DIPARTIMENTO DI ELETTRONICA, INFORMAZIONE E BIOINGEGNERIA

Politecnico di Milano

Machine Learning (Code: 097683)

July 28, 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 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 Gaussian Processes model for regression problems.

Exercise 2 (5 marks)

Describe the value iteration algorithm. Does the algorithm always return the optimal policy?

Exercise 3 (5 marks)

Describe the UCB1 algorithm. Is it a deterministic or a stochastic algorithm?

Exercise 4 (2 marks)

After training a logistic regression classifier with gradient descent on a given dataset, you find that it does not achieve the desired performance on the training set, nor the cross-validation one. Which of the following might be a promising step to take?

1. Use an SVM with a linear kernel.
2. Use an SVM with a quadratic kernel.
3. Add a set of features by basing on prior information on the problem.
4. Use lasso regularization.

Comment each one of the above choices.

1. NO: an SVM with a linear kernel is equivalent to a logistic regressor classifier, thus both the errors should not change with this modification.
2. YES: if we introduce new features it is likely that the training error would decrease. If the added quadratic features are also meaningful for the considered problem, we would also have a decrease in the test error.
3. YES: it it always a good idea to add a set of features by basing on prior information on the problem, since we are likely to help the model to better approximate the real process.
4. NO: since both the training and the test error are above what we expected, we are likely to need a larger amount of features, thus the use of lasso regularization (which in principle sparsifies the available set of features) would not give any improvement to the model.

Exercise 5 (2 marks)

Consider a linear regression with input x, target t and optimal parameter θ^*.

1. What happens to a model that uses as input variables x and $7x + 2$?
2. What do we expect on the uncertainty about the parameters we get by considering as input variables x and $7x + 2$?
3. What happens to a model that uses as input variables x and $x^2 - x$?
4. Is it possible that for some dataset (x, t) I get a lower training MSE if I use $\theta \neq \theta^*$ for prediction?

Motivate your answers.

1. The design matrix (i.e., $X^T X$) would be singular, thus its inversion would not be possible. This is because the introduction of a new linearly-dependent feature does not give any additional information to the model.
2. The parameter we estimate from the linear regression is likely to have high uncertainties, thus the p-values corresponding to the test $w_i = 0$ vs. $w_i \neq 0$ will be large.
3. This is a viable option for a linear model, i.e., to add polynomial features, to try to capture the real behaviour of the process.
4. If we used the closed form solution we are guaranteed that θ^* is the unique parameter to minimize the training MSE. In the case of singular design matrix we would have an infinite number of parameters, all giving the same MSE.

Exercise 6 (2 marks)

Consider the UCB1 and the Thompson Sampling algorithms. Tell if the following statements are true for the two aforementioned algorithms and motivate your answers.

1. Requires the knowledge of a pair of conjugate prior/posterior distributions;
2. Manages automatically the exploration/exploitation tradeoff;
3. Relies on the "optimism in the face of uncertainty" paradigm;
4. Is able to incorporate a priori knowledge about the problem.

1. TS: it relies on a Bayesian framework, it requires to have a prior and a way of updating this prior. One of the most common way is to use conjugate prior/posterior distributions.
2. TS and UCB1: all the MAB algorithm are specifically designed to minimize the regret and thus they manage automatically the exploration/exploitation tradeoff.
3. UCB1: it is based on the computation of the Hoeffding upper confidence bounds, which provides an optimistic estimates of the expected reward of the arms.
4. TS: since it is a Bayesian method, it can be naturally extended to include information about the problem by the modification of the prior.

Exercise 7 (2 marks)

Consider $x, y \in \mathbb{R}^d$ and assume that $k(x, y)$ is a valid kernel. Tell if the following are valid kernels.

1. $k(x, y) = sin(x)k(x, y)sin(y)$ $(d = 1)$;
2. $k(x, y) = ck(x, y) + 3$;
3. $k(x, y) = x^T Ay$ with $A = \begin{bmatrix} 4 & 3 \\ 3 & -6 \end{bmatrix}$ $(d = 2)$;
4. $k(x, y) = \begin{cases} x^T y & x < 0, y < 0 \\ (x^T y)^2 & x < 0, y > 0 \\ 1 - e^{x^T y} & x > 0, y < 0 \\ sin(x^T y) & x > 0, y > 0 \end{cases}$.

Motivate your answers.

From the Mercers theorem, any continuous, symmetric, positive semi-definite kernel function can be expressed as a dot product in a high-dimensional space. Moreover, it is possible to start from already valid kernels and transform them to have still valid kernels.

1. Yes, since it is a valid transformation of an already valid kernel (it would have been a valid kernel $f(x)k(x, y)f(y)$ for any $f(\cdot)$);
2. for some values of c it is a valid kernel, for instance if $c > 0$;
3. No, since one of the eigenvalues of the matrix A is negative the determinant of the matrix $Det(A) = \lambda_1\lambda_2 = 4(-6) - -3\ 3 = -33$.
4. No, since it is not symmetric, e.g., $k(-1, 2) = 4$ and $k(2, -1) = 1 - e^{-2}$.

