FUNDAMENTALS OF MACHINE LEARNING
AA 2025-2026
Prova Intermedia (FACSIMILE)

4 November, 2025

Istruzioni: Niente libri, niente appunti, niente dispositivi elettronici, e niente carta per appunti. Usare matita o penna di qualsiasi colore. Usare lo spazio fornito per le risposte.
Instructions: No books, no notes, no electronic devices, and no scratch paper. Use pen or pencil. Use the space provided for your answers.
This exam has 5 questions, for a total of 100 points and 10 bonus points.

Nome: _____

Matricola: _____

1. **Multiple Choice**: Select the correct answer from the list of choices.

(a) [5 points] True or False: Adding an L_2 regularizer to least squares regression will reduce bias. ◯ True √ **False**

(b) [5 points] True or False: A zero-mean Gaussian Prior (i.e. $p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \sigma I)$) on model parameters in a MAP estimate corresponds to adding an L2 regularization term (e.g. $||\mathbf{w}||_2$) to the loss in an MLE estimate. √ **True** ◯ False

(c) [5 points] True or False: In the Primal Form of the SVM, increasing hyperparameter C will decrease the complexity of the resulting classifier. ◯ True √ **False**

(d) [5 points] True or False: The Maximum a Priori (MAP) and Maximum Likelihood (ML) solution for linear regression are always equivalent. ◯ True √ **False**

(e) [5 points] If a hard-margin support vector machine tries to minimize $||\mathbf{w}||_2$ subject to $y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 2$, what will be the size of the margin?
◯ $\frac{1}{||\mathbf{w}||}$ √ $\frac{2}{||\mathbf{w}||}$ ◯ $\frac{1}{2||\mathbf{w}||}$ ◯ $\frac{1}{4||\mathbf{w}||}$

(f) [5 points] The posterior distribution of B given A is:
◯ $P(B \mid A) = \frac{P(A|B)P(A)}{P(B)}$
◯ $P(B \mid A) = \frac{P(A,B)P(B)}{P(A)}$
√ $P(B \mid A) = \frac{P(A|B)P(B)}{P(A)}$
◯ $P(B \mid A) = \frac{P(A|B)P(B)}{P(A,B)}$

(g) [5 points] Let $\mathbf{w}^*$ be the solution obtained using unregularized least-squares regression. What solution will you obtain if you scale all input features by a factor of c before solving?
◯ $c\mathbf{w}^*$ ◯ $c^2\mathbf{w}^*$ ◯ $\frac{1}{c^2}\mathbf{w}^*$ √ $\frac{1}{c}\mathbf{w}^*$

Total Question 1: 35

1 (c):

$$(w^*, b^*) = \underset{(w,b)}{\arg\max} \left\{ \frac{1}{2}\|\vec{w}\|^2 + C \sum_n \max(0, 1 - y_n f(\vec{x}_n)) \right\}$$

$f(\vec{x}_n) = \langle \vec{w}, \vec{x}_n \rangle + b$

hinge loss

C weights errors made on train set.

2. **Multiple Answer**: Select **ALL** correct choices: there may be more than one correct choice, but there is always at least one correct choice.

(a) [5 points] What are support vectors?

- √ **The examples $\mathbf{x}_n$ from the training set required to compute the decision function $f(\mathbf{x})$ in an SVM.**
- ◯ ~~The class means~~.
- ◯ ~~The training samples farthest from the decision boundary~~.
- √ **The training samples $\mathbf{x}_n$ that are on the margin (i.e. $y_n f(\mathbf{x}_n) = 1$).**

(b) [5 points] Which of the following are true about the relationship between the MAP and MLE estimators for linear regression?

- √ **They are equal in the limit of infinite training samples.**
- ◯ They are equal if $p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \sigma)$ for very small σ.
- √ **They are equal if** ~~$p(\mathbf{w}) = 1$~~.
- ◯ ~~They are never equal~~.

(c) [5 points] You train a linear classifier on 10,000 training points and discover that the training accuracy is only 67%. Which of the following, done in isolation, has a good chance of improving your training accuracy?

- √ **Add novel features.**
- ◯ ~~Train on more data.~~
- ◯ ~~Regularize the model.~~
- √ **Train on less data.**

(d) [5 points] What assumption does the quadratic Bayes generative classifier make about class-conditional covariance matrices?

