DIPARTIMENTO DI ELETTRONICA,
INFORMAZIONE E BIOINGEGNERIA
Politecnico di Milano

Machine Learning (Code: 097683)

June 23, 2021

Surname:	
Name:	
Student ID:	
Row:	Column:

Time: 2 hours | Prof. Daniele Loiacono | Maximum Marks: 33

- The following exam is composed of **8 exercises** (one per page). The first page needs to be filled with your **name, surname and student ID**. The following pages should be used **only in the large squares** present on each page. Any solution provided either outside these spaces or **without a motivation** will not be considered for the final mark.
- During this exam you are **not allowed to use electronic devices**, such as laptops, smartphones, tablets and/or similar. As well, you are not allowed to bring with you any kind of note, book, written scheme, and/or similar. You are also not allowed to communicate with other students during the exam.
- The first reported violation of the above-mentioned rules will be annotated on the exam and will be considered for the final mark decision. The second reported violation of the above-mentioned rules will imply the immediate expulsion of the student from the exam room and the **annulment of the exam**.
- You are allowed to write the exam either with a pen (black or blue) or a pencil. It is your responsibility to provide a readable solution. We will not be held accountable for accidental partial or total cancellation of the exam.
- The exam can be written either in **English** or **Italian**.
- You are allowed to withdraw from the exam at any time without any penalty. You are not allowed to keep the text of the exam with you while leaving the room.
- **Three of the points will be given on the basis on how quick you are in solving the exam.** If you finish earlier than 45 min before the end of the exam you will get 3 points, if you finish earlier than 30 min you will get 2 points and if you finish earlier than 15 min you will get 1 point (the points cannot be accumulated).
- The box on Page 10 can only be used to complete the Exercises 7, and/or 8.

Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7	Ex. 8	Time	Tot.
/ 7	/ 7	/ 2	/ 2	/ 2	/ 2	/ 4	/ 4	/ 3	/ 33

Exercise 1 (7 marks)

17'5

Describe and compare ordinary least squares and Bayesian linear regression.

Exercise 2 (7 marks)

17'5

Describe and compare Q-learning and SARSA.

Exercise 3 (2 marks) 5'

```
1  train_accuracy = []
2  test_accuracy = []
3  for k in np.arange(1, 11):
4          knn = neighbors.KNeighborsClassifier(k)
5          knn.fit(X_train, y_train)
6          accuracy = sum(knn.predict(X_train) == y_train) / len(y_train)
7          train_accuracy.append(accuracy)
8          accuracy = sum(knn.predict(X_test) == y_test) / len(y_test)
9          test_accuracy.append(accuracy)
10 selected = np.argmax(test_accuracy)
```

1. Describe the operations performed by the Python code snippet above. Provide the detailed description of the lines of code provided below.
2. Which kind of problem are we solving? What is the technique we are using?
3. Do you think there are some problems with the procedure described above? Do you think there exists a better way to perform the procedure described above?

1.
   - Line 4 – 5 The code is training a KNN classifier on a training set of input `X_train` and output `y_train`. In principle, since KNN is a non-parametric method it does not require training. This is only a requirement by the Python method. This approach is done using different values for the parameter k denoting the neighbourhood of each point.
   - Line 6 – 7 Computing the accuracy of the trained model over the training data.
   - Line 8 – 9 Computing the accuracy of the trained model over the the `X_test` data.
   - Line 10 Selecting the best model having the largest accuracy on the test data.
2. This procedure is a model selection approach to determine the best value for the hyperparameter k in the KNN classifier.
3. The procedure overall is correct. The olny problem might be that we are somehow biasing our decision on the specific test set we used. Since the training procedure in this case is light (i.e., it is not required) one might use a LOO approach to solve the model selection problem.

Exercise 4 (2 marks) 5

Are the following statements about Linear Regression True or False? Motivate your answers.

1. Linear regression can be used to model processes showing a non-linear behavior.
2. In the case we have a small number of samples, the use of Bayesian Linear Regression is advised as a substitute for classical Linear Regression.
3. After performing Ridge Regression on a dataset with regularization parameter $\lambda = k$, the eigenvalues of the matrix $(\phi^\top \phi + \lambda I)$ are smaller or equal than k.
4. Gaussian Process can be reformulated as a specific case of Linear Regression over a specific set of features.

