

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. $\|\mathbf{w}\|_2^2$) 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:
 $P(B | A) = \frac{P(A|B)P(A)}{P(B)}$
 $P(B | A) = \frac{P(A,B)P(B)}{P(A)}$
 $P(B | A) = \frac{P(A|B)P(B)}{P(A)}$
 $P(B | A) = \frac{P(A|B)P(B)}{P(A,B)}$
- (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?

- $c\mathbf{w}^*$ $c^2\mathbf{w}^*$ $\frac{1}{c^2}\mathbf{w}^*$ $\frac{1}{c}\mathbf{w}^*$

$$\vec{x} \rightarrow c\vec{x}$$

$$\vec{w}^* \vec{x} = y \Rightarrow \left(\frac{\vec{w}^*}{c} \right) (c\vec{x}) = \cancel{y}$$

Total Question 1: 35

$$\langle \vec{\omega}, \vec{x}_n \rangle + b$$

1(c):

$$(\omega^*, b^*) = \arg \max_{(\omega, b)} \left\{ \underbrace{\frac{1}{2} \|\vec{\omega}\|^2}_{\text{regularization term}} + C \sum_n \max(0, 1 - y_n f(\vec{x}_n)) \right\}$$

hinge loss

C weight 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 x_n from the training set required to compute the decision function $f(x)$ in an SVM.
- The class means.
- The training samples farthest from the decision boundary.
- The training samples x_n that are on the margin (i.e. $y_n f(x_n) = 1$).

Sparse Kernel Machine

(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 $p(w) = N(\vec{0}, \sigma)$ for very small σ .
- They are equal if $p(w) = \star \rightarrow$ Uninformative prior
- They are never equal.

$$p(\vec{\omega}) = N(\vec{0}, I e^{-10})$$

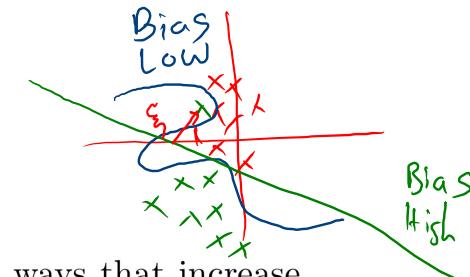
(c) [5 points] You train a linear classifier on 10,000 training points and discover that the training accuracy is only 67%. Which of the following, done in isolation, has a good chance of improving your training accuracy?

- Add novel features. $\Rightarrow +\text{features}$
- Train on more data. $\Rightarrow +\text{Model Size}$
- 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.

$$p(\vec{x} | C_1)$$



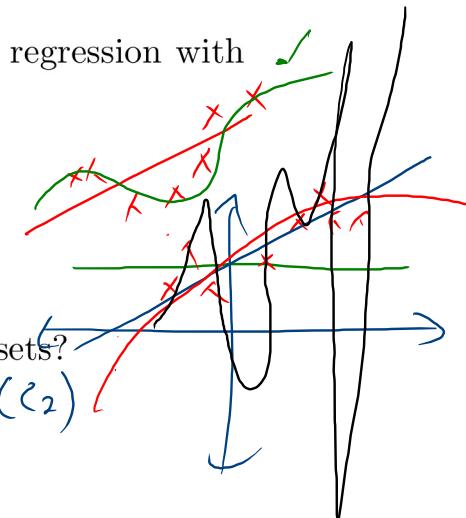
(e) [5 points] Which of the following are reasons why you might adjust your model in ways that increase the bias?

- You observe high training error and high validation error.
- You have few data points. \rightarrow maybe overfitting
- You observe low training error and high validation error.
- Your data are not linearly separable.

Definitely overfitting!

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

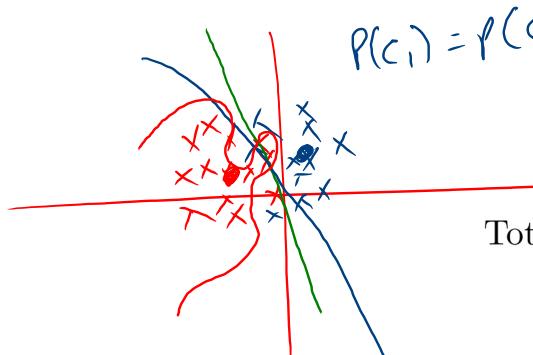
- If we increase the degree of polynomial, we increase variance.
- The regression function is nonlinear in the model parameters.
- The regression function is linear in the original input variables.
- If we increase the degree of polynomial, we decrease bias.



