DIPARTIMENTO DI ELETTRONICA,
INFORMAZIONE E BIOINGEGNERIA

Politecnico di Milano

Machine Learning (Code: 097683)
July 6, 2017

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 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 supervised learning technique denominated **Support Vector Machines** for classification problems.

Exercise 2 (5 marks)

Define the **VC dimension** and the describe the importance and usefulness of VC dimension in machine learning.

Exercise 3 (5 marks)

Describe the differences existing between the Q-learning and SARSA algorithms.

Exercise 4 **(2 marks)**

Categorize the following ML problems (supervised, unsupervised, RL):

1. Detect dangerous materials by means of an X-ray machine;
2. Determine which kind of users are accessing a server;
3. Teach a robot how to play tennis;
4. Predict the final degree of a student given only few exam marks.

For each one of them suggest **a set of features** that might be useful to solve the problem and **a method** to solve it.

1. Supervised: you have to classify the materials as dangerous/not dangerous. We could use logistic regression as method and material density and/or shape as features;
2. Unsupervised: if you do not know the classes i.e., the groups, you should determine them. We could use a clustering technique and use gender, interests, age as features;
3. RL: we need to provide the robot a policy which by basing on the current state selects a specific action. The methods we might use are Q-learning or SARSA and the features might be the position of the two players on the field and the ball position and speed;
4. Supervised: given the record of other students I want to predict the outcome of a new student. This is a regression problem, thus we can use linear regression and use the exam marks as input and the final degree as output.

Student's name:

Exercise 5 (2 marks)

For which dataset characteristics (e.g., large/small,) and which model (e.g., complex/simple) would you consider the following techniques for model selection?

1. Validation;
2. Crossvalidation;
3. AIC or BIC index;
4. Leave one out.

Justify you choices.

1. Validation: large dataset and complex model, if we are given a large dataset but the training of the model is complex, we might consider only to test it on an independent dataset to select the model hyperparameters;
2. Crossvalidation: large dataset and simple model, in the case we can multiply train the model without a large amount of computation, we might divide it into k folds and apply validation on each one of them;
3. AIC or BIC index: small dataset and complex model, the computation of these indexes does not require to retrain the model, is computed over the training data and automatically penalizes complex models;
4. Leave one out: small dataset and simple model, this method requires a high computational cost and thus it is feasible only if we have small dataset and the training time is limited.

Exercise 6 (2 marks)

Tell if the following statements are TRUE or FALSE. Motivate your answers.

1. Starting from any value function, we are not assured to converge to a solution when we apply repeatedly the Bellman optimality operator;
2. The closed form solution to the Bellman equation is always a good choice to compute the value function of an MDP;
3. The application of the Bellman optimality operator T for 5 times to a generic value function V_0 guarantees that $||V^* - T^5 V_0||_\infty \leq \gamma^3 ||V^* - V_0||_\infty$;
4. Starting from any value function, we are assured to converge to a solution when we apply repeatedly the Bellman expectation operator.

1. FALSE: the Bellman equation has a single fixed point and its recursive formulation is a contraction, thus no matter where we start we are assured to converge if we apply the operator enough times;
2. FALSE: if the MDP is too large it requires to invert a large matrix, which could be not feasible;
3. TRUE: we are assured that $||V^* - T^5 V_0||_\infty \leq \gamma^5 ||V^* - V_0||_\infty \leq \gamma^3 ||V^* - V_0||_\infty$ since $\gamma \leq 1$;
4. TRUE: the same reasoning for the optimality operator holds for the expectation operator.

Exercise 7 (2 marks)

Suppose you want to use a GP for a regression problem. You know that the input data varies a lot along some dimensions and less along others.

1. Provide the analytic form of a kernel suitable for this situation and motivate why you would choose it.
2. Do there exist other techniques that are able to handle this problem?
3. Why should not you consider such a model in the case you have the information that the dispersion is the same along all the input directions?
4. Do you think this problem could be solved by changing the usual prior distribution (i.e, Gaussian with zero mean and constant variance over the input space)?

1. A possible choice could be: $k(\mathbf{x}, \mathbf{x}') = \prod_k \exp\left\{\frac{(x_k - x'_k)}{2\sigma_k^2}\right\}$ where x_k and x'_k are the k-th component of the vectors $\mathbf{x}$ and $\mathbf{x}'$ respectively;
2. We might Z-score the input data, this way all the input would be of the same magnitude;
3. In the case we are using a model with more parameters and they are not needed we are more prone to overfitting, thus we should avoid using complex models when there is no necessity to do so;
4. With the prior we are regulating the GP in a specific point, while there is no way of determining the correlation two near points have by using the prior. Thus, the correct way of introducing this information is with the use of a different kernel.

