Aprendizagem 2021/22

Homework I

Deadline 15/10/2021 23:59 via Fenix as PDF

SOLUTION NOTES

I. Pen-and-paper [11v]

Considering the following training data:

	y_1	y_2	y_3	y_4	*class*
$\mathbf{x}_1$	0.6	A	0.2	0.4	0 (N)
$\mathbf{x}_2$	0.1	B	-0.1	-0.4	0
$\mathbf{x}_3$	0.2	A	-0.1	0.2	0
$\mathbf{x}_4$	0.1	C	0.8	0.8	0
$\mathbf{x}_5$	0.3	B	0.1	0.3	1 (P)
$\mathbf{x}_6$	-0.1	C	0.2	-0.2	1
$\mathbf{x}_7$	-0.3	C	-0.1	0.2	1
$\mathbf{x}_8$	0.2	B	0.5	0.6	1
$\mathbf{x}_9$	0.4	A	-0.4	-0.7	1
$\mathbf{x}_{10}$	-0.2	C	0.4	0.3	1

1) [4v] Train a Bayesian classifier assuming: i) independence and equal importance between {y1}, {y2} and {y3,y4} variable sets, and ii) numeric variables are normally distributed.

We need to estimate all necessary parameters to place decisions:

$$p(C=0 \mid y1,y2,y3,y4) = \frac{p(C=0)p(y1,y2,y3,y4|C=0)}{p(y1,y2,y3,y4)} = \frac{p\,(C=0)p(y1|C=0)p(y2|C=0)\,p(y3,y4|C=0)}{p(y1)p(y2)p(y3,y4)}$$

$$p(C=1 \mid y1,y2,y3,y4) = \frac{p(C=1)p(y1,y2,y3,y4|C=1)}{p(y1,y2,y3,y4)} = \frac{p\,(C=1)p(y1|C=1)p(y2|C=1)\,p(y3,y4|C=1)}{p(y1)p(y2)p(y3,y4)}$$

To place decisions, these probabilities need to be compared. To this end, numerators need to be estimated:

$$p(C=0)=0.4, \qquad p(C=1)=0.6$$

$$y1|C=0 \sim N(\mu_0=0.25, \sigma_0=0.238), \qquad y1|C=1 \sim N(\mu_1=0.05, \sigma_1=0.288)$$

$$p(y1=x|C=0) = \frac{1}{\sigma_0\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_0}{\sigma_0}\right)^2}, \quad p(y1=x|C=0) = \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2}$$

$$p(y2=x|C=0) = \begin{cases} 0.5 & x=A \\ 0.25 & x=B \\ 0.25 & x=C \end{cases}, \qquad p(y2=x|C=1) = \begin{cases} 1/6 & x=A \\ 1/3 & x=B \\ 0.5 & x=C \end{cases}$$

$$y3,y4\,\Big|\,C=0 \sim N\left(\boldsymbol{\mu}_0 = \begin{bmatrix} 0.2 \\ 0.25 \end{bmatrix}, \Sigma_0 = \begin{bmatrix} 0.18 & 0.18 \\ 0.18 & 0.25 \end{bmatrix}\right), \qquad y3,y4\,\Big|\,C=1 \sim N\left(\boldsymbol{\mu}_1 = \begin{bmatrix} 0.1167 \\ 0.083 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 0.11 & 0.12 \\ 0.12 & 0.21 \end{bmatrix}\right)$$

$$det(\Sigma_0) = 0.01262, {\Sigma_0}^{-1} = \begin{bmatrix} 19.84 & -14.286 \\ -14.286 & 14.286 \end{bmatrix}, \qquad det(\Sigma_1) = 0.00847, {\Sigma_1}^{-1} = \begin{bmatrix} 25.236 & -14.449 \\ -14.449 & 12.95 \end{bmatrix}$$

$$p(y3,y4=\mathbf{x}|C=0) = (2\pi)^{-\frac{2}{2}}\, det(\Sigma_0)^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_0)^T {\Sigma_0}^{-1}(\mathbf{x}-\boldsymbol{\mu}_0)}, \quad p(y3,y4=\mathbf{x}|C=1) = (2\pi)^{-\frac{2}{2}}\, det(\Sigma_1)^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T {\Sigma_1}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)}$$

Grading criteria:
- prior and model: 15%
- y1 parameters: 25%
- y2 parameters: 25%
- {y3,y4} parameters: 35%

Common discounts:
- model not shown (not even in succeeding items): -10%
- *population* instead of sample std deviation: -10% (per numeric set)
- final parameters only shown for one class

TÉCNICO LISBOA

Aprendizagem 2021/22

Homework I

Deadline 15/10/2021 23:59 via Fenix as PDF

2) [4v] Draw a confusion matrix for the training observations.

