TÉCNICO LISBOA

Aprendizagem 2023/24

Homework II

Non-exhaustive solution notes

I. Pen-and-paper [13v]

Consider the following dataset:

D	y_1	y_2	y_3	y_4	y_5	y_6
$\mathbf{x}_1$	0.24	0.36	1	1	0	A
$\mathbf{x}_2$	0.16	0.48	1	0	1	A
$\mathbf{x}_3$	0.32	0.72	0	1	2	A
$\mathbf{x}_4$	0.54	0.11	0	0	1	B
$\mathbf{x}_5$	0.66	0.39	0	0	0	B
$\mathbf{x}_6$	0.76	0.28	1	0	2	B
$\mathbf{x}_7$	0.41	0.53	0	1	1	B
$\mathbf{x}_8$	0.38	0.52	0	1	0	A
$\mathbf{x}_9$	0.42	0.59	0	1	1	B

1. Consider $\mathbf{x}_1$–$\mathbf{x}_7$ to be training observations, $\mathbf{x}_8$–$\mathbf{x}_9$ to be testing observations, $y_1 - y_5$ to be input variables and y_6 to be the target variable.

   a. [3.5v] Learn a Bayesian classifier assuming: i) $\{y_1, y_2\}$, $\{y_3, y_4\}$ and $\{y_5\}$ sets of independent variables (e.g., $y_1 \perp\!\!\!\perp y_3$ yet $y_1 \not\perp\!\!\!\perp y_2$), and ii) $y_1 \times y_2 \in \mathbb{R}^2$ is normally distributed. Show all parameters (distributions and priors for subsequent testing).

$$p(z|\mathbf{x}) = \frac{p(\mathbf{x}|z) \times p(z)}{p(\mathbf{x})}$$

$$p(z)\text{: } \textbf{priors}\text{: } p(A) = \frac{3}{7},\ \ p(B) = \frac{4}{7}$$

$$\textbf{PMFs}\text{: } p(y_3, y_4|z)\text{: } p(0,0|A) = 0,\ \ p(0,1|A) = \frac{1}{3},\ \ p(1,0|A) = \frac{1}{3},\ \ p(1,1|A) = \frac{1}{3}$$

$$p(0,0|B) = \frac{2}{4},\ \ p(0,1|B) = \frac{1}{4},\ \ p(1,0|B) = \frac{1}{4},\ \ p(1,1|B) = 0$$

$$p(y_5|z)\text{: } p(0|A) = \frac{1}{3},\ \ p(1|A) = \frac{1}{3},\ \ p(2|A) = \frac{1}{3}, p(0|B) = \frac{1}{4},\ \ p(1|B) = \frac{2}{4},\ \ p(2|B) = \frac{1}{4}$$

$$\textbf{PDFs}\text{: } N\left(\mathbf{u} = \begin{bmatrix}0.24\\0.52\end{bmatrix}, \Sigma = \begin{bmatrix}0.0064 & 0.0096\\0.0096 & 0.0336\end{bmatrix} \mid A\right),\ \ N\left(\mathbf{u} = \begin{bmatrix}0.5925\\0.3275\end{bmatrix}, \Sigma = \begin{bmatrix}0.0229 & -0.00976\\-0.00976 & 0.0315\end{bmatrix} \middle| B\right)$$

$p(\mathbf{x})$ is <u>optional</u> for classification purposes

   b. [2.5v] Under a MAP assumption, classify each testing observation showing all your calculus.

Let us compute the $p(z|\mathbf{x})$ for each observation:

$$\mathbf{x}_8\text{: } p(0.38, 0.52, 0, 1, 0|A) = p(0.38, 0.52|A)p(0,1|A)p(0|A) = 0.9847 \times \frac{1}{3} \times \frac{1}{3} = 0.1094,$$

$$p(0.38, 0.52, 0, 1, 0|B) = 1.96237 \times \frac{1}{4} \times \frac{1}{4} = 0.12265$$

$$p(A|\mathbf{x}_8) = 0.1094 \times \frac{3}{7} \times k,\ \ p(B|\mathbf{x}_8) = 0.12265 \times \frac{4}{7} \times k$$

$$\text{normalization: } p(A|\mathbf{x}_8) = \frac{p(\mathbf{x}|A)p(A)}{p(\mathbf{x}|A)p(A) + p(\mathbf{x}|A)p(A)} = \frac{p(A|\mathbf{x})}{p(A|\mathbf{x}) + p(B|\mathbf{x})} = 0.40,\ \ p(B|\mathbf{x}_8) = 0.60$$