Exercise 8 (2 marks)

Provide an example for which a policy considering lower confidence bounds to the expected reward of each arm is failing to converge to the optimum in a MAB setting with Bernoulli rewards. More specifically, this algorithm chooses the arm by selecting the one with the largest:

$$\hat{R}_t(a_i) = \frac{1}{N_t(a_i)} \sum_{j=1}^{t} r_{i,j} \, \mathbb{I}\left\{a_i = a_{i_j}\right\} - \sqrt{\frac{2\log(t)}{N_t(a_i)}},$$

where $r_{i,j}$ is the reward for arm a_i at time j, $N_t(a_i)$ is the number of times the algorithm selected arm a_i, t is the current round and $\mathbb{I}$ is the indicator function.

How often does the algorithm does not converge?

Even if we only rely on the empirical mean we are not assured to converge, thus using an even pessimistic index we incur in similar cases. The simplest example is when we have a negative outcome for the optimal arm in its first pulls and positive ones for all the other ones. In this case we would have negative index $\hat{R}_t(a_i)$ for the optimal arm (we need $t > 2K \log t$ total pulls to ensure that all the other arms have positive index). Here I am assuming that if the arms have negative value I just pull them randomly.

This event occurs with probability $(1-\mu^*)^{\frac{t}{K}} \prod_{k \neq k^*} \mu_k^{\frac{t}{K}}$, where μs are the expected values of the arms. Thus the convergence will fail with even higher probability.

Exercise 9 (3 marks)

Consider the following results from a `Matlab` script performing linear regression:

```
Linear regression model:
    y ~ 1 + x1 + x2
Estimated Coefficients:
                   Estimate        SE        tStat      pValue
                   ---------    --------    --------    -------

    (Intercept)    -0.025569    0.042526    -0.60126     0.5486
    x1               0.42564     0.76242     0.55828     0.5775
    x2               0.26882     0.38142     0.70478    0.48206
Number of observations: 150, Error degrees of freedom: 147
Root Mean Squared Error: 0.272
R-squared: 0.927,  Adjusted R-Squared 0.926
F-statistic vs. constant model: 935, p-value = 2.45e-84
```

1. Do you think that all the considered features are significant for the problem?
2. Do you think that at least one feature is significant for the analysed problem?
3. What might be the problem of this linear model?

1. We cannot tell from the specific p-values for each one of the input, but it is likely that there are some input variables which might be not significant;
2. Since the p-value for the F-test is low there is statistical evidence that at least one of the input is significant for the problem (at least more significant than using the empirical mean);
3. Probably the two variables are highly correlated, thus the use of a single input could be enough for this problem.

Exercise 10 (3 marks)

Starting from the formula of the softmax classifier for k classes:

$$y_k(\mathbf{x}) = \frac{\exp(\mathbf{w}_k^T \mathbf{x})}{\sum_j \exp(\mathbf{w}_j^T \mathbf{x})}$$

derive the formula for the sigmoid logistic regression parameter $\mathbf{w}$ for the two-class problem.

Assume that the estimated parameter is $\mathbf{w} = [4\ 2\ 1]$ and the input vector is of the form $\mathbf{x} = [x_1\ x_2\ 1]$. Draw the boundary of the logistic regression in the input space and in the parameter space.

Since we are considering only two classes we have that summation is only over two parameter vectors $\mathbf{w}_1$ and $\mathbf{w}_2$. if we consider class C_1 we may write:

$$\begin{aligned} y_1(x) &= \frac{\exp(\mathbf{w}_1^T x)}{\exp(\mathbf{w}_1^T x) + \exp(\mathbf{w}_2^T x)} \\ &= \frac{\frac{\exp(\mathbf{w}_1^T x)}{\exp(\mathbf{w}_1^T x)}}{\frac{\exp(\mathbf{w}_1^T x)+\exp(\mathbf{w}_2^T x)}{\exp(\mathbf{w}_1^T x)}} \\ &= \frac{1}{1+\exp[(\mathbf{w}_2 - \mathbf{w}_1)^T x]} \end{aligned}$$

The boundary in the parameter space is the point $[4\ 2\ 1]$ in a 3D space, while in the input space it is the line $\mathbf{w}^T \mathbf{x} = 0 \rightarrow x_2 = -2x_1 - \frac{1}{2}$.