(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.

$$p(C_1) = p(C_2)$$



Total Question 2: 35

Generative Bayes is
"More Work".

$$\parallel P(C_1), P(C_2), p(\vec{x} | C_1), p(\vec{x} | C_2), p(\vec{x} | C_1) \cdot p(C_1)$$

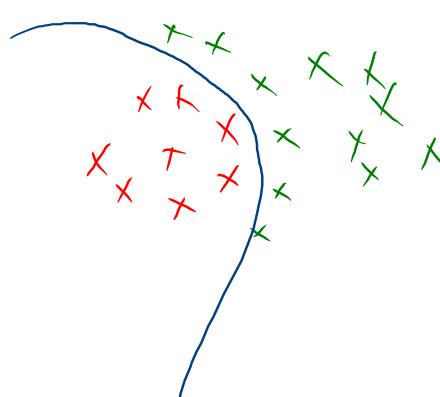
$2D + \Theta(D^2)$ parameters
Means Cov

Logistic regression

Directly estimate

Page 2

Posterior: $p(C_1 | \vec{x})$ with only D parameters



3. [15 points] Assume the class conditional distributions for a two-class classification problem are $p(\mathbf{x} | \mathcal{C}_1) = \mathcal{N}(\boldsymbol{\mu}_1, \beta^{-1}I)$ and $p(\mathbf{x} | \mathcal{C}_2) = \mathcal{N}(\boldsymbol{\mu}_2, \beta^{-1}I)$. 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 $H = \{\mathbf{x} | \mathbf{w}^T \mathbf{x} + b = 0\}$ for some \mathbf{w} and b .

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

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

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

$$\begin{aligned} p(\mathcal{C}_1 | \mathbf{x}) &= p(\mathcal{C}_2 | \mathbf{x}) \\ \frac{p(\mathbf{x} | \mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x})} &= \frac{p(\mathbf{x} | \mathcal{C}_2)p(\mathcal{C}_2)}{p(\mathbf{x})} \\ \underline{p(\mathbf{x} | \mathcal{C}_1)p(\mathcal{C}_1)} &= \underline{p(\mathbf{x} | \mathcal{C}_2)p(\mathcal{C}_2)} \end{aligned}$$

Now let:

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

$$\begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{a} & 0 \\ 0 & \frac{1}{b} \end{bmatrix}$$

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

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

So, we may write the optimal decision boundary as:

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

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}_{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:

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

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:

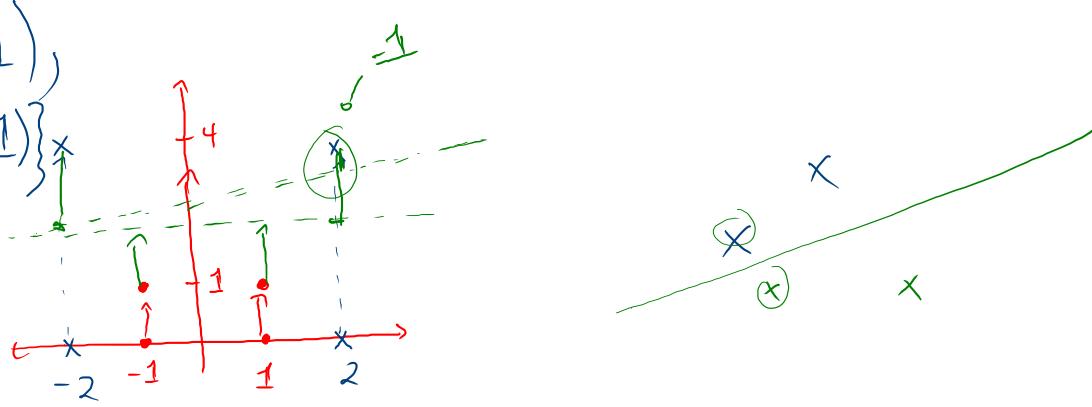
$$\begin{aligned} \ln p(\mathcal{D} | \mathbf{w}) &= \ln \prod_{n=1}^N \mathcal{N}(y_n | \mathbf{w}^T \mathbf{x}_n, \beta^{-1}) \\ &= \frac{N}{2} (\ln \beta - \ln(2\pi)) - \frac{\beta}{2} \sum_{n=1}^N \{y_n - \mathbf{w}^T \phi(\mathbf{x}_n)\}^2. \end{aligned}$$

Now, the maximum likelihood solution is the \mathbf{w}_{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}_{ml} is precisely the solution that minimizes the sum of squared errors.

$$\phi(D) = \{([1 1]^T, +1), ([-1 1]^T, +1), ([2 4]^T, -1), ([-2 4]^T, -1)\}$$



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

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

Find a nonlinear embedding $\phi : \mathbb{R} \rightarrow \mathbb{R}^2$ that makes 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:

$$+1 \left([\begin{matrix} w_1 & w_2 \end{matrix}] \begin{bmatrix} 1 \\ 1 \end{bmatrix} + b \right) \quad \begin{array}{l} w_1 + w_2 + b = 1 \\ -w_1 + w_2 + b = 1 \\ -2w_1 - 4w_2 - b = 1 \\ 2w_1 - 4w_2 - b = 1. \end{array} \quad +1 \left([\begin{matrix} w_1 & w_2 \end{matrix}] \begin{bmatrix} -1 \\ +1 \end{bmatrix} + b \right) \quad \begin{array}{l} (1) \\ (2) \\ (3) \\ (4) \end{array}$$

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:

$$w_2 + b = 1 \quad (5)$$

$$-4w_2 - b = 1. \quad (6)$$

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.

$$\begin{aligned} w_1 &\geq 0 \\ w_1 &\leq 0 \end{aligned} \quad \parallel = 0$$

$$b \geq \frac{2}{3}$$

$$b \leq -\frac{2}{3}$$

$$\frac{1}{2} \|\mathbf{w}\|^2 + \sum \max(-\dots)$$