# FUNDAMENTALS OF MACHINE LEARNING

## AA 2023-2024

### Prova Intermedia (FACSIMILE)

2 Novembre, 2023

> **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: ______________________________

$\ln [P(D|\vec{w})\, P(w)] - \ln P(D)$

All $w_i$ "small" on average

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 variance. √ **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

   Soft-margin

   (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~~ Posterior (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:
   - ○ $P(B \mid A) = \frac{P(A|B)P(A)}{P(B)}$
   - ○ $P(B \mid A) = \frac{P(A,B)P(B)}{P(A)}$
   - √ $P(B \mid A) = \frac{P(A|B)P(B)}{P(A)}$ ✓
   - ○ $P(B \mid A) = \frac{P(A|B)P(B)}{P(A,B)}$

   $\boxed{\lambda x}$

   (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

Ridge Regression

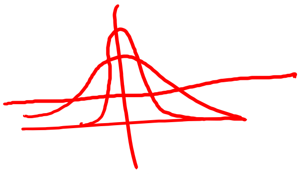
$\mathcal{L} = \sum_{i=1}^{N} (\vec{w}^T\vec{x}_i + b - y_i)^2 + \lambda \|\vec{w}\|_2^2$

Lasso Regression

$(+\lambda \|\vec{w}\|_1$ $= \sum_{i=1}^{d} |w_i|)$

Most $w_i = 0$ Sparse

$N(\vec{0}, \sigma^2)$

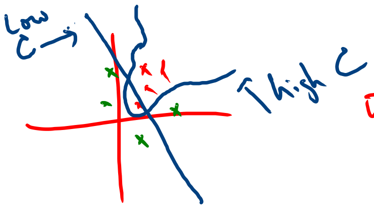


$w^{*\prime} = \frac{1}{c}\vec{w}^*$

$w^T(c\vec{x}) + b$

$\frac{1}{c}\vec{w}^T(c\vec{x}) + b$

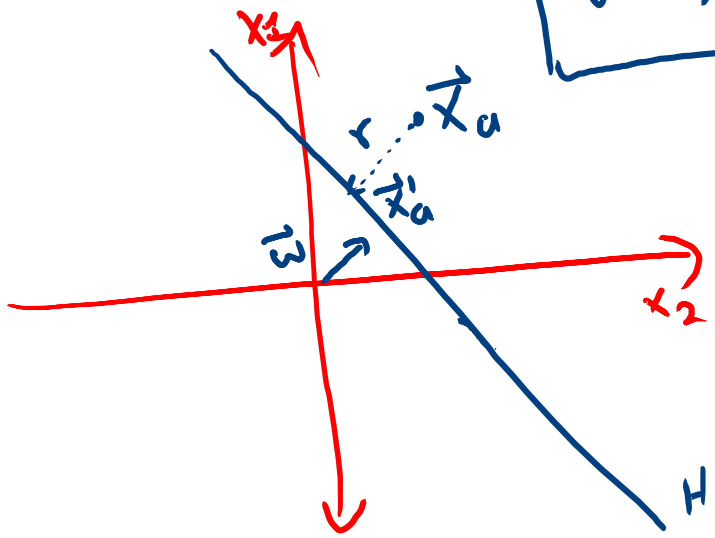
$\min \left[\frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{N} \xi_i\right]$ subject to ...




$\vec{w}_{ML} = \arg\max_{\vec{w}} P(D|\vec{w})$

$\vec{w}_{MAP} = \arg\max_{\vec{w}} P(\vec{w}|D) = \arg\max \frac{P(D|\vec{w})\,P(\vec{w})}{P(D)}$

#1 (e)




$$\boxed{\vec{w}^T \vec{x}_a + b = 2}$$

$$\vec{w} = [w_1 \quad w_2]^T$$

$$H_{\vec{w}} = \{ x \mid \underbrace{\vec{w}^T \vec{x} + b}\ = 0 \}$$

$$x_a' = \vec{x}_a - r \frac{\vec{w}}{\|\vec{w}\|}$$

$$\vec{w}^T \left( \vec{x}_a - r \frac{\vec{w}}{\|\vec{w}\|} \right) + b = 0$$

$$\Rightarrow \underbrace{\vec{w}^T \vec{x}_a + b}_{2} - r \frac{\vec{w}^T \vec{w}}{\|\vec{w}\|} = 0$$

$$2 - r \frac{\|\vec{w}\|^2}{\|\vec{w}\|} = 0$$

$$2 - r \|w\| = 0$$

$$r = 2 / \|\vec{w}\|$$

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?

Dual

- $\surd$ **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.
- $\surd$ **The training samples $\mathbf{x}_n$ that are on the margin (i.e. $y_n f(\mathbf{x}_n) = 1$).** Primal

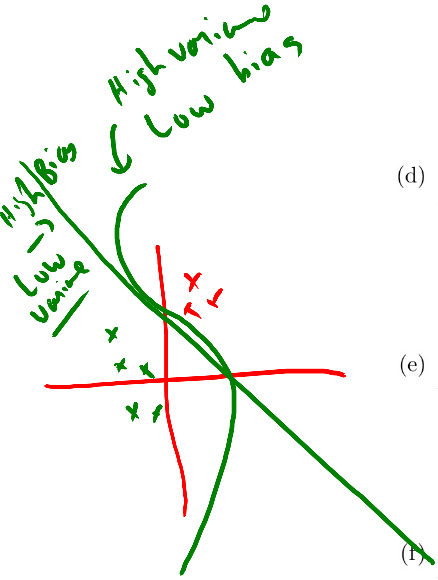
(b) [5 points] Which of the following are true about the relationship between the MAP and MLE estimators for linear regression?

