Aprendizagem 2022

Lab 4: Linear Regression and *k*NN

Practical exercises

I. Lazy learning

1. Consider the following data:

	input		output	
	y_1	y_2	y_3	y_4
$\mathbf{x}_1$	1	1	A	1.4
$\mathbf{x}_2$	2	1	B	0.5
$\mathbf{x}_3$	2	3	B	2
$\mathbf{x}_4$	3	3	B	2.2
$\mathbf{x}_5$	2	2	A	0.7
$\mathbf{x}_6$	1	2	A	1.2

Assuming a *k*-nearest neighbor with *k*=3 applied within a leave-one-out schema:

a) Let y_3 be the output variable (*categoric*). Considering an Euclidean (*l*2) distance, provide the classification estimates for x_1.

b) Let y_4 be the output variable (*numeric*). Considering cosine similarity, provide the mean regression estimate for x_1.

c) Consider a weighted-distance *k*-nearest neighbor with Manhattan (l1) distance, identify the:

i. weighted mode estimate of x_1 for the y_3 outcome

ii. weighted mean estimate of x_1 for the y_4 outcome

II. Linear regression

2. Consider the following training data:

	y_1	y_2	output
$\mathbf{x}_1$	1	1	1.4
$\mathbf{x}_2$	2	1	0.5
$\mathbf{x}_3$	1	3	2
$\mathbf{x}_4$	3	3	2.5

a) Find the closed form solution for a linear regression . minimizing the sum of squared errors

b) Predict the target value for $x_{new} = [2\ 3]^T$

c) Sketch the predicted three-dimensional hyperplane

d) Compute the MSE and MAE produced by the linear regression

e) Are there biases on the residuals against y1? And y2?

f) Compute the closed form solution considering Ridge regularization term with $\lambda = 0.2$.

g) Compare the hyperplanes obtained using ordinary least squares and Ridge regression.

h) Why is Lasso regression suggested for data spaces of higher dimensionality?

3. Consider the following training data where *output* is an ordinal variable

a) Find a linear regression using the closed form solution

b) Assuming the output threshold θ=0.5, use the regression to classify $x_{\text{new}} = [2\ 2.5]^T$

	y_1	y_2	output
$\mathbf{x}_1$	1	1	1
$\mathbf{x}_2$	2	1	1
$\mathbf{x}_3$	1	3	0
$\mathbf{x}_4$	3	3	0

4. Considering the following data to learn a model $z = w_1 y_1 + w_2 y_2 + \varepsilon$, where $\varepsilon \sim N(0,5)$

Compare:

	y_1	y_2	output
$\mathbf{x}_1$	3	-1	2
$\mathbf{x}_2$	4	2	1
$\mathbf{x}_3$	2	2	1

a) $\boldsymbol{w} = [w_1\ w_2]^T$ using the maximum likelihood approach

b) $\boldsymbol{w}$ using the Bayesian approach, assuming $p(\boldsymbol{w}) = N\left(\boldsymbol{w} \mid \boldsymbol{u} = [0\ 0], \boldsymbol{\sigma} = \begin{bmatrix} 0.2 & 0 \\ 0 & 0.2 \end{bmatrix}\right)$

5. Identify a transformation to aid the linearly modelling of the following data points.

Sketch the predicted surface.

	y_1	y_2	output
x_1	-0.95	0.62	0
x_2	0.63	0.31	0
x_3	-0.12	-0.21	1
x_4	-0.24	-0.5	0
x_5	0.07	-0.42	1
x_6	0.03	0.91	0
x_7	0.05	0.09	1
x_8	-0.83	0.22	0

6. Consider logarithmic and quadratic transformations: $\varphi_1(x_1) = log(x_1)$, $\varphi_2(x_1) = {x_1}^2$

a) Plot both of the closed form regressions.

b) Which one minimizes the sum of squared errors on the original training data

	input	*output*
$\mathbf{x}_1$	3	1.5
$\mathbf{x}_2$	4	9.3
$\mathbf{x}_3$	6	23.4
$\mathbf{x}_4$	10	45.8
$\mathbf{x}_5$	12	60.1

7. Select the criteria that promotes a smoother regression model:

a) Applying Lasso and Ridge regularization to linear regression models

b) Increasing the depth of a decision tree regressor

c) Increasing the *k* of a *k*NN regressor

d) Parameterizing a *k*NN regressor with uniform weights instead of distance-based weights

Programming quest

8. Consider the *housing* dataset available at https://web.ist.utl.pt/~rmch/dscience/data/housing.arff and the *Regression* notebook available at the course's webpage:

a) Compare the determination coefficient of the non-regularized, Lasso and Ridge linear regression

b) Compare the MAE and RMSE of linear, kNN and decision tree regressors on housing