer classifier. According to (b) the error rate of majority inducer using 2-fold stratified cross-validation will be 0.5. On the other hand, leave-one-out is quite susceptible to changes in the number of positive and negative instances in the training set, even by a single count, leading to a high error rate of 1 for the majority inducer.

4. **4.7:**

    (a) $P(A = 1|+) = \frac{3}{5} = 0.6$
    $P(A = 1|-) = \frac{2}{5} = 0.4$
    $P(A = 0|+) = \frac{2}{5} = 0.4$
    $P(A = 0|-) = \frac{3}{5} = 0.6$
    $P(B = 1|+) = \frac{1}{5} = 0.2$
    $P(B = 1|-) = \frac{2}{5} = 0.4$
    $P(B = 0|+) = \frac{4}{5} = 0.8$
    $P(B = 0|-) = \frac{3}{5} = 0.6$
    $P(C = 1|+) = \frac{4}{5} = 0.8$
    $P(C = 1|-) = \frac{5}{5} = 1$
    $P(C = 0|+) = \frac{1}{5} = 0.2$
    $P(C = 0|-) = \frac{4}{5} = 0$

    (b) Here, $P(A = 0, B = 1, C = 0) = Q$
    $P(+|A = 0, B = 1, C = 0)$
    $= \frac{P(A=0,B=1,C=0|+)P(+)}{P(A=0,B=1,C=0)}$
    $= \frac{P(A=0|+)P(B=1|+)P(C=0|+)P(+)}{P(A=0,B=1,C=0)}$
    $= \frac{0.4 \times 0.2 \times 0.2 \times 0.5}{Q} = \frac{0.008}{Q}$

$P(-|A=0,B=1,C=0)$
$= \frac{P(A=0,B=1,C=0|-)P(-)}{P(A=0,B=1,C=0)}$
$= \frac{P(A=0|-)P(B=1|-)P(C=0|-)P(-)}{P(A=0,B=1,C=0)}$
$= \frac{0.6\times0.4\times0\times0.5}{Q} = \frac{0}{Q}$
Here, as the $P(+|A=0,B=1,C=0)$ have greater value so the class label should be "+".

(c) Given that, $p = \frac{1}{2}, m = 4$.
$P(A=1|+) = \frac{5}{9} = 0.55$
$P(A=1|-) = \frac{4}{9} = 0.44$
$P(A=0|+) = \frac{4}{9} = 0.44$
$P(A=0|-) = \frac{5}{9} = 0.55$
$P(B=1|+) = \frac{3}{9} = 0.33$
$P(B=1|-) = \frac{4}{9} = 0.44$
$P(B=0|+) = \frac{6}{9} = 0.66$
$P(B=0|-) = \frac{5}{9} = 0.55$
$P(C=1|+) = \frac{6}{9} = 0.66$
$P(C=1|-) = \frac{7}{9} = 0.77$
$P(C=0|+) = \frac{3}{9} = 0.33$
$P(C=0|-) = \frac{2}{9} = 0.22$

(d) Here, $P(A=0,B=1,C=0) = Q$
$P(+|A=0,B=1,C=0)$
$= \frac{P(A=0,B=1,C=0|+)P(+)}{P(A=0,B=1,C=0)}$
$= \frac{P(A=0|+)P(B=1|+)P(C=0|+)P(+)}{P(A=0,B=1,C=0)}$
$= \frac{0.44\times0.33\times0.33\times0.5}{Q} = \frac{0.02469}{Q}$

$P(-|A=0,B=1,C=0)$
$= \frac{P(A=0,B=1,C=0|-)P(-)}{P(A=0,B=1,C=0)}$
$= \frac{P(A=0|-)P(B=1|-)P(C=0|-)P(-)}{P(A=0,B=1,C=0)}$
$= \frac{0.55\times0.44\times0.22\times0.5}{Q} = \frac{0.02743}{Q}$
Here, as the $P(-|A=0,B=1,C=0)$ have greater value so the class label should be "-".

(e) The estimation for conditional probabilities using the m-estimate probability approach is better when one of the conditional probability is zero, as we don't want our whole expression to become zero.

