DIPARTIMENTO DI ELETTRONICA, INFORMAZIONE E BIOINGEGNERIA

Politecnico di Milano

Machine Learning (Code: 097683)
July 4, 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 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** 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 logistic regression algorithm and compare it with the perceptron algorithm.

Exercise 2 (5 marks)

Describe the SVM algorithm for classification problems. Which algorithm can we use to train an SVM? Provide an upper bound to the generalization error of an SVM.

Exercise 3 (5 marks)

Describe what are eligibility traces and how they are used in the TD(λ) algorithm. Explain what happens when $\lambda = 0$ and when $\lambda = 1$.

Exercise 4 (2 marks)

Consider the following code lines in MATLAB:

```
load iris_dataset.mat
x = zscore(irisInputs(1,:)');
[t, ~] = find(irisTargets == 1); % 1 setosa, 2 virginica, 3
   versicolor
phi = zscore([x x.^2]);
lin_model = fitlm(x, t);
qua_model = fitlm(phi, t);
if qua_model.Rsquared.Adjusted > lin_model.Rsquared.Adjusted;
  y = predict(lin_model, x);
else
  y = predict(qua_model, x);
end
```

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.

This snippet performs classification on the Iris dataset. The script evaluates two different models, one linear in the original input and one with the addition of a quadratic feature, by looking at the adjusted R^2 of the fitted model.
There are two issues about the model we trained:

1. A model with larger adjusted R^2 is better, therefore the `if` conditions should be the other way round.
2. It is not a good idea to solve a classification problem with a regression one. This solution is not robust to outliers.

Exercise 5 (2 marks)

Answer the following questions about kernels. Motivate your answers.

1. Can you define a kernel over a feature set composed of propositional logic formulas? Provide an explicit example of kernel.
2. Assume to have a non-linearly separable dataset and you know which mapping is able to project the data in a linearly separable space. Are there still reasons to consider the use of kernels?
3. You decided to use a kernel approach. Do you prefer to have a large dataset or a small one?
4. Can you define a kernel over sets? Provide an explicit example of kernel.

1. YES: for instance we might count the number of operations to transform a logic formula into another and use this number as a similarity measure.
2. YES: the mapping, even if is known, might be too computationally heavy to be computed. Therefore, the use of kernel might reduce this complexity (a.k.a. kernel trick).
3. SMALL: the kernels are non-parametric method and their complexity depends on the dimension of the training set. None the less, if the dataset is too small we do not have enough information on the problem, thus we require it to be informative enough for the considered task.
4. YES: for instance, if A and B are two sets: $k(A, B) = 2^{|A \cap B|}$.

Exercise 6 (2 marks)

Which of the following statements are true? Provide motivations for your answers.

1. One advantage of using linear models is that the process generating data is often linear;
2. If the F-statistic of a linear regression is significant ($p \ll 1$), all the predictors have statistically significant effects;
3. In a linear regression with several variables, a variable has a negative regression coefficient if and only if its correlation with the response is negative.
4. Depending on the optimization approach (LS or likelihood maximization), you have different solutions for the optimal paramter vector w^*.

1. FALSE: usually the real life process are complex and linear models are a way of providing a simple approximation of such processes.
2. FALSE: the F statistics is low if at least one of the coefficient is significant, or equivalently it tests the current model versus the constant one.
3. FALSE: there might be spurious effect deriving with the interaction with other variables. Conversely, if the regression has a single input this statement is true.
4. TRUE/FALSE: if we assume to have a Gaussian heteroscedastic noise in the data, the two solutions coincides. If we assume a different distributions on the output we have different oprimal parameter vectors w^*.

Exercise 7 (2 marks)

Which method would you choose to solve the following **control problems**:

- Advertisement problem (the one with three states we saw during classes);
- Atari Games (hint: we are able to generate as many episodes as we want);
- Poker;
- Black Jack.

Motivate your answer.

The first problem has few states and the transition model and the rewards are known, therefore can be solved in a closed form. We might use the solution of the Bellman expectation equation and use brute force over all the possible policies.
In the other problems we do not have information about the transition (it is stochastic or unknown) or the rewards. Therefore, we need to use some RL control technique, like SARSA or Q-learning.