- ◯ ~~That they are equal~~.
- ◯ ~~That they are diagonal.~~
- ◯ ~~That their determinants are equal~~.
- √ **None of the above.**

(e) [5 points] Which of the following are reasons why you might adjust your model in ways that increase the bias?

- ◯ You observe high training error and high validation error.
- √ **You have few data points.**
- √ **You observe low training error and high validation error.**
- ◯ ~~Your data are not linearly separable~~.

(f) [5 points] Which of the following are true of polynomial regression (i.e. least squares regression with polynomial basis mapping)?

- √ **If we increase the degree of polynomial, we increase variance.**
- ◯ ~~The regression function is nonlinear in the model parameters.~~
- ◯ ~~The regression function is linear in the original input variables~~.
- √ **If we increase the degree of polynomial, we decrease bias.**

(g) [5 points] Which of the following classifiers can be used on non linearly separable datasets?

- ◯ The hard margin SVM.
- √ **Logistic regression.**
- √ **The linear generative Bayes classifiers.**
- √ **Fisher's Linear Discriminant.**

Total Question 2: 35

3. [15 points] Assume the class conditional distributions for a two-class classification problem are $p(\mathbf{x} \mid \mathcal{C}_1) = \mathcal{N}(\boldsymbol{\mu}_1, \beta^{-1}I)$ and $p(\mathbf{x} \mid \mathcal{C}_2) = \mathcal{N}(\boldsymbol{\mu}_2, \beta^{-1}I)$. Moreover, assume that the class priors are equal: $p(\mathcal{C}_1) = p(\mathcal{C}_2)$. Show that the optimal decision boundary is *linear*, i.e. that it can be written as $H = \{\mathbf{x} \mid \mathbf{w}^T\mathbf{x} + b = 0\}$ for some $\mathbf{w}$ and b.

Hint: Remember that points $\mathbf{x}$ on the optimal decision boundary will satisfy $p(\mathcal{C}_1 \mid \mathbf{x}) = p(\mathcal{C}_2 \mid \mathbf{x})$, and that the formula for the multivariate Gaussian density is:

$$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\}$$

Solution: We must find the hypersurface where the class posterior densities are equal:

$$\begin{aligned} p(\mathcal{C}_1 \mid \mathbf{x}) &= p(\mathcal{C}_2 \mid \mathbf{x}) \\ \frac{p(\mathbf{x} \mid \mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x})} &= \frac{p(\mathbf{x} \mid \mathcal{C}_2)p(\mathcal{C}_2)}{p(\mathbf{x})} \\ p(\mathbf{x} \mid \mathcal{C}_1)p(\mathcal{C}_1) &= p(\mathbf{x} \mid \mathcal{C}_2)p(\mathcal{C}_2) \end{aligned} \tag{1}$$

Now let:

$$Z = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}}$$

and substitute this and the class-conditional densities into equation (1):

$$\begin{aligned} Z^{-1}\exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)^T(\beta^{-1}I)^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)\}p(\mathcal{C}_1) &= Z^{-1}\exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)^T(\beta^{-1}I)^{-1}(\mathbf{x} - \boldsymbol{\mu}_2)\}p(\mathcal{C}_2) \\ \exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)^T(\beta I)(\mathbf{x} - \boldsymbol{\mu}_1)\} &= \exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)^T(\beta I)(\mathbf{x} - \boldsymbol{\mu}_2)\} \\ -\frac{\beta}{2}(\mathbf{x} - \boldsymbol{\mu}_1)^T(\mathbf{x} - \boldsymbol{\mu}_1) &= -\frac{\beta}{2}(\mathbf{x} - \boldsymbol{\mu}_2)^T(\mathbf{x} - \boldsymbol{\mu}_2) \\ (\mathbf{x} - \boldsymbol{\mu}_1)^T(\mathbf{x} - \boldsymbol{\mu}_1) &= (\mathbf{x} - \boldsymbol{\mu}_2)^T(\mathbf{x} - \boldsymbol{\mu}_2) \\ \mathbf{x}^T\mathbf{x} - 2\boldsymbol{\mu}_1^T\mathbf{x} + \boldsymbol{\mu}_1^T\boldsymbol{\mu}_1 &= \mathbf{x}^T\mathbf{x} - 2\boldsymbol{\mu}_2^T\mathbf{x} + \boldsymbol{\mu}_2^T\boldsymbol{\mu}_2 \\ 2(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T\mathbf{x} + \boldsymbol{\mu}_1^T\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^T\boldsymbol{\mu}_2 &= 0 \end{aligned}$$

So, we may write the optimal decision boundary as:

$$H = \{\mathbf{x} \mid \mathbf{w}^T\mathbf{x} + b = 0\}$$

for $\mathbf{w} = 2(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T$ and $b = \boldsymbol{\mu}_1^T\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^T\boldsymbol{\mu}_2$.

