CSE343/CSE543/ECE363/ECE563: Machine Learning Sec A (Monsoon 2023)
Mid-sem

Date of Examination: 26.09.2023 Duration: 1 hour Total Marks: 25 marks

---

Instructions –

- Attempt all questions.MCQs have a single correct option.
- State any assumptions you have made clearly.
- Standard institute plagiarism policy holds.
- No evaluation without suitable justification.

---

0 marks if the option or justification of MCQs is incorrect.

1. What does the term "naive" in Naive Bayes refer to? **[1 mark]**

   (A) The algorithm's simplicity and ease of implementation.
   (B) The assumption that all features are independent, given the class.
   (C) The use of Bayesian statistics.
   (D) The ability to handle continuous and categorical data simultaneously.

   Solution: B) The assumption that all features are independent, given the class.
   Explanation - The term "naive" in Naive Bayes refers to the assumption that all features are independent of each other, given the class label. This is a simplifying assumption that makes the algorithm computationally tractable and is why it is called "naive."

2. You picked a parameter theta using K-fold cross validation. The value of K in this case is 5. What do you think is the best way to pick the final model and estimate the error? **[1 mark]**

   (A) Pick any of the 5 models you built for your model; use its error estimate on the held-out data.
   (B) Train a new model on the full data set, using the θ you found; use the average CV error its error estimate.
   (C) Pick any of the 5 models you built for your model; use the average CV error for the 5 models as its error estimate.
   (D) Train a new model on the full data set, using the θ you found; use the total CV error as its error estimate.

   Solution: B) After having picked the best parameter, we train a new model on the whole dataset, and use the average CV error as the error estimate.

3. Which of the following statements is true about lasso and ridge regularization? **[1 mark]**

   (A) Lasso regularization can produce sparse weights and can perform automated feature selection.
   (B) Ridge regularization can produce sparse weights and can perform automated feature selection.
   (C) Lasso regularization cannot produce sparse weights but can perform automated feature selection.
   (D) Ridge regularization cannot produce sparse weights but can perform automated feature selection.

Solution- A: Unlike ridge, lasso regularization turns weights to zero which indirectly performs feature selection.

4. In your experiment, you wish to strongly penalize outliers. Consider the formula for mean squared error (MSE), which is often used as a loss function:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Is it a good idea to use mean cubed error (consider power 3 instead of 2) as a loss function instead? (Yes/No). **[1 mark]**
Solution- Some of the possible problems -

(A) Cubed error will make each term positive or negative, and errors will cancel each other out

(B) Higher degrees of the exponent gives more penalty to outliers, one bad point will compromise the fit

5. You are solving a biometric authentication task (modeled as binary classification) that uses fingerprint data to help users log into their devices. You train a classification model for user A until it achieves $> 95\%$ accuracy on a held-back development set (for the same user). However, upon deployment, you get complaints that the model fails to correctly authenticate user A about half the time (50% misclassification rate). List one factor you think could have contributed to the mismatch in misclassification rates between the dev set and deployment, and how you'd go about fixing this issue. **[2 marks]**

The model does really well on the dev set, but fails upon deployment (can be thought of as a test set). This reduces to a distribution discrepancy between the test and dev distributions.
There are multiple factors that could contribute to this, but a couple of important (and popular) ones are:

- Class imbalance in dev set: Model could deterministically predict the same label all the time, yet achieve 95% dev accuracy if 95% of the dev set samples belonged to that label. This could be solved by using a balanced dev set, or using an alternate evaluation metric such as F1 score.

- Distribution shift between dev and deployment: Key change in biometric attribute being monitored. Eg.- alteration to fingerprint (fingerprint recognition), etc.. Solution would be to retrain the model with data from new conditions.

1 mark without explanation, 2 mark for answer with proper explanation
consider other valid points as well

6. What is the perpendicular distance of the point x = [6,7] from the SVM hyperplane wTx + b where w = [2,5] and b=1. Also, find the projection of the point onto the hyperplane. **[2 marks]**

To find the perpendicular distance of the point $x = [6, 7]$ from the SVM hyperplane $w^T x + b$ where $w = [2, 5]$ and $b = 1$, you can use the following formula for the distance of a point from a hyperplane:

$$\text{Distance} = \frac{|w^T x + b|}{\|w\|}$$

where w^T is the transpose of vector w, x is the point, b is the bias term, and $\|w\|$ is the magnitude (norm) of vector w.