Note: you can use `scipy` or `excel` to support/check your estimates, yet show intermediary results.

Observation $\mathbf{x}_1$:

$$p(C=0 \mid \mathbf{x}_1) = p(C=0 \mid y1=0.6, y2=\mathrm{A}, y3=0.2, y4=0.4) = \frac{p\,(C=0)p(y1=0.6|C=0)p(y2=\mathrm{A}|C=0)\;p(y3=0.2, y4=0.4|C=0)}{p(y1)p(y2)p(y3,y4)}$$

$$p(C=1 \mid \mathbf{x}_1) = p(C=1|y1=0.6, y2=A, y3=0.2, y4=0.4) = \frac{p(C=1)p(y1=0.6|C=1)p(y2=A|C=1)p(y3=0.2, y4=0.4|C=1)}{p(y1)p(y2)p(y3,y4)}$$

$$p\,(C=0)p(y1=0.6|C=0)p(y2=A|C=0)\;p(y3=0.2, y4=0.4 \mid C=0)$$
$$= 0.6 \times p\big(x=0.6 \mid N(\mu_0=0.25, \sigma_0=0.238)\big) \times 0.5 \times p\left(\mathbf{x}=[0.2{,}0.4] | \boldsymbol{\mu}_0 = \begin{bmatrix}0.2\\0.25\end{bmatrix}, \Sigma_0 = \begin{bmatrix}0.18 & 0.18\\0.18 & 0.25\end{bmatrix}\right) = 8.84\mathrm{E}-09$$

$$p\,(C=1)p(y1=0.6|C=1)p(y2=A|C=1)\;p(y3=0.2, y4=0.4 \mid C=1)$$
$$= 0.6 \times p\big(x=0.6 \mid N(\mu_1=0.05, \sigma_1=0.288)\big) \times 0.5 \times p\left(\mathbf{x}=[0.2{,}0.4] | \boldsymbol{\mu}_1 = \begin{bmatrix}0.1167\\0.083\end{bmatrix}, \Sigma_1 = \begin{bmatrix}0.11 & 0.12\\0.12 & 0.21\end{bmatrix}\right) = 1.70\mathrm{E}-10$$

Class estimate for $\mathbf{x}_1$ is $C=1$ as $p(C=1|x1) > p(C=0|x1)$

Repeating for the remaining observations:

	p(C=0)	p(x|y1,C=0)	p(x|y2,C=0)	p(x|y3,y4,C=0)	numerator	p(C=1)	p(x|y1,C=1)	p(x|y2,C=1)	p(x|y3,y4,C=1)	numerator	class
$\mathbf{x}_1$	0.4	0.56862374	0.5	1.2074	0.13731	0.6	0.22385196	0.167	1.2119	0.02713	**0**
$\mathbf{x}_2$	0.4	1.37412425	0.25	0.4603	0.06325	0.6	1.36405047	0.33	0.9567	0.261	**1**
$\mathbf{x}_3$	0.4	1.63932932	0.5	0.7066	0.23167	0.6	1.20922134	0.17	0.6079	0.0735	**0**
$\mathbf{x}_4$	0.4	1.37412425	0.25	0.5124	0.07041	0.6	1.36405047	0.5	0.2030	0.08308	**1**
$\mathbf{x}_5$	0.4	1.63932932	0.25	1.1743	0.1925	0.6	0.95029081	0.33	1.2071	0.22942	**1**
$\mathbf{x}_6$	0.4	0.56862374	0.25	0.3338	0.01898	0.6	1.20922134	0.5	0.6698	0.24299	**1**
$\mathbf{x}_7$	0.4	0.11616176	0.25	0.7066	0.00821	0.6	0.66203755	0.5	0.6079	0.12073	**1**
$\mathbf{x}_8$	0.4	1.63932932	0.25	1.0847	0.17782	0.6	1.20922134	0.33	0.8408	0.20334	**1**
$\mathbf{x}_9$	0.4	1.37412425	0.50	0.2174	0.05976	0.6	0.66203755	0.167	0.3880	0.02569	**0**
$\mathbf{x}_{10}$	0.4	0.28071407	0.25	1.0804	0.03033	0.6	0.95029081	0.5	1.1252	0.32078	**1**

