TÉCNICO LISBOA

Aprendizagem 2022/23

Homework III

Deadline 28/10/2022 23:59 via Fenix as PDF

- Submission Gxxx.PDF in Fenix where xxx is your group number. Please note that it is possible to submit several times on Fenix to prevent last-minute problems. Yet, only the last submission is considered valid
- Use the provided report template. Include your programming code as an Appendix
- Exchange of ideas is encouraged. Yet, if copy is detected after automatic or manual clearance, homework is nullified and IST guidelines apply for content sharers and consumers, irrespectively of the underlying intent
- Please consult the FAQ before posting questions to your faculty hosts

I. Pen-and-paper [12v]

Consider the problem of learning a regression model from 5 univariate observations $\big((0.8),(1),(1.2),(1.4),(1.6)\big)$ with targets (24,20,10,13,12).

1) [5v] Consider the basis function, $\phi_j(x) = x^j$, for performing a 3-order polynomial regression,

$$\hat{z}(x, \mathbf{w}) = \sum_{j=0}^{3} w_j \phi_j(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 \,.$$

Learn the Ridge regression (l_2 regularization) on the transformed data space using the closed form solution with $\lambda = 2$.

Hint: use `numpy` matrix operations (e.g., `linalg.pinv` for inverse) to validate your calculus.

2) [1v] Compute the training RMSE for the learnt regression model.
3) [6v] Consider a multi-layer perceptron characterized by one hidden layer with 2 nodes. Using the activation function $f(x) = e^{0.1x}$ on all units, all weights initialized as 1 (including biases), and the half squared error loss, perform one batch gradient descent update (with learning rate $\eta = 0.1$) for the first three observations (0.8), (1) and (1.2).

II. Programming and critical analysis [8v]

Consider the following three regressors applied on `kin8nm.arff` data (available at the webpage):

- linear regression with Ridge regularization term of 0.1
- two MLPs – MLP_1 and MLP_2 – each with two hidden layers of size 10, hyperbolic tangent function as the activation function of all nodes, a maximum of 500 iterations, and a fixed seed (`random_state=0`). MLP_1 should be parameterized with early stopping while MLP_2 should not consider early stopping. Remaining parameters (e.g., loss function, batch size, regularization term, solver) should be set as default.

Using a 70-30 training-test split with a fixed seed (`random_state=0`):

4) [4v] Compute the MAE of the three regressors: linear regression, MLP_1 and MLP_2.
5) [1.5v] Plot the residues (in absolute value) using two visualizations: boxplots and histograms. *Hint*: consider using `boxplot` and `hist` functions from `matplotlib.pyplot` to this end
6) [1v] How many iterations were required for MLP_1 and MLP_2 to converge?
7) [1.5v] What can be motivating the unexpected differences on the number of iterations? Hypothesize one reason underlying the observed performance differences between the MLPs.

END

Design Matrix :

$$X = \begin{pmatrix} 1 & x_1^1 & x_1^2 & x_1^3 \\ 1 & x_2^1 & x_2^2 & x_2^3 \\ 1 & x_3^1 & x_3^2 & x_3^3 \\ 1 & x_4^1 & x_4^2 & x_4^3 \\ 1 & x_5^1 & x_5^2 & x_5^3 \end{pmatrix} = \begin{pmatrix} 1 & 0.8 & 0.64 & 0.512 \\ 1 & 1 & 1 & 1 \\ 1 & 1.2 & 1.44 & 1.728 \\ 1 & 1.4 & 1.96 & 2.744 \\ 1 & 1.6 & 2.56 & 4.096 \end{pmatrix}$$

Bias

$$X^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0.8 & 1 & 1.2 & 1.4 & 1.6 \\ 0.64 & 1 & 1.44 & 1.96 & 2.56 \\ 0.512 & 1 & 1.728 & 2.744 & 4.096 \end{pmatrix}$$