1. TRUE: using a set of nonlinear basis functions one might approcximate also nonlinear processes.
2. TRUE: in this case we might include the prior information we have in the Bayesian prior and update it basing on the data. It is known that Bayesian method are usually outperforming the frequentist counterparts in the case the sample are scarce.
3. FALSE: Ridge regression ensures that the eigenvectors of the resulting matrix are larger or equal than the value λ. Indeed, this method is used to avoid having an ill-conditioned design matrix.
4. TRUE: in this case the basis has an infinite number of features. In general using a valid kernel corresponds to the projection of our original dataset in a higher dimensional feature space.

Exercise 5 (2 marks) 5'

You want to apply RL to train an AI agent to play a single-player videogame. The state of the game is fully observable and, at each step, the agent has to select an action from a discrete set of possibilities. The interaction ends as soon as the agent reaches the end of the level or fails. To optimize the policy for your AI, you have a set of recorded trajectories (i.e., sequences of state, action, and reward) of the AI agent playing the game following a suboptimal policy. Unfortunately, most of these trajectories are not complete (i.e., they do not cover all the interactions from the beginning of the level to either the end, or to a game-over state).

Indicate if the following methods can be applied to this problem, motivating your answer.

1. Monte Carlo Policy Iteration;
2. Value Iteration;
3. Sarsa;
4. Q-Learning.

To solve provide a solution for the described scenario we need to solve a control problem. We need to use an online method since we do not have trajectories which are complete. Moreover, one might not resort to a dynamic programming approach, since we do not have a full description of the environment, but only trajectories. Finally, we need to have an offpolicy method since we can only rely on trajectories collected in the past. Therefore:

1. Not a viable option since Monte Carlo requires complete trajectories;
2. Not a viable option since we would require a complete description of the environment;
3. Not a viable option since it requires to follow the policy provided by it. (Using importance sampling might be a solution)
4. Viable option since it is offpolicy, online and uses only trajectories.

Exercise 6 (2 marks) 5[1]

Tell which of the following methods is a parametric method and which is not. Motivate your answers.

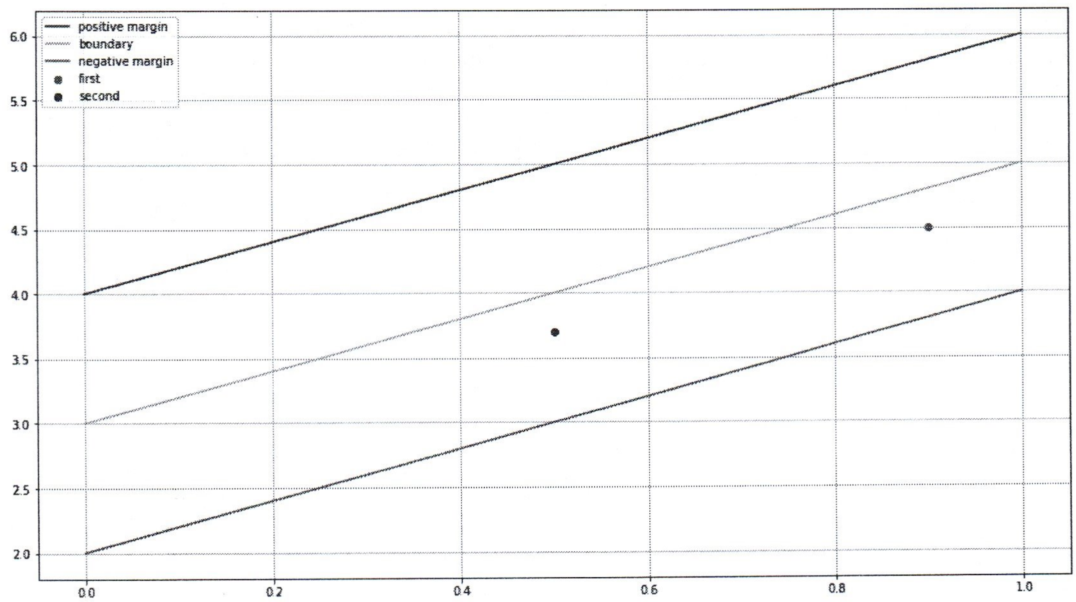
1. Gaussian Processes;
2. Logistic Regression;
3. Ridge Regression;
4. K-Nearest Neighbors.

