DIPARTIMENTO DI ELETTRONICA, INFORMAZIONE E BIOINGEGNERIA

Politecnico di Milano

Machine Learning (Code: 097683)
June 19, 2018

Name:	
Surname:	
Student ID:	
Row:	Column:

Time: 2 hours 30 minutes | Prof. Marcello Restelli | Maximum Marks: 34

- The following exam is composed of **10 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 outside these spaces will not be considered for the final mark.
- During this exam you are **not allowed to use electronic devices** like 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 allowed to leave the room not early than half the time of the duration of the exam. 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).**

Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7	Ex. 8	Ex. 9	Ex. 10	Time	Tot.
/ 5	/ 5	/ 5	/ 2	/ 2	/ 2	/ 2	/ 2	/ 3	/ 3	/ 3	/ 34

Exercise 1 (5 marks)

Describe the ridge regression algorithm and compare it with the Bayesian linear regression approach.

Exercise 2 (5 marks)

Define the VC dimension of a hypothesis space. What is the VC dimension of linear classifiers?

Exercise 3 (5 marks)

Describe which methods can be used to compute the value function V^π of a policy π in a discounted Markov Decision Process.

Exercise 4 (2 marks)

Consider the following code lines in MATLAB:

```
1 load iris_dataset;
2 irisInputs = zscore(irisInputs);
3 [loadings, scores, variance] = pca(irisInputs');
4 selPC = find(cumsum(variance) / sum(variance) > 0.99, 1);
5 pc = scores(:, 1:selPC);
```

Describe the process and purpose of what is implemented in this snippet. Tell if the method is sound or if it is necessary to modify the procedure to follow the classic ML guideline regarding this method.

The previous snippet is applying the Principal Component analysis to the iris dataset to perform dimensionality reduction.
At first, the data are loaded, at Line 1, and normalized (the mean is removed and are divided by the standard deviation), at Line 2. After that, the PCA is applied to the input, at Line 3. At Line 4 only the principal components providing at least 99% of the variance present in the input are selected and at Line 5 only the dimensions of the input corresponding to these components have been saved.
The procedure is overall correct, apart from the normalization part. Indeed, the data should have zero mean before performing the PCA, but it is not required to divide by the standard deviation, which could remove most of the information about the variance we want to exploit in the PCA. This is true only if the scales of data are not clearly different from each other.

Exercise 5 (2 marks)

For each one of the following statements, tell if it is true for the UCB1 and/or TS algorithms.

1. It relies on the assumption to know the family of the arms reward distributions (e.g., Bernoulli);
2. It modifies the statistics of all the arms;
3. It incorporates knowledge about the arms;
4. It is a randomized algorithm.

1. TS: we need to know a conjugate pair prior-posterior specific for the distribution of the rewards before executing the method.
2. UCB1: it updates the bound of all the arms, since in its numerator we have $logt$, t being the current round.
3. TS: by choosing the appropriate prior we are able to introduce some a priori information about the process in the prior.
4. TS: it extracts a sample from each posterior distribution of the arms at each turn.

Exercise 6 (2 marks)

Consider a linear regression with input x and target t.

1. Do you think there could be a problem if we consider as features x and $200x^2$?
2. Provide a technique to solve the problem.
3. Do you think there could be a problem if we consider as features x and $5x - 2$?
4. Propose a technique to overcome the problem of choosing the features one should include in the model.

Motivate your answers.

1. There might be problems in the scale of the two features, which might result way different in terms of magnitude from each others.
2. We might solve the problem by normalizing the features before performing the linear regression.
3. In this case, the two features are linearly dependent, therefore the corresponding design matrix $\Phi^T\Phi$ will be singular. To solve this issue one can remove one of the two features, since it does not provide any additional information to the regressor, or apply some regularization method, which is able to select automatically the features which are important to this problem.
4. One of the technique used not to test many feature choices is the use of Kernels, which are able to incorporate in the model a large number of features, without even having to make them explicit.

Exercise 7 (2 marks)

Tell if the following statements are true or false and motivate your answers.

1. Generally, first-visit estimation is better than every-visit if you use a small amount of episodes;
2. With MC estimation you can extract a number of samples for the value function equal to the length of the episode you consider for prediction;
3. Stochasticity in the rewards requires the use of a larger number of episodes to have precise prediction of the MDP value in the case we use MC estimation;
4. MC estimation works better than TD if the problem is not Markovian.

1. FALSE: MC every-visit is a biased estimator, but has lower variance than first-visit MC. This makes it suitable for situations in which we have a small amounts of episodes.
2. TRUE/FALSE: with MC every-visit you have a sample from each state/action pair we visit, with first-visit MC we have at most one sample per state.
3. TRUE: to reduce the uncertainty in estimators we need to reduce their variance, that decreases as the number of samples increases.
4. TRUE: TD exploits the Markovian property explicitly, therefore is better suited for Markovian problems.

Exercise 8 (2 marks)

Consider a classification problem. For each of the following characteristics, provide a classification algorithm.

