Instituto Superior Técnico

Aprendizagem 2021/22

Exam 2

25 February 2022

- A duração deste exame é de 2 horas, com uma tolerância adicional de 30 minutos.
- Desistências e entregas após os primeiros 90 minutos.
- É permitida apenas a consulta de uma folha (frente e verso).
- O número total de pontos é 20.
- Marque as suas respostas NA FOLHA DE EXAME.
- Escreva o seu número e nome no topo de cada página.
- Escreva todas as fórmulas. Apresente *todos* os cálculos e justificações para as suas respostas.

- The exam duration is 2h with an additional 30 minutes tolerance. *Panic not.*
- Withdrawals and deliveries after the first 90 minutes.
- The consultation of a single sheet note is allowed.
- 20 points total.
- Mark your answers ON THE EXAM ITSELF.
- Write your number and name at the top of each page.
- Write all formulas. Present *all* the computations and justifications for your answers.

Não preencher: para o uso oficial / not fill: for official use only

1	**2**	**3**	**4**	**5**	**SUM**
5v	6v	5.5v	2v	1.5v	20v

1. [5 pts] **Neural Networks**

Consider a MLP classifier characterized by the following weights

$$W^{[1]} = \begin{pmatrix} 2 & 0 & -3 & 1 \\ -1 & 1 & 1 & -1 \\ 1 & -1 & 2 & -2 \end{pmatrix}, \quad b^{[1]} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad W^{[2]} = \begin{pmatrix} -0.1 & 1 & 0 \\ -2 & 0 & -1 \end{pmatrix}, \quad b^{[2]} = \begin{pmatrix} 0.65 \\ 0.40 \end{pmatrix}$$

and activation function $f(x) = \frac{1}{1+e^{-2x}} = \text{sigmoid}(2x) = \sigma(2x)$ for every unit of the hidden layer.

Consider the example $\mathbf{x} = (1,1,1,1)^{\mathrm{T}}$, $z = (1,0)^{T}$ and the cross-entropy error.

To make the math easier, consider $\sigma(2.2) = 0.9$ and $\sigma(-\mathrm{x}) = 1 - \sigma(\mathrm{x})$.

a) [1v] Do forward propagation

b) [1v] Compute $\delta^{[2]} = \frac{\partial \text{Error}}{\partial \text{NET}^{[2]}}$

c) [0.5v] Update $\mathrm{W}^{[2]}$ using a learning rate of 0.1

d) [1v] Compute $\delta^{[1]} = \frac{\partial \text{Error}}{\partial \text{NET}^{[1]}}$

e) [0.5v] Update $b^{[1]}$ using a learning rate of 0.1

f) [1v] Under certain conditions, the phenomenon of "*vanishing gradients*" appears when learning starts progressing very slowly. Looking at $\delta^{[2]}$, what are these conditions? What can we change in the architecture to avoid this? Answer each question with math and a short sentence.
Hint: look into the limits of $\delta^{[2]}$ for an increasing $NET^{[2]}$

2. [6 pts] **Clustering**, **regression and PCA**

Consider the following observations in a Euclidean space:

	y_1	y_2	z
$\mathbf{x}_1$	0	0	0.1
$\mathbf{x}_2$	1	2	0.4
$\mathbf{x}_3$	2	2	0.4
$\mathbf{x}_4$	4	4	0.9

a) [1.5v] Using input variables, identify the k-means clustering solution with $k = 3$, and $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$ observations as initial centers (seeds). How many iterations are necessary for convergence? Show the results per iteration.

Hint: you can compute k-means visually, indicating the results per iteration

b) [1.5v] Compute the silhouette of the larger cluster.

c) [1v] Estimate the quantities of training observations, $\hat{z} = g(\mathbf{x})$ where g is given by a linear regression model with $\phi(\mathbf{x}) = \|\mathbf{x}\|_1$ and $\mathbf{w} = (-0.2, 0.2)^T$

d) [0.5v] Using the estimates, identify the training root mean squared error (RMSE) of g.

e) [1.5v] The following covariance matrix and eigenvectors were produced for the given dataset:

$$C = \begin{pmatrix} 2.917 & 2.667 \\ 2.667 & 2.667 \end{pmatrix}, \quad \mathbf{u}_A = \begin{pmatrix} -0.690 \\ 0.723 \end{pmatrix}, \quad \mathbf{u}_B = \begin{pmatrix} 0.723 \\ 0.690 \end{pmatrix}$$

Which eigenvector, $\mathbf{u}_A$ or $\mathbf{u}_B$, explains more data variability?

3. [5.5 pts] **Bayes and tree learning**

Consider a decision tree learned from a dataset with two binary input variables and 10 observations. The following association rules define this decision tree:

- If $y_1 = 0$ then $class = Positive$ (3 observations correctly predicted at this leaf)
- If $y_1 = 1 \wedge y_2 = 0$ then $class = Positive$ (2 observations correctly predicted and 2 observations incorrectly predicted at this leaf)
- If $y_1 = 1 \wedge y_2 = 1$, then $class = Negative$ (3 observations correctly predicted at this leaf)

a) [1.5v] Compute the confusion matrix and precision of the provided decision tree.

b) [2v] Given $\mathbf{x} = [1,0]^T$, estimate $P(\mathbf{x} \mid Positive)$.
Hint: recover the original data.

c) [1v] Classify $\mathbf{x} = [0,0]^T$ using kNN with Hamming distance, $k = 7$ and uniform weights.

d) [1v] Is the given data separable by a SVM with an RBF kernel? Justify

4. [2 pts] **Model complexity**

Consider the regression problem of estimating the spam degree of a document described by m features.

Estimate the complexity of a cluster-based RBF network, and a MLP with two hidden layers of 10 nodes each. Both networks have RELU activations on hidden and output nodes.

Identify the number of clusters k that guarantees that the RBF network has lower (parameter) complexity than the MLP.

5. [1.5 pts] **Deep learning**

Consider a multilayer perceptron for classification with H hidden layers. Assume that all weights and biases are initialized to zero. Now consider the following scenarios:

- **i.** all activation functions are equal to $f(net) = net^2$ and the error function is squared error
- **ii.** all activation functions are equal to $f(net) = e^{net}$ and the error function is squared error

For each scenario, what do you predict will happen? Is learning possible? Why?
Justify by computing the deltas and one short sentence.

END