Given the values: $x = [6, 7]$, $w = [2, 5]$, $b = 1$ Calculate $w^T x$:

$$w^T x = [2, 5] \cdot [6, 7] = 2 \cdot 6 + 5 \cdot 7 = 12 + 35 = 47$$

Calculate $\|w\|$, which is the magnitude (norm) of vector w:

$$\|w\| = \sqrt{2^2 + 5^2} = \sqrt{4 + 25} = \sqrt{29}$$

Now, you can calculate the perpendicular distance:

$$\text{Distance} = \frac{|w^T x + b|}{\|w\|} = \frac{|47 + 1|}{\sqrt{29}} = \frac{48}{\sqrt{29}}$$

To find the projection of the point $x = [6, 7]$ onto the hyperplane

$$\text{Projection} = x - \frac{(w^T x + b)}{\|w\|^2} w$$

Substitute the values:

$$\text{Projection} = [6, 7] - \frac{(47 + 1)}{(\sqrt{29})^2}[2, 5] = [6, 7] - \frac{48}{29}[2, 5]$$

$$\text{Projection} = [6, 7] - \left[\frac{96}{29}, \frac{240}{29}\right]$$

$$\text{Projection} = \left[\frac{78}{29}, \frac{-37}{29}\right]$$

1 mark for correct distance
1 mark for correct projection

7. Derive the inverse logit function. **[2 marks]**

$$\text{Let } y = \text{logit}(x) = \log \tfrac{x}{1-x}$$

$$e^y = \frac{x}{1 - x}$$

$$1 + e^y = \frac{1 - x}{1 - x} + \frac{x}{1 - x} = \frac{1}{1 - x}$$

$$\frac{1}{1 + e^y} = 1 - x$$

$$x = 1 - \frac{1}{1 + e^y} = \frac{e^y}{1 + e^y}$$

0.5 mark for correct logit definition, 1.5 mark for steps

8. Discuss the various possibilities that can arise in the use of the leave-one-out method by eliminating any one pattern from the training sample and constructing an SVM solution based on the remaining patterns. **[2 marks]**

    1. During LOOM, one might pick up a data point which is a support vector. This will alter the model's decision boundary, change the margin width, and lead to misclassification in that iteration.
    2. If any point other than the support vector is left out, then it would not affect the decision boundary.
    3. If a class has only one data point, and that point is left out, the algorithm will not work.

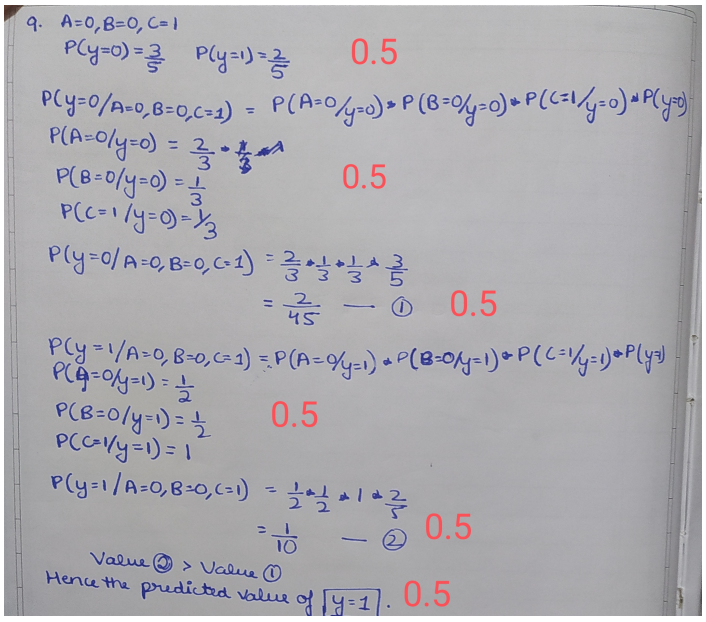

Figure 1: Solution for Q9

A	B	C	y
0	0	1	0
0	1	0	0
1	1	0	0
0	0	1	1
1	1	1	1

Table 1:

Land Area (sq. ft)	House Price (USD)
90	85
105	90
85	80

Table 2:

1.5 mark for option 1, 0.5 mark for option 2.

If option 3 is mentioned, give 0.5 if no marks have been given in the question.

9. How would a naive Bayes classifier predict y given this input: $A = 0, B = 0, C = 1$? Assume that in case of a tie, the classifier always prefers to predict 0 for y. Refer table 1. **[4 marks]**

10. The following plot represents four different cases both high and low bias and variance. The red colored circle represents the model and the dark blue dots represent the predictions. Which of the following fig represents the given cases. **[2 marks]**
A) High Bias, Low variance B) High Variance, Low bias
Solution: A) Figure (iii) represents high bias,low variance B) Figure (i) represents low bias, high variance

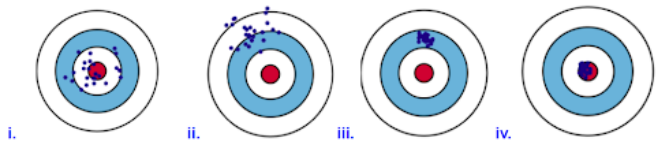

Figure 2: Q10

11. Suppose that the lifetime of Badger brand light bulbs is modeled by an exponential distribution with an

unknown parameter λ. We test 5 bulbs and find they have lifetimes of 2, 3, 1, 3, and 4 years, respectively. What is the Maximum Likelihood Estimator for λ given that the PDF of X_i is $f(x_i) = \lambda e^{-\lambda x_i}$? **[2 marks]**