1. Non-parametric;
2. Generative;
3. Kernel-based;
4. Provides nonlinear separating boundaries.

1. K-Nearest Neighbor: this method relied directly on the data we have to classify points. In this specific case we do not require any kind of training, since the method does not require to estimate any parameter.
2. Naive Bayes: it directly models the joint probability $p(x, t)$ of inputs and targets. This way it is able to provide samples of data similar to the one present in the training set.
3. SVM: they are naturally fitting kernel methods by the dual formulation of the optimization problem raising from the constrained minimization used to train the SVMs.
4. SVM with Gaussian Kernels of K-Nearest Neighbor: in this case the separating hyperplane is not linear anymore. The input space is divided into regions whose shape is not defined a priori to the training procedure.

Exercise 9 (3 marks)

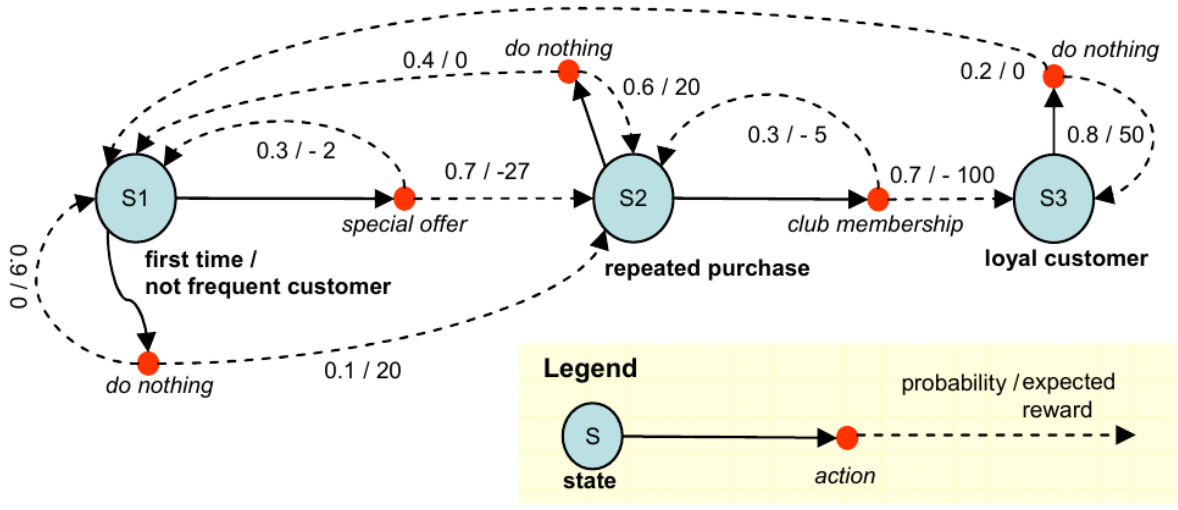
Consider the linear two-class SVM classifier defined by the parameters $w = [2\ 4]$ and bias $b = 2$. Answer the following questions providing adequate motivations.

- Give an example of a point which is in the positive class according to the SVM.
- How the point $x_2 = [3\ -1]$ is classified according to the trained SVM?
- Should I retrain the SVM if I add to the training set the point $x_1 = [-2\ 4]$, $t = 1$?

1. A point is in the positive class if $w^T x + b \geq 0$. Therefore $x = [0\ 0]$, having $w^T x + b = 2$, is classified in the positive class.
2. We have $w^T x_2 + b = 4$ and the point x_2 is in classified in the positive class.
3. One should retrain the SVM only if the point x_1 would have been a support vector in the current one. A point is a support vector if it is between the margins, or, formally, $|w^T x + b| \leq 1$. Since $|w^T x_1 + b| = 14$ we do not need to retrain the SVM.

Exercise 10 (3 marks)

Consider the following MDP.



- Provide the optimal policy for a discount factor of $\gamma = 1$;
- Provide the optimal policy for a discount factor of $\gamma = 0.5$ (you can justify your answer basing on what has been shown during the lectures and exercise sessions);
- Provide the equations which computes the state-value function of state $S2$ in the case we follow a policy:

$$(do\ nothing, club\ membership, do\ nothing)$$

and a discount factor of $\gamma = 0.1$ (you are not required to invert a matrix).

1. A discount factor of $\gamma = 1$ is not advisable for an infinite time horizon MDP, since it could lead to infinite reward. Moreover, using $\gamma = 1$ we are not able to provide a closed form solution to the Bellman equation. Therefore, the question is ill-posed.
2. We saw during classes that even using $\gamma = 0.9$ we would use the most myopic strategy $(do\ nothing, do\ nothing, do\ nothing)$. If we have an even smaller discount factor could only have an different optimal strategy.
3. $V = (I - \gamma P)^{-1} R = \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \frac{1}{10} \begin{bmatrix} 0.9 & 0.1 & 0 \\ 0 & 0.3 & 0.7 \\ 0.2 & 0 & 0.8 \end{bmatrix} \right)^{-1} \begin{bmatrix} 2 \\ -71.5 \\ 40 \end{bmatrix}$