5. **4.13:**

Marital Status			
Class	Single	Married	Divorced
Yes	2	0	1
No	2	4	1

Home Owner		
Class	Yes	No
Yes	0	3
No	3	4

Here, i) d(status= Single, status=Married)
$=|\frac{2}{4} - \frac{0}{4}| + |\frac{2}{4} - \frac{4}{4}|$
$=|\frac{1}{2} - 0| + |\frac{1}{2} - 1| = 1$

ii) d(status= Single, status=Divorced)
$=|\frac{2}{4} - \frac{1}{2}| + |\frac{2}{4} - \frac{1}{2}| = 0$

iii) d(status= Married, status=Divorced)
$=|\frac{0}{4}-\frac{1}{2}|+|\frac{4}{4}-\frac{1}{2}|=1$

iv) d(HomeOwner= Yes, HomeOwner= No)
$=|\frac{0}{3}-\frac{3}{7}|+|\frac{3}{3}-\frac{4}{7}|$
$=\frac{6}{7}$

2 2.2 General Refresher

1. **1.** Given that, $g(x) = -3x^2 + 24x - 30$
   Here,
   b= 24
   a= -3
   So, $x = -\frac{b}{2a}$
   $= -\frac{24}{2(-3)}$
   $= 4$
   So value for x=4 maximizes g(x).

2. **2.** Given that,$f(x) = 3x_0^3 - 2x_0x_1^2 + 4x_1 - 8$
   So,the partial derivative of $f(x)$ in respect to x_0: $f(x_0) = 9x_0^2 - 2x_1^2$
   And the partial derivative of $f(x)$ in respect to x_1: $f(x_1) = -4x_0x_1 + 4$

3. **3.** The matrix Python code can be found here:
   https://colab.research.google.com/drive/12MyS1TrVs4WWeBn83uviWpBdmeDzxaNU?usp=sharing

4. **5.** A Bernoulli random variable has two possible outcomes: 0 or 1. A binomial distribution is the sum of independent and identically distributed Bernoulli random variables.
   A bernoulli random variable X is a random variable that satisfies $P(X = 1) = p, P(X = 0) = 1 - p$. If $X = X_1 + X_2 + \ldots + X_n$ and each of $X_1, X_2, \ldots, X_n$ has a bernoulli distribution with the same value of p and they are independent,then X has a binomial distribution, and the possible values of X are $\{0, 1, 2, 3, \ldots, n\}$. So a bernoulli distribution is a special case of binomial distribution. Specifically, when $n = 1$ the binomial distribution becomes bernoulli distribution. Another way to say all bernoulli distributions are binomial distributions, but most binomial distributions are not bernoulli distributions.

5. **7.** Given that, random variable Y has distribution

$$P(Y = y) = \begin{cases} \mathrm{e}^{-y}, & \text{if } y \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

   (a) $\int_{y=-\infty}^{\infty} p(Y = y)\, dy$
   $=\int_{y=-\infty}^{0}$P(Y=y) dy + $\int_{y=0}^{\infty}$P(Y=y) dy
   $= \int_{y=0}^{\infty}\mathrm{e}^{-y}$ dy
   $= [\ \frac{-1}{e^y}]_0^\infty$
   = -0+1 = 1

   (b) $\mu = \int\limits_{-\infty}^{\infty} yp(Y = y)dy$
   $=\int\limits_0^\infty y\mathrm{e}^{-y}\ dy = [-y\mathrm{e}^{-y}\ ]_0^\infty + \int\limits_0^\infty \mathrm{e}^{-y}\ dy$
   $= [-y\mathrm{e}^{-y}\ ]_0^\infty + [-\mathrm{e}^{-y}\ ]_0^\infty$
   $= [-y\mathrm{e}^{-y}\ - \mathrm{e}^{-y}\ ]_0^\infty$

$= \lim_{y\to\infty} [-y\mathrm{e}^{-y} \ - \mathrm{e}^{-y}\ ]$ - $\left[0\mathrm{e}^{-0} \ - \mathrm{e}^{-0}\ \right]$
$= [0 - 0]$ - $[0 - 1]$
$= 1$

(c) $\sigma^2 = \int_{y=-\infty}^{\infty} \mathrm{y}^2 \mathrm{p(Y{=}y)}\ \ \mathrm{dy} - \mu^2$
$= \int_0^{\infty} \mathrm{y}^2 e^{-y}\ \ \mathrm{dy} - \mu^2$
$= [-2e^{-y} - 2ye^{-y} - y^2e^{-y}]_0^{\infty} - \mu^2$
$= 2 - 1^2 = 1$

(d) $E[Y|Y \geq 10]$
$= \int_{10}^{\infty} ye^{-y}\, dy$
$= [-ye^{-y}]_{10}^{\infty} + [-e^{-y}]_{10}^{\infty}$
$= [-ye^{-y} - e^{-y}]_{10}^{\infty}$
$= [\frac{-y}{e^y} - \frac{1}{e^y}]_{10}^{\infty}$
$= 0 + 0 + \frac{10}{e^{10}} + \frac{1}{e^{10}}$
$= \frac{11}{e^{10}}$
$= 4.99 \times 10^{-4}$
$= 5 \times 10^{-4}$