- $\surd$ **They are equal if $p(\mathbf{w}) = \cancel{x}$** c
- ◯ They are equal if $p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \sigma)$ for very small $\sigma$.
- ◯ They are never equal.
- $\surd$ **They are equal in the limit of infinite training samples.** $\sim p(x, y)$

(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?

- $\surd$ **Add novel features.**
- ◯ Train on more data.
- $\surd$ **Train on less data.**
- ◯ Regularize the model.

$\Rightarrow$ Model with not enough Variance




(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.
- $\surd$ **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.
- $\surd$ **You have few data points.**
- $\surd$ **You observe low training error and high validation error.** — overfitting!
- ◯ 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)?

- $\surd$ **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.
- $\surd$ **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.
- $\surd$ **Logistic regression.**
- $\surd$ **The linear generative Bayes classifiers.**
- $\surd$ **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 $p(\mathbf{x} \mid \mathcal{C}_2) = \mathcal{N}(\boldsymbol{\mu}_2, \beta^{-1}I)$. 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:

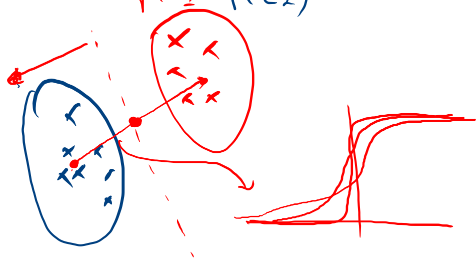
$$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\}$$

if $p(C_1) = p(C_2)$ is "easier"

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

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

$p(C_1) > p(C_2)$



Now let:

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

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

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

Something New!

So, we may write the optimal decision boundary as:

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

for $\mathbf{w} = -\frac{\beta}{2} \cdot 2(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T$ and $b = -\frac{\beta}{2}(\boldsymbol{\mu}_1^T\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^T\boldsymbol{\mu}_2) + \ln p(\mathcal{C}_1) - \ln p(\mathcal{C}_2)$.

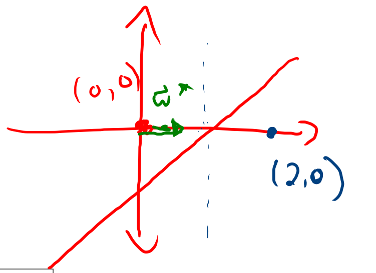
$+\ln p(S)$

$$-\frac{\beta}{2}(x^T\vec{x} - 2\vec{\mu}_1^T\vec{x} + \mu_1^T\vec{\mu}_1) + \ln p(C_1) = -\frac{\beta}{2}(x^T\vec{x} - 2\vec{\mu}_2^T\vec{x} + \vec{\mu}_2^T\vec{\mu}_2)$$

4. [15 points] Assume we have a training set of only two points (one from each class):

$$\mathcal{D} = \{([0,0], -1), ([2,0], +1)\}$$

Solve for the optimal hard margin primal SVM parameters $\mathbf{w}$ and $b$ for this dataset.




**Solution:** Since there are only two samples – one from each class – in $\mathcal{D}$, we know that both will be support vectors. Thus we can write the primal form of the hard-margin SVM learning problem for this dataset as:

$$\begin{aligned}(\mathbf{w}^*, b^*) = &\arg\min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 \\ \text{subject to} \quad & -1(\mathbf{w}^T[0,0]^T + b) = 1 \\ & 1(\mathbf{w}^T[2,0]^T + b) = 1\end{aligned}$$

$$-1\left[[w_1 \ w_2]\begin{bmatrix}0\\0\end{bmatrix} + b\right] = 1$$

From the first constraint we see that:

$$0 \times w_1^* + 0 \times w_2^* - b^* = 1$$

from which we can conclude that $b^* = -1$. Plugging this into the second constraint, we see that:

$$\begin{aligned}2 \times w_1^* + 0 \times w_2^* - 1 &= 1 \Rightarrow \\ 2 \times w_1^* &= 2 \Rightarrow \\ w_1^* &= 1\end{aligned}$$

Thus, to minimize $||\mathbf{w}^*||$ we must set $w_2^* = 0$ and the optimal solution to this problem is:

$$(\mathbf{w}^*, b^*) = ([1,0]^T, -1).$$

$C \in \mathbb{R}^+$

5. [10 points (bonus)] Show that the Maximum a Posteriori (MAP) solution to a supervised learning problem is equivalent to the Maximum Likelihood solution if $p(\mathbf{w}) = C$ for some constant $C \in \mathbb{R}$.

**Solution:** We can begin from either formulation and arrive at equivalence with the other. Let's start from the Maximum Likelihood solution $\mathbf{w}_{\text{ML}}$ that maximizes the data likelihood:

multiply by constant doesn't change argmax

$$
\begin{aligned}
\mathbf{w}_{\text{ML}} &= \arg\max_{\mathbf{w}} p(\mathcal{D} \mid \mathbf{w}) \\
&= \arg\max_{\mathbf{w}} p(\mathcal{D}|\mathbf{w}) \frac{C}{p(\mathcal{D})}, \text{ (multiplying by constant in } \mathbf{f} \text{ won't change argmax)} \\
&= \arg\max_{\mathbf{w}} p(\mathcal{D}|\mathbf{w}) \frac{p(\mathbf{w})}{p(\mathcal{D})} \\
&= \arg\max_{\mathbf{w}} \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})} \\
&= \mathbf{w}_{\text{MAP}}.
\end{aligned}
$$

by assumption

$$\arg\max_{x} \left[ \frac{x^2 - 3}{10} \, 100 \right]$$