$\mathbf{x}_8$ is classified as B

Aprendizagem 2023/24

Homework II

Non-exhaustive solution notes

$$\mathbf{x}_9\text{: } p(0.42, 0.59, 0, 1, 1|A) = 0.0448, \ \ p(0.42, 0.59, 0, 1, 1|B) = 0.216$$

$$p(A|\mathbf{x}_9) = 0.0448 \times \frac{3}{7} \times k, \ \ p(B|\mathbf{x}_9) = 0.216 \times \frac{4}{7} \times k$$

$$\text{normalization: } p(A|\mathbf{x}_9) = 0.135, \ \ p(B|\mathbf{x}_9) = 0.865$$

$\mathbf{x}_9$ is classified as B

c. [2v] Consider that the default decision threshold of $\theta = 0.5$ can be adjusted according to

$$f(\mathbf{x}|\theta) = \begin{cases} \text{A} & P(\text{A}|\mathbf{x}) > \theta \\ \text{B} & \text{otherwise} \end{cases} .$$

Under a maximum likelihood assumption, what thresholds optimize testing accuracy?

Maximum likelihood estimates assume uniform prior information for posterior calculus:

$$p(A|\mathbf{x}_8) = 0.47, \ \ p(B|\mathbf{x}_8) = 0.53$$

$$p(A|\mathbf{x}_9) = 0.17, \ \ p(B|\mathbf{x}_9) = 0.83$$

To optimize testing accuracy $0.17 < \theta < 0.47$, so that $\mathbf{x}_8$ is classified as A and $\mathbf{x}_9$ as B

2. Let y_1 be the target numeric variable, y_2-y_6 be the input variables where y_2 is binarized under an equal-width (equal-range) discretization. For the evaluation of regressors, consider a 3-fold cross-validation over the full dataset ($\mathbf{x}_1$- $\mathbf{x}_9$) without shuffling the observations.

a. [1v] Identify the observations and features per data fold after the binarization procedure.

Accepted binarization under either $y_2 \in [0,1]$ or $y_2 \in [0.11, 0.72]$ assumption

fold		y_{out}	y'_2	y_3	y_4	y_5	y_6
	$\mathbf{x}_1$	0.24	0	1	1	0	A
1	$\mathbf{x}_2$	0.16	0	1	0	1	A
	$\mathbf{x}_3$	0.32	1	0	1	2	A
	$\mathbf{x}_4$	0.54	0	0	0	1	B
2	$\mathbf{x}_5$	0.66	0	0	0	0	B
	$\mathbf{x}_6$	0.76	0	1	0	2	B
	$\mathbf{x}_7$	0.41	1	0	1	1	B
3	$\mathbf{x}_8$	0.38	1	0	1	0	A
	$\mathbf{x}_9$	0.42	1	0	1	1	B

b. [4v] Consider a distance-weighted kNN with $k = 3$, the Hamming distance (d), and $1/d$ weights. Compute the MAE of this kNN regressor for the 1st iteration of the cross-validation (where the train observations have the lower indices).

Pairwise distances table with 3 nearest neighbors highlighted

Hamming	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$
$\mathbf{x}_7$	4	4	**2**	**2**	**3**	4
$\mathbf{x}_8$	**2**	4	**1**	4	**3**	5
$\mathbf{x}_9$	4	4	**2**	**2**	**3**	4

Weighted means:

$$\hat{z}_7 = 0.375 \times \hat{z}_3 + 0.375 \times \hat{z}_4 + 0.25 \times \hat{z}_5 = 0.4512$$

$$\hat{z}_8 = 0.3896, \ \ \hat{z}_9 = 0.4512$$

$$MAE = \frac{|\hat{z}_7 - z_7| + |\hat{z}_8 - z_8| + |\hat{z}_9 - z_9|}{3} = 0.027$$

TÉCNICO LISBOA

Aprendizagem 2023/24

Homework II

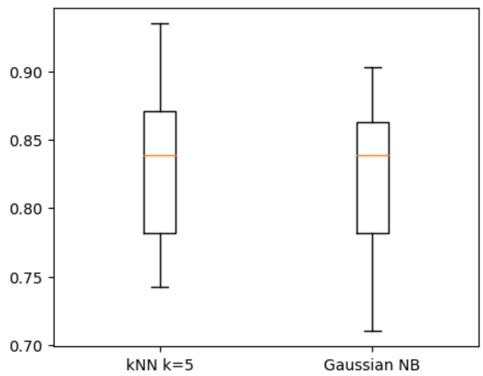
Non-exhaustive solution notes

II. Programming and critical analysis [7v]

Considering the `column_diagnosis.arff` dataset available at the course webpage's homework tab. Using `sklearn`, apply a 10-fold stratified cross-validation with shuffling (`random_state=0`) for the assessment of predictive models along this section.