1. NONPARAMETRIC: it require to store the initial data and compute the value of the kernel to provide a prediction;
2. PARAMETRIC: the final result of the optimization problem is the parameter vector of the weights $\mathbf{w}$. The training set can be discarded once we have this vector.
3. PARAMETRIC and NONPARAMETRIC: during the course we saw that this approach can be casted in both ways depending if the number of parameterts is large or small.
4. NONPARAMETRIC: the provide a prediction one needs the entire dataset. No data from the training can be discarded.

Exercise 7 (4 marks) 10'

Consider a linear, hard-margin, two-class SVM classifier defined by parameters $\mathbf{w} = [-2, 1]$, $b = -3$. Answer the following questions providing adequate motivations.

1. Provide the analytical formula of the boundary and the margins.
2. How is the point $\mathbf{x}_1 = [9/10; 9/2]$ classified according to the trained SVM?
3. Assume to collect a new sample $\mathbf{x}_2 = [1/2; 37/10]$ in the negative class. Do you need to retrain the SVM?
4. Which additional information would you require to classify the point $\mathbf{x}_1$ in the case we have a Kernel SVM classifier?



1.

$$
\begin{aligned}
-2 * x_1 + 1 * x_2 - 3 &= 0 \text{ (boundary)} \\
-2 * x_1 + 1 * x_2 - 4 &= 0 \text{ (positive margin)} \\
-2 * x_1 + 1 * x_2 - 2 &= 0 \text{ (negative margin)}
\end{aligned}
$$

2. $-2 \cdot 0.9 + 1 \cdot 4.5 - 3 = -0.3 \rightarrow$ classified as negative
3. $-2 \cdot 0.5 + 1 \cdot 3.7 - 3 = -0.3 \rightarrow$ classified as negative. The point is correctly classified, however, it is inside the margin, hence, we should retrain.
4. I would need the support vectors and the formula of the kernel, including all its hyperparameters.

Exercise 8 (4 marks) 10

1. Show that the VC-dimension of the class H of axis aligned rectangles is $VC(H) = 4$. Provide a proof resorting to textual and/or visual explanations.

2. How many samples do you need to guarantee that this classifier provides you with an error larger than $\varepsilon = 0.1$ with probability smaller than $\delta = 0.2$?

1. To show that $VC(H) \geq 4$ one might choose four points on a cross and show by enumeration that the set in the power set of these points can be shattered by the axis aligned rectangles. To show that $VC(H) < 5$ we need to consider three different cases. Assume that the classifiers is s.t. the positive instance should be included in the rectangle and the negative placed outside.
   (a) Three points are aligned: by alternating the labels on the aligned points and the points cannot be shattered.
   (b) Four points can be placed on the edges of an axis-aligned rectangle and the fifth one is inside: using a negative label on the inside point one should see that the set cannot be shattered.
   (c) The points are all on the edges of an axis-aligned rectangle: at least two points are on a single edge. Using alternating labels on the aforementioned points and one on an adjacent edge, provides evidence that these points cannot be shattered.

   Since no set of 5 points can be shattered, we proved that $VC(H) < 5$ and also the initial statement.

2. We have that we need a number of points:

$$N \geq \frac{1}{\epsilon}\left(4\log_2\left(\frac{2}{\delta}\right) + 8VC(H)\log_2\left(\frac{13}{\epsilon}\right)\right)$$
$$N \geq \lceil 10(4\log_2(10) + 32\log_2(75)\rceil$$
$$N \geq \lceil 40 \cdot 4 + 32 \cdot 7\rceil$$
$$N \geq 384$$