Aprendizagem 2023

Lab 4: *k*NN and Evaluation

Practical exercises

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
$\mathbf{x}_5$	1	0	A	0.7
$\mathbf{x}_6$	1	4	A	1.2

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

a) Let y_3 be the output variable (*categoric*). Classify $\mathbf{x}_1$ when considering uniform weights and:

i. Euclidean (*l*2) distance (real input variables)

$\|\mathbf{x}_i - \mathbf{x}_j\|_2$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$
$\mathbf{x}_1$	-	1	$\sqrt{5}$	$\sqrt{8}$	1	3

$$\hat{z}_1 = mode(B, B, A) = B$$

ii. Hamming distance (categorical input variables)

$\mathrm{H}(\mathbf{x}_i, \mathbf{x}_j)$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$
$\mathbf{x}_1$	-	1	2	2	1	1

$$\hat{z}_1 = mode(B, A, A) = A$$

b) Let y_4 be the output variable (*numeric*). Considering cosine similarity, provide the mean regression estimate for $\mathbf{x}_1$

cos	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$
$\mathbf{x}_1$	-	0.95	0.98	1	0.70	0.86

$$\hat{z}_2 = mean(0.5, 2, 2.2) = 1.5(6)$$

c) Consider a weighted-distance *k*NN with Euclidean (*l*2) distance, identify:

i. the weighted mode estimate of $\mathbf{x}_1$ for the y_3 outcome

l_1	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$
$\mathbf{x}_1$	-	1	$\sqrt{5}$	$\sqrt{8}$	1	3

$$\hat{z}_1 = weighted\_mode\left(1 \times A, \left(\frac{1}{1} + \frac{1}{\sqrt{5}}\right) B\right) = weighted\_mode(A, 1.45 \times B) = B$$

ii. the weighted mean estimate of $\mathbf{x}_1$ for the y_4 outcome

$$\hat{z}_1 = \frac{\frac{1}{1}0.5 + \frac{1}{\sqrt{5}}2 + \frac{1}{1}0.7}{\frac{1}{1} + \frac{1}{\sqrt{5}} + \frac{1}{1}} = 0.86$$

2. Let x_j be the measurement on variable y_j for observation $\mathbf{x}$.

Given the learnt regression model $\hat{x}_4 = 1 - 0.8x_1 + 0.2{x_2}^2 + 0.2x_1x_2$:

a) Compute the y_4 regression estimates for the observations of the aforementioned dataset

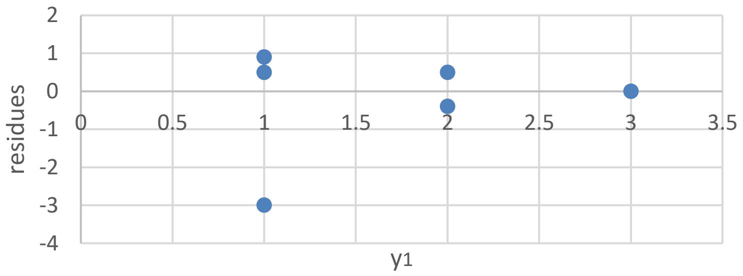
$$\hat{\mathbf{z}} = (0.6 \quad 0 \quad 2.4 \quad 2.2 \quad 0.2 \quad 4.2)$$

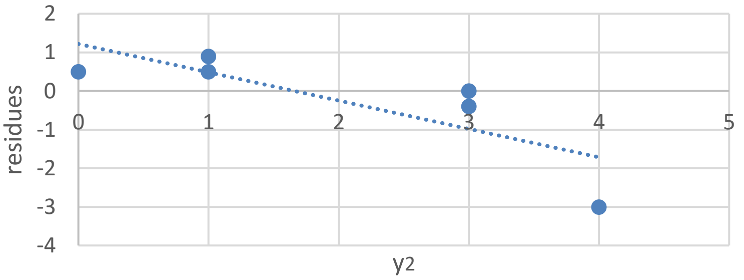
b) Compute the training Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE)

$$\mathbf{z} - \hat{\mathbf{z}} = (0.8 \quad 0.5 \quad \text{-}0.4 \quad 0 \quad 0.5 \quad \text{-}3)$$

$$MAE = 0.8(6), \quad RMSE = 1.31$$

c) Perform a residue analysis to assess the presence of systemic biases against y_1 and y_2





