CSE 5523: Machine Learning - Midterm

11:59 pm 10/23/2025

Policy:

- You have **one day** to complete the midterm exam. The exam will be released at 12:00 AM on Oct 23rd (Thursday). You need to submit it before Oct 23rd at 11:59 PM to Carmen as PDF file. Please make sure your submission is **recognizable**.
- It will be take home exam. If you have any questions about the questions, you can send me an email.
- You are allowed to use lecture slides, class notes, review materials, and homework assignments, but **not allowed** to use AI tools or search for answers directly from the Internet.
- You are allowed to directly use the results we derived in class (e.g., MLE for Bernoulli, Gaussian distributions, closed-form solutions of linear regression)
- You are **not allowed** to discuss with other students during the exam. You must complete the exam on your own.
- Any violation may lead to 0 points for your midterm exam. I have to report to university if there is a violation of University's academic misconduct and integrity policy.
- The contents of the exam are not allowed to be reproduced, distributed, or transmitted **at any time even after the exam**, in any form or by any means, without the permission of the instructor.

Exam content and grading: There are seven written questions in total (100 points). You should write down the detailed derivations and explain your answers. Partial credits will be given based on your justification.

1) **Bayes Optimal Classifier (10 pts).**

Consider one-dimensional feature $X \in \mathbb{R}$ and binary $Y \in \{+1, -1\}$. Given the following GDA model:

$$\Pr(Y = +1) = 0.7; \quad \Pr(Y = -1) = 0.3$$

$$\Pr(X = x|Y = +1) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$$

$$\Pr(X = x|Y = -1) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{(x-8)^2}{2})$$

Suppose we are given a new feature $x = 4$ and we want to find its prediction $\hat{Y}$ that minimizes the expected loss $\mathbb{E}[\mathbf{1}(\hat{Y} \neq Y)]$, what is the prediction of $x = 4$?

2) **MLE (10 pts).** Consider an exponential distribution. The density function is given by

$$P(x) = \begin{cases} \lambda \exp(-\lambda x), & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

Given a dataset $\{x_1, x_2, ..., x_n\}$, what is the maximum likelihood estimate $\hat{\lambda}_{ML}$ of the parameter λ?

3) **Linear regression (15 pts) .** Consider a linear regression problem, where we have four data points

$$x_1 = [0, 0]^T; \;\; y_1 = 0$$
$$x_2 = [0, 1]^T; \;\; y_2 = 1.5$$
$$x_3 = [1, 0]^T; \;\; y_3 = 2$$
$$x_4 = [1, 1]^T; \;\; y_4 = 2.5$$

Suppose we want to find $\widetilde{w} \in \mathbb{R}^3$ to minimize the following:

$$\min_{\widetilde{w} \in \mathbb{R}^3} \frac{1}{4} \sum_{i=1}^{4} (y_i - \widetilde{w}^T \widetilde{x}_i)^2$$

where $\widetilde{x}_i = \begin{bmatrix} 1 \\ x_i \end{bmatrix} \in \mathbb{R}^3$.

(a) (8 pts) What is the optimal value for $\widetilde{w}$?

(a) (7 pts) If we want to solve the problem using gradient descent, what is the gradient descent update with learning rate $\eta > 0$, i.e., write the update in the form of $\widetilde{w}_{t+1} \leftarrow f(\widetilde{w}_t)$ (you should find f).

4) **Linear Discriminant Analysis (15 pts).** Consider a binary classification with the following dataset:

$$
\begin{aligned}
x_1 &= [1, 0]^T; \;\; y_1 = 0 \\
x_2 &= [0, 1]^T; \;\; y_2 = 0 \\
x_3 &= [1, 1]^T; \;\; y_3 = 0 \\
x_4 &= [-1, 0]^T; \;\; y_4 = 1 \\
x_5 &= [0, -1]^T; \;\; y_5 = 1 \\
x_6 &= [-1, -1]^T; \;\; y_6 = 1
\end{aligned}
$$

We want to train a linear discriminant analysis (LDA) model from the above dataset.

(a) (7 pts) To find LDA, we need to estimate $P(Y = y)$ and $P(X|Y = y)$ from labeled data. Let's use maximum likelihood estimators, what are estimated $P(Y = y)$ and $P(X|Y = y)$?

(b) (4 pts) Given $\tilde{x} = [0, -2]^T$, what is the predicted label $\hat{y} = \arg\max_{y \in \{0,1\}} P(Y = y|X = \tilde{x})$?

(c) (2 pts) Is LDA a generative model or discriminative model?

(d) (2 pts) Is the following statement true or false: LDA **cannot** be applied if the true class-conditional density $P(X|Y = y)$ for each class is not Gaussian.

5) **Naive Bayes (15 pts).** Consider the following dataset

$$x_1 = [0, 0, 1]^T; \;\; y_1 = 0$$
$$x_2 = [0, 1, 0]^T; \;\; y_2 = 0$$
$$x_3 = [1, 1, 0]^T; \;\; y_3 = 0$$
$$x_4 = [0, 0, 1]^T; \;\; y_4 = 1$$
$$x_5 = [1, 1, 1]^T; \;\; y_5 = 1$$
$$x_6 = [1, 0, 0]^T; \;\; y_6 = 1$$
$$x_7 = [1, 1, 0]^T; \;\; y_7 = 1$$

We want to train a Naive Bayes classifier from the dataset.