Exercise 8 (2 marks)

Consider the following statement regarding PCA and tell if they are true or false. Provide motivations for your answers.

1. PCA is susceptible to local optima, thus trying multiple random initializations for the process of finding the principal components might help;
2. If all the input features are on very similar scales, we should not perform mean normalization (so that each feature has zero mean) before running PCA;
3. Given the scores $\tilde{x}_i \in \mathbb{R}^k$ and the loadings W, it is possible to reconstruct perfectly $x_i \in \mathbb{R}^d$;
4. PCA can be used for unsupervised feature extraction, for data visualization and data compression.

1. FALSE: the process of computing the PCA, in its original implementation, is a deterministic inversion of a matrix.
2. FALSE: it is always a good idea to do mean normalization, otherwise we could get "weird" principal components, for instance we could get a first principal component near to $e_1 = (1, 0)$ in 2D if the mean of the first component is way larger than the second one.
3. TRUE/FALSE: if $k = d$ yes, since we kept all the information from the original dataset. If $k \leq d$ we might have some reconstruct error;
4. TRUE: if we keep only $k \leq d$ principal components we might used it as an unsupervised feature extraction technique, if we keep $k \leq 3$ we can also visualize the principal component and since by keeping $k < d$ we would have a good representatives for the initial dataset, it can also used for data compression.

Exercise 9 (3 marks)

Assume to have two different linear models (with the intercept) working on the same dataset of $N = 206$ samples.

- The first model has $k_1 = 5$ inputs, considers linear features and has a residual sum of squares of $RSS_1 = 0.6$ on a validation set;
- The second model has $k_2 = 10$ inputs, considers only quadratic features (e.g., x_1^2 and $x_1 x_2$) and has a residual sum of squares of $RSS_2 = 0.1$ on a validation set;

Would you choose the second over the first one? Why? Recall that the F-test statistic is:

$$\hat{F} = \frac{N - p_2}{p_2 - p_1} \frac{RSS_1 - RSS_2}{RSS_2} \sim F(p_2 - p_1, N - p_2),$$

where p_1 and p_2 are the two parameters of the two models and $F(a, b)$ is the Fisher distribution with a and b degrees of freedom and that the 99th quantile of the Fisher distribution is $F_{99}(50, 150) = 1.6648$.

We have that the first model has $p_1 = 5 + 1$ parameters, while the second one has $p_2 = \frac{10(10-1)}{2} + 10 + 1 = 56$, thus:

$$\hat{F} = \frac{206 - 56}{56 - 6} \frac{0.6 - 0.1}{0.1} = 15 \gg 1.6648,$$

thus there is statistical evidence at level at least 0.99 that the second model is better than the first one.

Exercise 10 (3 marks)

Assume to have an MDP with four states $\mathcal{S} = \{H, M, L, F\}$ (F is terminal), two actions $\mathcal{A} = \{r, w\}$ and a discount factor $\gamma = 1$. Given the following trajectories:

$$(H, r, 2) \to (L, r, 3) \to (M, r, 2) \to (F)$$
$$(H, w, 2) \to (H, r, 3) \to (M, w, 1) \to (F)$$

1. Compute the values of the different states by resorting to first-visit and every-visit MC.
2. Using a learning rate $\alpha = 0.5$, compute the state values by resorting to TD. Assume to start from zero values for each state.
3. Can you tell if the previously defined MDP is deterministic or stochastic?

MC First visit	MC Every visit
$V(H) = \frac{7+6}{2} = \frac{13}{2}$	$V(H) = \frac{7+6+4}{3} = \frac{17}{3}$
$V(L) = 5$	$V(L) = 5$
$V(M) = \frac{2+1}{2} = \frac{3}{2}$	$V(M) = \frac{2+1}{2} = \frac{3}{2}$

The TD update rule is:

$$V(s_t) \leftarrow V(s_t) + \alpha(R_t + \gamma V(s_{t+1} - V(s_t)) = V(s_t)$$
$$V(s_t) \leftarrow V(s_t) + 0.5(R_t + V(s_{t+1} - V(s_t))$$

thus we have:

$$V(H) \leftarrow 0 + 0.5(2 + 0 - 0) = 1$$
$$V(L) \leftarrow 0 + 0.5(3 + 0 - 0) = 1.5$$
$$V(M) \leftarrow 0 + 0.5(2 + 0 - 0) = 1$$
$$V(H) \leftarrow 1 + 0.5(2 + 1 - 1) = 2$$
$$V(H) \leftarrow 2 + 0.5(3 + 1 - 2) = 3$$
$$V(M) \leftarrow 1 + 0.5(1 + 0 - 1) = 1$$

Since there is a state action pair with different rewards i.e., (H,r), the MDP can not be deterministic.