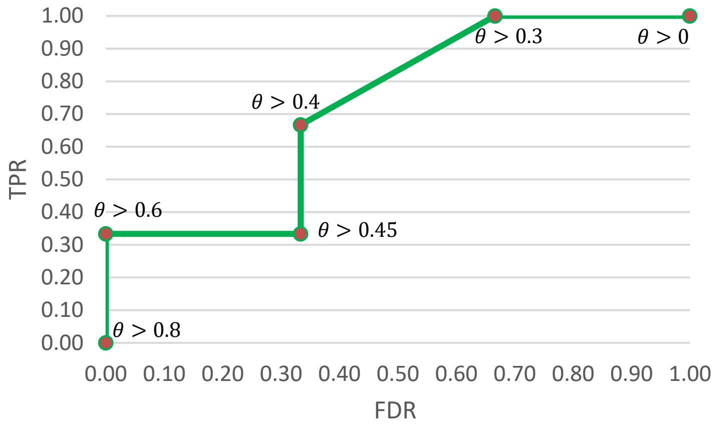
There is no evidence towards the presence of biases on y1. However, as residues appear to be correlated against y2, we can hypothesize that the learnt regressor is moderately biased against y2 for the given data.

3. [*optional*] Consider the probabilistic outcome of a classifier for the given six observations to be

$$\mathbf{p}(y_3 = A \mid \mathbf{x}) = [p(y_3 = A \mid \mathbf{x_1}), \ldots, p(y_3 = A \mid \mathbf{x_6})] = [0.45 \quad 0.4 \quad 0.3 \quad 0.6 \quad 0.8 \quad 0.4]$$

a) Draw the training ROC curve

z	$\hat{z}$	0	>0.3	>0.4	>0.45	>0.6	>0.8
1	0.45	TP	TP	TP	FN	FN	FN
0	0.4	FP	FP	TN	TN	TN	TN
0	0.3	FP	TN	TN	TN	TN	TN
0	0.6	FP	FP	FP	FP	TN	TN
1	0.8	TP	TP	TP	TP	TP	FN
1	0.4	TP	TP	FN	FN	FN	FN
FPR=FP/N		1.00	0.67	0.33	0.33	0.00	0.00
TPR=TP/P		1.00	1.00	0.67	0.33	0.33	0.00
F1		2/3	0.75	2/3	0.5	0.5	NA



b) Compute the training AUC

$$AUC = \left(\frac{1}{3}\times\frac{1}{3}\right) + \left(\frac{2}{3}\times\frac{1}{3} + \frac{1}{2}\times\frac{1}{3}\times\frac{1}{3}\right) + \left(1\times\frac{1}{3}\right) = 0.72$$

c) Would you change the default 0.5 probability threshold for this classifier in order to maximize training F1?

Yes, training F1 is maximal when the probability threshold $\theta \in]0.3, 0.4]$

Programming quest

1. Consider the accuracy estimates collected under a 5-fold CV for two predictive models M1 and M2, acc_{M1}=(0.7,0.5,0.55,0.55,0.6) and acc_{M2}=(0.75,0.6,0.6,0.65,0.55).
Using **scipy**, assess whether the differences in predictive accuracy are statistically significant.
Resource: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_rel.html

4. 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. Using a 10-fold cross-validation:

a) Assess the MAE of a kNN regressor for $k \in \{1,5,9\}$ (remaining parameters as default)

b) Compare the RMSE of the default kNN and decision tree regressors