

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: _____

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
- (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 and Maximum likelihood 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?
☐ $c\mathbf{w}^*$ ☐ $c^2\mathbf{w}^*$ ☐ $\frac{1}{c^2}\mathbf{w}^*$ ☐ $\frac{1}{c}\mathbf{w}^*$

Total Question 1: 35

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 if $p(\mathbf{w}) = 1$.
 - ☐ They are equal if $p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \sigma)$ for very small σ .
 - ☐ They are never equal.
 - ☐ They are equal in the limit of infinite training samples.
- (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.
 - ☐ Train on less data.
 - ☐ Regularize the model.
- (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 $p(\mathbf{x} \mid \mathcal{C}_2) = \mathcal{N}(\boldsymbol{\mu}_2, \beta^{-1}I)$. 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:

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

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.

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 $C \in \mathbb{R}$.