Exercise 8 (2 marks)

Consider the following statements and tell if they are true or false. Motivate your answers.

1. The computation of the bias-variance decomposition is possible only theoretically. No algorithm provides an explicit decomposition of the twos.
2. An error which is comparable on the training and the test, but larger than what is required by the application, means that the used method has a large variance.
3. The cross-validation error provides slightly larger estimates of the prediction error on newly seen data.
4. If a model results in being too complex, to solve the problem we need to carefully remove some of the input features.

1. TRUE: it is possible only if we know the true process. Some method have an explicit decomposition (e.g., KNN).
2. FALSE: if the two errors are comparable the variance of the model is low. Conversely, since we are not able to reach the desired performance it might be because the real process is more complex than the model. Therefore, we might be in a case where a large bias is present.
3. TRUE: it is a slightly pessimistic estimates of the test error.
4. FALSE: this is an option. We might also resort to regularization (e.g., ridge regression) or to feature extraction (e.g., PCA).

Exercise 9 (3 marks)

Evaluate the value for the MDP with four states $\mathcal{S} = \{A, B, C, D\}$ (D is terminal), two actions $\mathcal{A} = \{h, r\}$ given the policy π, given the following trajectories:

$$(A, h, 3) \to (B, r, 2) \to (C, h, 1) \to (D)$$
$$(C, h, 2) \to (A, h, 1) \to (D)$$
$$(B, r, 1) \to (A, h, 1) \to (D)$$

1. Do you think that a total reward maximization ($\gamma = 1$) is possible in this MDP?
2. Compute the approximation of the state-value function of the MDP by using MC first-visit and every-visit.
3. Assume to consider a discount factor $\gamma = 0.5$. Compute the state-value function by resorting to TD(0). Assume to start from zero values for each state and $\alpha = 0.5$.

1. The MDP corresponding to the provided episodes might have either a finite time horizon or an indefinite time horizon (there is no way of discriminating between the two cases by looking at a finite number of episodes). In the former case the use of a total reward maximization is a viable option, in the latter one no.
2. Since there are no repeated states the FV and EV Monte Carlo estimation coincides.

$$V(A) = \frac{6+1+1}{3} = \frac{8}{3} \qquad V(B) = \frac{3+2}{2} = \frac{5}{2} \qquad V(C) = \frac{1+3}{2} = 2$$

3. The formula for the $TD(0)$ update is the following:

$$V(S) \leftarrow V(S) + \alpha(R(S) + \gamma V(S') - V(S))$$

where S' is the next state we visit. Therefore:

$$V(A) \leftarrow 0 + 0.5(3 + 0.5 \cdot 0 - 0) = \frac{3}{2}$$
$$V(B) \leftarrow 0 + 0.5(2 + 0.5 \cdot 0 - 0) = 1$$
$$V(C) \leftarrow 0 + 0.5(1 + 0.5 \cdot 0 - 0) = \frac{1}{2}$$
$$V(C) \leftarrow \frac{1}{2} + 0.5\left(2 + 0.5 \cdot \frac{3}{2} - \frac{1}{2}\right) = \frac{13}{8}$$
$$V(A) \leftarrow \frac{3}{2} + 0.5\left(1 + 0.5 \cdot 0 - \frac{3}{2}\right) = \frac{5}{4}$$
$$V(B) \leftarrow 1 + 0.5\left(1 + 0.5 \cdot \frac{5}{4} - 1\right) = \frac{21}{16}$$
$$V(A) \leftarrow \frac{5}{4} + 0.5\left(1 + 0.5 \cdot 0 - \frac{5}{4}\right) = \frac{9}{8}$$

Exercise 10 (3 marks)

Starting from the formula of the softmax classifier:

$$y_k(x) = \frac{\exp(w_k^T x)}{\sum_j \exp(w_j^T x)},$$

derive the formula for the sigmoid logistic regression for the two classes problem (with a single parameter w and a single target t).

After that, assume to have a dataset composed of N samples (x_n, y_n). Derive the gradient of the following loss function for the prediction of a logistic regression t_n:

$$L(w) = -\ln\left(\prod_{n=1}^{N} y_n^{t_n}(1-y_n)^{1-t_n}\right).$$

See the theory slides.