4. [15 points] Suppose that the data $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$ in a regression problem are generated as:

$$y_n = \mathbf{w}^T \mathbf{x}_n + \varepsilon,$$

where $\varepsilon \sim \mathcal{N}(0, \beta^{-1})$ is a zero-mean random variable. Show that the maximum likelihood solution $\mathbf{w}_{\text{ML}}$ to this problem is equivalent to the solution that minimizes the squared error on $\mathcal{D}$.

Hint: Recall that the formula for the univariate Gaussian density is given by:

$$\mathcal{N}(y \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}.$$

Solution: Assuming the samples $(\mathbf{x}_n, y_n)$ are iid, we can write the likelihood of dataset $\mathcal{D}$ given model parameters $\mathbf{w}$ as:

$$\begin{aligned} \ln p(\mathcal{D} \mid \mathbf{w}) &= \ln \prod_{n=1}^N \mathcal{N}(y_n \mid \mathbf{w}^T \mathbf{x}_n, \beta^{-1}) \\ &= \frac{N}{2}(\ln \beta - \ln(2\pi)) - \frac{\beta}{2} \sum_{n=1}^N \{y_n - \mathbf{w}^T \phi(\mathbf{x}_n)\}^2. \end{aligned}$$

Now, the maximum likelihood solution is the $\mathbf{w}_{\text{ml}}$ that maximizes this:

$$\begin{aligned} \mathbf{w}_{\text{ml}} &= \arg\max_{\mathbf{w}} \left\{ \frac{N}{2}(\ln \beta - \ln(2\pi)) - \frac{\beta}{2} \sum_{n=1}^N \{y_n - \mathbf{w}^T \phi(\mathbf{x}_n)\}^2 \right\} \\ &= \arg\max_{\mathbf{w}} -\frac{\beta}{2} \sum_{n=1}^N \{y_n - \mathbf{w}^T \phi(\mathbf{x}_n)\}^2 \\ &= \arg\min_{\mathbf{w}} \sum_{n=1}^N \{y_n - \mathbf{w}^T \phi(\mathbf{x}_n)\}^2, \end{aligned}$$

and so $\mathbf{w}_{\text{ml}}$ is precisely the solution that minimizes the sum of squared errors.

5. [10 points (bonus)] Consider the dataset with 1-dimensional inputs:

$$\mathcal{D} = \{(+1, +1), (-1, +1), (+2, -1), (-2, -1)\}.$$

Find a nonlinear embedding $\phi : \mathbb{R} \to \mathbb{R}^2$ that makes $\mathcal{D}$ linearly separable. Show that the resulting embedded dataset is indeed linearly separable by solving for the optimal hard-margin SVM solution $(\mathbf{w}^*, b^*)$.

Solution: Let $\phi(x) = [x\ x^2]^T$ be our explicit embedding from $\mathbb{R}$ into $\mathbb{R}^2$. Any solution $(\mathbf{w}, b)$ must satisfy the following constraints:

$$\begin{aligned} w_1 + w_2 + b &= 1 \quad (1)\\ -w_1 + w_2 + b &= 1 \quad (2)\\ -2w_1 - 4w_2 - b &= 1 \quad (3)\\ 2w_1 - 4w_2 - b &= 1. \quad (4) \end{aligned}$$

Or, more precisely, if there exists $(\mathbf{w}, b)$ satisfying the above *equality* constraints, then the data are linearly separable.

Question: How should we interpret a solution satisfying the above *equality* constraints?

Adding (1) to (2) and adding (3) to (4) we get:

$$\begin{aligned} w_2 + b &= 1 \quad (5)\\ -4w_2 - b &= 1. \quad (6) \end{aligned}$$

And so, after adding these, we get $w_2 = -\frac{2}{3}$. Plugging this into (1) and (2) and subtracting (1) from (2), we then get $w_1 = 0$. Plugging both of these into any of the above equations, we then get $b = \frac{5}{3}$.

Thus $(\mathbf{w}, b) = ([0, \frac{2}{3}]^T, \frac{5}{3})$ satisfies the primal SVM constraints and the embedded dataset is linearly separable.