FUNDAMENTALS OF MACHINE LEARNING
AA 2022-2023
Prova Intermedia

3 Novembre, 2022

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: The hard margin Support Vector Machine can be used to learn decision boundaries for non linearly separable datasets. ◯ True √ **False**

   Solution: The hard margin SVM will fail to converge on non linearly separable datasets.

   (b) [5 points] True or False: Logistic regression is a method for nonlinear regression. ◯ True √ **False**

   Solution: Logistic regression is a classification model.

   (c) [5 points] True or False: Least squares regression using a polynomial basis mapping results in a model that is nonlinear in the original input variables $\mathbf{x}$. √ **True** ◯ False

   Solution: Should be clear.

   (d) [5 points] True or False: Adding a zero-mean Gaussian Prior (i.e. $p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \sigma I)$) increases bias in our model. √ **True** ◯ False

   Solution: Adding a prior means we need to use the MAP solution which biases the model towards $\mathbf{w} = \mathbf{0}$ in this case.

   (e) [5 points] True or False: In the Primal Form of the SVM, decreasing hyperparameter C will increase bias in our model. √ **True** ◯ False

   Solution: If we decrease C we are giving less weight to making errors and more to minimizing $||\mathbf{w}||^2$, so more bias again towards $\mathbf{w} = \mathbf{0}$.

(f) [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 they are isotropic.
√ **None of the above.**

Solution: The quadratic Bayes discriminant makes no assumptions about the class-conditional covariance matrices.

(g) [5 points] In linear regression, adding an L2 regularizer is equivalent to:

√ **A Gaussian prior on model parameters.**
◯ Nonlinear regression.
◯ Polynomial regression.
◯ A uniform prior on model parameters.

Solution: See solution to bonus question.

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] Which of the following are true about the C hyperparameter in the SVM?

- ◯ As C approaches 0, the soft margin SVM approaches the hard margin SVM solution.
- ◯ C can be negative, as long as each of the slack variables are nonnegative.
- √ **As C approaches ∞, the soft margin SVM approaches the hard margin SVM solution.**
- ◯ None of the above.

Solution: The answer to this question is HIGHLY subjective and implementation dependent, so everyone received full credit for this question.

(b) [5 points] You are training a soft-margin SVM on a binary classification problem. You find that your model's training accuracy is very high, while your validation accuracy is very low. Which of the following are likely to improve your model's performance on the validation data?

- √ **Training your model on more data.**
- ◯ Increasing the hyperparameter C.
- ◯ Adding a nonlinear (e.g. quadratic) feature to each sample point.
- √ **Decreasing the hyperparameter C.**

Solution: If your training accuracy is high and validation accuracy low, they you are probably overfitting. The best solution is to collect more training data or decrease C so that model bias is increased (e.g. regularization).

(c) [5 points] What is the behavior of the width of the SVM margin ($\frac{1}{||w||}$) as $C \to 0$?

- ◯ No change.
- ◯ It goes to 0.
- √ **It goes to ∞.**
- ◯ None of the above.

Solution: If the regularization coefficient C goes to zero, the only term left in the objective is the minimization of $\frac{1}{2}||\mathbf{w}||^2$, so the margin goes to infinity.

(d) [5 points] Which of the following statements about SVMs are true?

- √ **If a finite set of training points from two classes is linearly separable, a hard-margin SVM will always find a decision boundary correctly classifying every training point.**
- ◯ If a finite set of training points from two classes is not linearly separable, a soft-margin SVM will always find a decision boundary correctly classifying every training point.
- √ **Every trained two-class hard-margin SVM model has at least one point of each class at a distance of exactly $\frac{1}{||\mathbf{w}||}$ from the decision boundary.**
- ◯ None of the above.

Solution: Shouldn't need much explanation, but a non-linearly separable dataset can never be correctly classified (unless maybe if we use the kernel trick).

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

- √ **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.

Solution: If you have little training data or low training error, increasing variance is a very bad idea. If the data is not linearly separable, you might want to increase variance via an explicit embedding or the kernel trick – the answer was considered correct whether or not the last choice was selected.

(f) [5 points] If we use a nonlinear basis mapping of our input variables for linear regression, which of the following are true?

√ **The resulting problem is linear in the model parameters.**

◯ The resulting problem is nonlinear in the model parameters.