$4 \times 5 \quad 5 \times 4$

$$X^T X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0.8 & 1 & 1.2 & 1.4 & 1.6 \\ 0.64 & 1 & 1.44 & 1.96 & 2.56 \\ 0.512 & 1 & 1.728 & 2.744 & 4.096 \end{pmatrix} \begin{pmatrix} 1 & 0.8 & 0.64 & 0.512 \\ 1 & 1 & 1 & 1 \\ 1 & 1.2 & 1.44 & 1.728 \\ 1 & 1.4 & 1.96 & 2.744 \\ 1 & 1.6 & 2.56 & 4.096 \end{pmatrix}$$

"B^T" "B"

$$= \begin{pmatrix} 5 & 6 & 7.6 & 10.08 \\ 6 & 7.6 & 10.08 & 13.8784 \\ 7.6 & 10.08 & 13.8784 & 19.68 \\ 10.08 & 13.8784 & 19.68 & 28.55488 \end{pmatrix}$$

$$\left(X^TX+\lambda I\right)^{-1}=\left[\begin{pmatrix}5 & 6 & 7.6 & 10.08\\ 6 & 7.6 & 10.08 & 13.8784\\ 7.6 & 10.08 & 13.8784 & 19.68\\ 10.08 & 13.8784 & 19.68 & 28.55488\end{pmatrix}+\begin{pmatrix}2 & 0 & 0 & 0\\ 0 & 2 & 0 & 0\\ 0 & 0 & 2 & 0\\ 0 & 0 & 0 & 2\end{pmatrix}\right]^{-1}$$

$$=\begin{pmatrix}7 & 6 & 7.6 & 10.08\\ 6 & 9.6 & 10.08 & 13.8784\\ 7.6 & 10.08 & 15.8784 & 19.68\\ 10.08 & 13.8784 & 19.68 & 30.55488\end{pmatrix}^{-1}$$

$A=$

$$=\begin{pmatrix}0.3417 & -0.1214 & -0.0749 & -0.0093\\ -0.1214 & 0.3892 & -0.0967 & -0.0745\\ -0.0749 & -0.0967 & 0.3726 & -0.1714\\ -0.0093 & -0.0745 & -0.1714 & 0.18\end{pmatrix}$$

$A^{-1}=$ $\qquad B=X$

$$\underbrace{\left(X^TX+\lambda I\right)}_{A}{}^{-1}\underbrace{X}_{B}{}^T=\begin{pmatrix}0.3417 & -0.1214 & -0.0749 & -0.0093\\ -0.1214 & 0.3892 & -0.0967 & -0.0745\\ -0.0749 & -0.0967 & 0.3726 & -0.1714\\ -0.0093 & -0.0745 & -0.1714 & 0.18\end{pmatrix}\begin{pmatrix}1 & 1 & 1 & 1 & 1\\ 0.8 & 1 & 1.2 & 1.4 & 1.6\\ 0.64 & 1 & 1.44 & 1.96 & 2.56\\ 0.512 & 1 & 1.728 & 2.744 & 4.096\end{pmatrix}$$

$$=\begin{pmatrix}0.1918 & 0.136 & 0.072 & -0.0007 & -0.082\\ 0.9 & 0.0966 & 0.0777 & 0.0296 & -0.051\\ -0.001 & 0.0296 & 0.0495 & 0.05 & 0.0223\\ -0.086 & -0.075 & -0.034 & 0.0444 & 0.1701\end{pmatrix}$$

$$W=\underbrace{\left(X^TX+\lambda I\right)}_{A}{}^{-1}\underbrace{X}_{B}{}^T\underbrace{Z}_{C}=\begin{pmatrix}0.1918 & 0.136 & 0.072 & -0.0007 & -0.082\\ 0.9 & 0.0966 & 0.0777 & 0.0296 & -0.051\\ -0.001 & 0.0296 & 0.0495 & 0.05 & 0.0223\\ -0.086 & -0.075 & -0.034 & 0.0444 & 0.1701\end{pmatrix}\begin{pmatrix}24\\ 20\\ 10\\ 13\\ 12\end{pmatrix}$$

$$=\begin{pmatrix}7.0451\\ 4.6409\\ 1.9673\\ 1.301\end{pmatrix}$$