- **(a) (7 pts)** To find Naive Bayes classifier, we need to estimate $P(Y = y)$ and $P(X[d]|Y = y)$ from labeled data. Let's use maximum likelihood estimators, what are estimated $P(Y = y)$ and $P(X[d]|Y = y)$?

- **(b) (4 pts)** Given $\tilde{x} = [0, 0, 1]^T$, what is the predicted label $\hat{y} = \arg\max_{y \in \{0,1\}} P(Y = y|X = \tilde{x})$?

- **(c) (2 pts)** Is Naive Bayes a generative model or discriminative model?

- **(d) (2 pts)** Is the following statement about Naive Bayes true or false: The core assumption of Naive Bayes classifiers is that all observed variables (features) are independent, i.e., $P(X[1], \cdots, X[D]) = \prod_{d=1}^{D} P(X[d])$

6) **Logistic regression (15 pts).** Consider a binary classification problem with the following dataset:

$$\begin{aligned} x_1 &= [1,1]^T; \;\; y_1 = 0 \\ x_2 &= [2,2]^T; \;\; y_2 = 0 \\ x_3 &= [3,3]^T; \;\; y_3 = 0 \\ x_4 &= [2,3]^T; \;\; y_4 = 1 \\ x_5 &= [3,4]^T; \;\; y_5 = 1 \\ x_6 &= [4,5]^T; \;\; y_6 = 1 \end{aligned}$$

We want to train a logistic regression model from the dataset. To find the model, we need to estimate $P(Y = 1|X = x) = \sigma(w^T x + b) = \sigma(\widetilde{w}^T \widetilde{x})$ from labeled data, where $\sigma(\widetilde{w}^T \widetilde{x}) = \frac{1}{1+\exp(-\widetilde{w}^T \widetilde{x})}$ is the sigmoid function and we denote $\widetilde{w} = \begin{bmatrix} b \\ w \end{bmatrix}, \widetilde{x} = \begin{bmatrix} 1 \\ x \end{bmatrix}$ to simplify notations.

In the class, we show that the parameter $\widetilde{w}$ can be found by solving the following minimization problem:

$$\min_{\widetilde{w}} - \sum_{i=1}^{N} \big(y_i \log \sigma(\widetilde{w}^T \widetilde{x}_i) + (1 - y_i) \log(1 - \sigma(\widetilde{w}^T \widetilde{x}_i)) \big)$$

- **(a) (7 pts)** Using gradient descent with learning rate η, describe the steps to optimize $\widetilde{w}$ of the logistic regression model.

- **(b) (6 pts)** If the logistic regression model is trained and the learned parameter is $\widetilde{w} = [-4, 1.2, 0.8]^T$. Given a data point $\hat{x} = [3,3]^T$, what is the predicted label $\hat{y} = \arg\max_{y \in \{0,1\}} P(Y = y|X = \hat{x})$?

- **(c) (2 pts)** Is logistic regression a generative model or discriminative model?

7) **Maximum Margin Classifier (20 pts).**

Consider a dataset with two data points $(x_1, y_1), (x_2, y_2)$:

$$x_1 = [0, 0]^T; \quad y_1 = -1$$
$$x_2 = [2, 1]^T; \quad y_2 = +1$$

(a) (8 pts) Find the parameters w^*, b^* of maximum margin classifier by solving the following optimization:

$$\begin{aligned} \min_{w,b} \quad & \tfrac{1}{2}||w||^2 \\ \text{s.t.} \quad & y_i(w^T x_i + b) \geq 1, \forall i \end{aligned}$$

(b) (4 pts) Explain why minimizing $||w||^2$ is equivalent to maximizing the margin.

(c) (4 pts) From the lecture we know that the optimal w^* can be written as a linear combination of data points:

$$w^* = \sum_{i=1}^{2} \alpha_i^* y_i x_i$$

Find α_2^*.

(d) (2 pts) In class we learnt that SVM can be used to classify linearly inseparable data by transforming it to a higher dimensional space with a kernel $k(x; z) = \phi(x)^T \phi(z)$, where $\phi(x)$ is a feature mapping. Let $k_1 : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ be a valid kernel function, and $c \in \mathbb{R}_+$ be a positive constant. $\phi_1 : \mathbb{R}^n \to \mathbb{R}^d$ is feature mapping of k_1. Explain how to use ϕ_1 to obtain the kernel $k(x, z) = ck_1(x, z)$.

(e) (2 pts) Suppose we have another dataset that is not linearly separable. We want to find optimal soft-margin hyperplane by solving the following optimization:

$$\begin{aligned} \min_{w,b,\{\xi_i\}_{i=1}^n} \quad & \tfrac{1}{2}||w||^2 + \tfrac{C}{n} \textstyle\sum_{i=1}^n \xi_i \\ \text{s.t.} \quad & y_i(w^T x_i + b) \geq 1 - \xi_i, \forall i \\ & \xi_i \geq 0, \forall i \end{aligned}$$

Is the following statement true or false: The optimal soft-margin hyperplane classifier tends to have a larger margin when the parameter C increases.