√ **The resulting problem is nonlinear in the input variables.**

◯ None of the above.

Solution: An explicit embedding of the input data yields a solution that is still linear in the (increased number of) model parameters, but nonlinear in the original inputs.

(g) [5 points] Which of the following are true of K-fold cross-validation?

√ **We must fit K models to K different training datasets.**

◯ The average performance over the K folds is an upper bound on the true performance of our model.

√ **The average performance over the K folds is a lower bound on the true performance of our model.**

◯ None of the above.

Solution: K-fold cross validation gives us a lower bound on performance because if we add more training data the performance should improve (or at least not degrade). It requires us to fit the model K times, though.

Total Question 2: 35

3. [15 points] Consider the dataset:

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

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

Solution: We can use the embedding: $\phi([x_1, x_2]) = x_1 x_2$. Then the optimal hard-margin SVM problem is:

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

But these constraints then simplify to just (after substituting $\phi(\cdot)$ for all four samples):

$$\begin{aligned}w + b = 1 \\ w - b = 1\end{aligned}$$

Adding these two equations together gives us $w^* = 1$, and then substituting this into either gives $b^* = 0$. □

4. [15 points] Suppose that $(\mathbf{w}^*, b^*)$ is the solution to a two-class classification problem. Show that $(c\mathbf{w}^*, cb^*)$ for any $c \in \mathbb{R} \setminus \{0\}$ defines the same linear discriminant as the original solution.

Solution: The discriminant defined by $(\mathbf{w}^*, b^*)$ are all points $\mathbf{x}$ satisfying ${\mathbf{w}^*}^T\mathbf{x} + b^* = 0$. Now, for any $c \neq 0$:

$$\begin{aligned} c{\mathbf{w}^*}^T\mathbf{x} + cb^* &= 0 \iff \text{(by linearity)} \\ c({\mathbf{w}^*}^T\mathbf{x} + b^*) &= 0 \iff \text{(since } c \neq 0\text{)} \\ {\mathbf{w}^*}^T\mathbf{x} + b^* &= 0 \end{aligned}$$

And so the two discriminants are equal. □

5. [10 points (bonus)] Show that the Maximum a Posteriori (MAP) solution to linear regression with a Gaussian prior $p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \frac{1}{\beta} I)$ is equivalent to the regularized Maximum Likelihood Estimate (MLE) solution for a specific value of the regularization coefficient λ in the MLE solution.

   Hint: λ should be a function of β, and recall that the multivariate Gaussian density is:

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

Solution: For dataset $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$, the MAP solution is the $\mathbf{w}^*$ that maximizes the posterior:

$$\begin{aligned} p(\mathbf{w} \mid \mathcal{D}) &= \frac{p(\mathcal{D} \mid \mathbf{w}) p(\mathbf{w})}{p(\mathcal{D})} \\ &\propto \left\{ \prod_{n=1}^N \mathcal{N}(y_n \mid \mathbf{w}^T \phi(\mathbf{x}_n), \alpha^{-1}) \right\} \mathcal{N}(\mathbf{w} \mid \mathbf{0}, \beta^{-1}) \end{aligned}$$

We can take the logarithm of this and not change the maximum, so we have:

$$\begin{aligned} & \left\{ \sum_{n=1}^N \ln \mathcal{N}(y_n \mid \mathbf{w}^T \phi(\mathbf{x}_n), \alpha^{-1}) \right\} + \ln \mathcal{N}(\mathbf{w} \mid \mathbf{0}, \beta^{-1}) \\ = & -\frac{\alpha}{2} \sum_{n=1}^N (y_n - \mathbf{w}^T \phi(\mathbf{x}_n))^2 - \frac{\beta}{2} \mathbf{w}^T \mathbf{w} + \text{constants.} \end{aligned}$$

The maximizer of this will be the same as that of regularized least squares MAP solution for $\lambda = \beta/\alpha$:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \left\{ \frac{1}{2} \sum_{n=1}^N (y_n - \mathbf{w}^T \phi(\mathbf{x}_n))^2 - \frac{\beta}{\alpha} \mathbf{w}^T \mathbf{w} \right\}.$$

□

[THIS PAGE INTENTIONALLY LEFT BLANK]