Solution: Rubric - 1.5 marks for correct method of derivation, 0.5 mark for correct answer

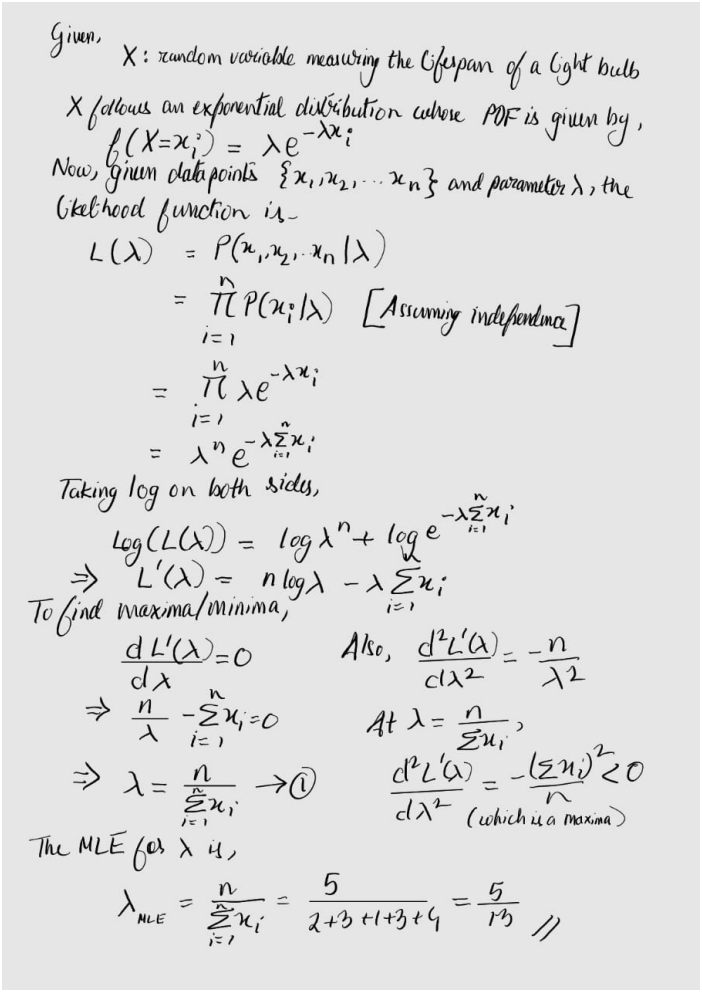

Figure 3: Solution for Q11

12. Given a kernel function $k(\cdot, \cdot)$ that maps the points to some feature space defined by an unknown mapping $\phi(\cdot)$, how would you compute the Euclidean distance in the feature space, i.e., between $\phi(x)$ and $\phi(y)$, using only the kernel function $k(\cdot, \cdot)$ and the points x and y? **[2 marks]**

Solution: The distance between $\phi(x)$ and $\phi(y)$ is computed as follows:

$$\sqrt{(\phi(x) - \phi(y))^T (\phi(x) - \phi(y))} \tag{1}$$

This can be expanded as:

$$\sqrt{\phi(x)^T \phi(x) + \phi(y)^T \phi(y) - 2\phi(x)^T \phi(y)} \tag{2}$$

This can be written in terms of the kernel function as:

$$\sqrt{k(x,x)+k(y,y)-2k(x,y)} \tag{3}$$

13. Three randomly chosen land areas were sold for the prices mentioned in the table below. A real estate dealer wants to predict the house price for any given land area based on these samples. Using the data provided, come up with a linear regression equation that best predicts the house prices.

    Note: There are other alternative approaches as well (The best one would be the normal equation method if you had access to scientific calculators). Please provide your solution and justification. Refer table 2 **[3 marks]**

    Solution: 2 marks for correct method, 0.5 marks for each correct parameter value

Method 1

If the estimated function is $\hat{y} = f(x) = a_0 + a_1 x$

then the optimal parameters is given by,

$$a_1 = \frac{n\Sigma x_i \Sigma y_i - \Sigma x_i \Sigma y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2}$$

$$= \frac{3 \times 23900 - 280 \times 255}{3 \times 26350 - 280^2}$$

$$= 0.4615$$

Also $a_0 = \frac{\Sigma y_i - a_1 \Sigma x_i}{n} = \frac{255 - 0.4615 \times 280}{3} = 41.923$

Thus the linear regression model is $y = 41.923 + 0.4615x$

Method 2

$\theta = (X^T X)^{-1} X^T Y$ where $X = \begin{pmatrix} 1 & 90 \\ 1 & 105 \\ 1 & 85 \end{pmatrix}$ and $Y = \begin{pmatrix} 85 \\ 90 \\ 80 \end{pmatrix}$

Solving this we get,

$$\theta = \begin{pmatrix} 41.923 \\ 0.4615 \end{pmatrix} = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix}$$

The linear regression model is $y = 41.923 + 0.4615x$

Figure 4: Solution for Q13