Comparing predictions against truth ground:

		predicted	
		P (1)	N (0)
true	P (1)	TP=5	FN=1
	N (0)	FP=2	TN=2

3) [0.5v]Evaluate the training F1 score.

$$F1 = \frac{2}{recall^{-1} + precision^{-1}} = \frac{TP}{TP + \frac{1}{2}(FP+FN)} = \mathbf{0.769}$$

Grading criteria:

- probability of the Gaussians: 20%
- overall posterior: 20%
- soundness of all observations: 20%
- confusion matrix TP/TN: 20%
- confusion matrix FP/FN: 20%

Common discount: lack of intermediate calculations for an illustrative instance: -1%

4) [2.5v] Identify the decision probability threshold that optimizes training accuracy. Comment.

To estimate $p(C = k \mid \mathbf{x})$, we can compute the probabilities of the denominator. However, we can also notice:

$$p(C=0|\mathbf{x}) = 1 - p(C=1|\mathbf{x}) \Leftrightarrow \frac{p(C=0)p(\mathbf{x}|C=0)}{p(\mathbf{x})} = 1 - \frac{p(C=1)p(\mathbf{x}|C=1)}{p(\mathbf{x})} \Leftrightarrow$$

$$p(\mathbf{x}) = p\ (C=0)p(\mathbf{x}|C=0) + p\ (C=1)p(\mathbf{x}|C=1)$$

In fact, when we divide our numerator by $p(\mathbf{x})$, we are simply normalizing. So let us normalize the numerators:

	$p(C=0\|x)$	$p(C=0\|x)$	*Class*
$\mathbf{x}_1$	0.83502	0.16498	0
$\mathbf{x}_2$	0.19507	0.80493	0
$\mathbf{x}_3$	0.75914	0.24086	0
$\mathbf{x}_4$	0.45872	0.54128	0
$\mathbf{x}_5$	0.45625	0.54375	1
$\mathbf{x}_6$	0.07245	0.92755	1
$\mathbf{x}_7$	0.06366	0.93634	1
$\mathbf{x}_8$	0.46651	0.53349	1
$\mathbf{x}_9$	0.69934	0.30066	1
$\mathbf{x}_{10}$	0.08638	0.91362	1

Now, we can compute the ROC curve to identify the best threshold:

	class	0.165	0.241	0.301	0.533	0.541	0.544	0.805	0.914	0.928	0.936	1
$\mathbf{x}_1$	0	FP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN
$\mathbf{x}_2$	0	FP	FP	FP	FP	FP	FP	FP	TN	TN	TN	TN
$\mathbf{x}_3$	0	FP	FP	TN	TN	TN	TN	TN	TN	TN	TN	TN
$\mathbf{x}_4$	0	FP	FP	FP	FP	FP	TN	TN	TN	TN	TN	TN
$\mathbf{x}_5$	1	TP	TP	TP	TP	TP	TP	FN	FN	FN	FN	FN
$\mathbf{x}_6$	1	TP	TP	TP	TP	TP	TP	TP	TP	TP	FN	FN
$\mathbf{x}_7$	1	TP	TP	TP	TP	TP	TP	TP	TP	TP	TP	FN
$\mathbf{x}_8$	1	TP	TP	TP	TP	FN	FN	FN	FN	FN	FN	FN
$\mathbf{x}_9$	1	TP	TP	TP	FN	FN	FN	FN	FN	FN	FN	FN
$\mathbf{x}_{10}$	1	TP	TP	TP	TP	TP	TP	TP	TP	FN	FN	FN
accuracy		0.6	0.7	**0.8**	0.7	0.6	0.7	0.6	0.7	0.6	0.5	0.4