1) [3v] Compare the performance of kNN with $k = 5$ and Naïve Bayes with Gaussian assumption (consider all remaining parameters for each classifier as `sklearn`'s default):

   a. Plot two boxplots with the fold accuracies for each classifier.




```
kNN k=5 accuracies: 0.839 ± 0.063
[0.935, 0.806, 0.871, 0.935, 0.742, 0.871,
0.839, 0.839, 0.774, 0.774]
Gaussian NB accuracies: 0.823 ± 0.054
[0.839, 0.871, 0.839, 0.871, 0.774, 0.839,
0.903, 0.806, 0.774, 0.71]
```

```
folds = StratifiedKFold(n_splits=10,shuffle=True,random_state=0)
predictors = [KNeighborsClassifier(n_neighbors=1),KNeighborsClassifier(n_neighbors=5),GaussianNB()]
accs, cms = [[],[],[]], [np.zeros((3,3))]*3

# A: iterate per fold
for train_k, test_k in folds.split(X, y):
    X_train, X_test = X.iloc[train_k], X.iloc[test_k]
    y_train, y_test = y.iloc[train_k], y.iloc[test_k]

    # B: train and assess
    for i in range(len(predictors)):
        predictors[i].fit(X_train, y_train)
        pred = predictors[i].predict(X_test)
        accs[i].append(round(metrics.accuracy_score(y_test, pred),3))
        cms[i] = cms[i] + np.array(metrics.confusion_matrix(y_test, pred))

# C: print results
sort_cols = ['Hernia','Normal','Spondylolisthesis']
for i in range(len(predictors)):
    print(predictors[i],"\nAccuracies:",round(np.mean(accs[i]),3),"±",round(np.std(accs[i]),3),"\n",accs[i])
    cm_df = pd.DataFrame(cms[i],index=np.char.add('True ',sort_cols),columns=np.char.add('Pred ',sort_cols))
    print("Confusion matrix:\n",cm_df)

plt.boxplot(accs)
plt.xticks([1,2,3],["kNN k=1","kNN k=5","NB"])
```

Aprendizagem 2023/24

Homework II

Non-exhaustive solution notes

b. Using `scipy`, test the hypothesis "kNN is statistically superior to Naïve Bayes regarding accuracy", asserting whether is true.

```
res = stats.ttest_rel(knn_accs, nb_accs, alternative='greater')
```

For the specifications in the statement, p-value=0. 19. One cannot reject the null hypothesis at common significance levels (e.g., $\alpha = 0.05$), and thus we cannot assert the given hypothesis as true. Note that we should refrain from stating that the given hypothesis is false or that equality holds in the absence of additional statistical tests.

2) [2.5v] Consider two kNN predictors with $k = 1$ and $k = 5$ (uniform weights, Euclidean distance, all remaining parameters as default). Plot the differences between the two cumulative confusion matrices of the predictors. Comment.

kNN k=1 confusion matrix:

	Pred Hernia	Pred Normal	Pred Spondylolisthesis
True Hernia	37.0	23.0	0.0
True Normal	14.0	80.0	6.0
True Spondylolisthesis	1.0	7.0	142.0

kNN k=5 confusion matrix:

	Pred Hernia	Pred Normal	Pred Spondylolisthesis
True Hernia	39.0	21.0	0.0
True Normal	19.0	78.0	3.0
True Spondylolisthesis	1.0	6.0	143.0

Differences:

	Pred Hernia	Pred Normal	Pred Spondylolisthesis
True Hernia	**+2**	**-2**	0
True Normal	**+5**	**-2**	**-3**
True Spondylolisthesis	0	**-1**	**+1**

Small-to-moderate differences are observed, including a moderate increase in Hernia's recall and Spondylolisthesis' precision, and a moderate deterioration of Normal's recall.

3) [1.5v] Considering the unique properties of the given dataset, identify three major difficulties of naïve Bayes learning.

1) variable dependencies (inadequacy of independence assumption)

2) variables not normally distributed (inadequacy of Gaussian assumption)

3) locality of decisions (preference towards local classifiers such as kNN with small k)

4) probability estimates from a moderate number of observations (e.g., inadequate estimates, null probabilities)

5) moderate imbalance between classes creating biases in MAP estimates via priors

…

END