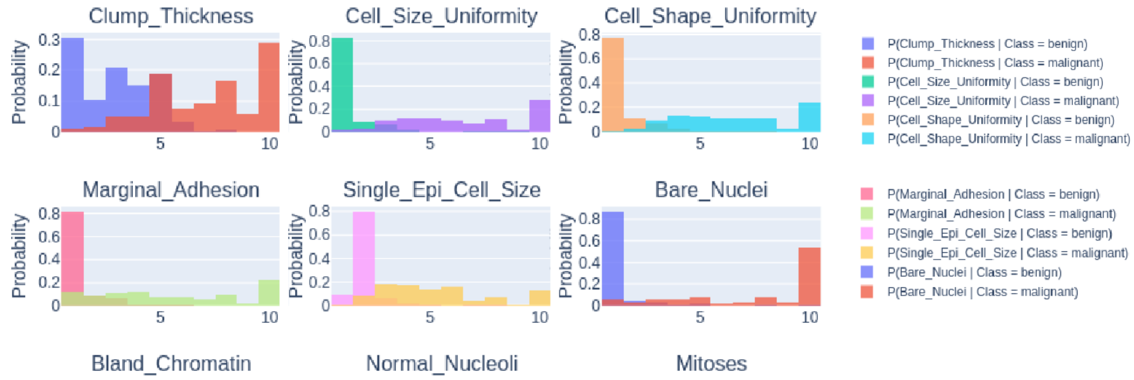
The optimal probability threshold is 0.301.

Common discounts: F1 score computed for the negative class: -20%

II. Programming and critical analysis [9v]

Considering the `breast.w.arff` dataset available at

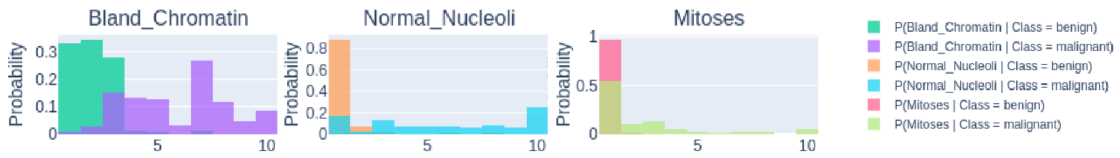
5) [2v] Draw the class-conditional distributions per variable. Suggestion: use 3x3 plot grid.



Aprendizagem 2021/22

Homework I

Deadline 15/10/2021 23:59 via Fenix as PDF



Common discounts: separate presentation of class-conditional distributions per variable -10%

6) [3v] Using a 10-fold cross validation with seed=`<group number>`, assess the accuracy of kNN under $k \in \{3,5,7\}$, Euclidean distance and uniform weights to identify which k is, empirically, less susceptible to the overfitting risk.

Solution notes:
- comparison of training and testing accuracies (or their differences) along $k \in \{3,5,7\}$ using performance estimates obtained from the assessed 5 or 10 fold estimates
- assess the variability of estimates to further check whether differences are significant
- critical analysis of the gathered performance estimates
- final decisions may vary depending on your seeds, generally results favor k=5 although with loose statistical significance

Common discounts:
- incomplete or incorrect code (programming)
- not using training estimates to assess overfitting -25%
- absence of training estimates yet careful assessment of variability estimates -15%

7) [2v] Fixing $k = 3$, and assuming accuracy estimates are normally distributed, test the hypothesis "kNN is statistically superior to Naïve Bayes (multinomial assumption)".

Solution notes: paired (single-tailed) t-test based on the 10-fold testing estimates to assess $k\text{NN} > NB$.

Null hypothesis (equal means) is rejected at 1%, confirming statistical superiority of kNN.

Common discounts:
- absence of statistical testing
- lacking answer after statistical testing
- incorrect interpretation of p-value against null hypothesis
- absence of statistical testing yet careful assessment of variability estimates -40%

8) [2v] Given the empirical data collected along 5-7, enumerate two reasons that can underlie the differences in performance between kNN and Naïve Bayes.

A few valid reasons include:
1. inadequacy variable independence assumption (drawn from answer 5)
2. imbalanced priors in naive Bayes biasing MAP estimates
3. moderate data size, affecting pdf/pmf approximations in naive Bayes
4. inadequacy of the underlying multinomial assumption in naive Bayes (drawn from answer 5)
5. scarcity of specific class-conditional variable measurements (zero probs), affecting naive Bayes decisions
6. adequacy of local patterns in favor of kNN (empirical evidence drawn from pairwise similarities)

Common discounts:
- efficiency considerations were not fully counted as a reason
- reasons presented in favor of the worse performing method
- only one valid reason presented

END