Pattern Recognition/Pattern Recognition Techniques

2014/2015

Exame Recurso 26 June 2015 Duration: 2h30

Name: Tiago Silva **Number:** 2022216215 **Practical Class:**

AVISO

The Exam has a duration of 2h30m. The test is composed by five questions. The last question is a Matlab practical question. Each question must be answered in the framed box below it. Questions may be answered in Portuguese or English. This is a closed book test. You are allowed to use a calculator machine. As consultation you may use only 1 Page A4 with your own notes. Violation of the last rule ends up with exam cancellation, course failure and eventually you may be subject to disciplinary procedure. If you have any questions, you may ask. Good Luck!

Question	pts	Results	Graded by:
1)	20		
2)	20		
3)	10		
4)	20		
5)	30		

Graded by:

Name: **Number:** **Practical Class:**

Question 1 - Principal Component Analysis

□ **20 pts**

Consider the dataset represented in the table bellow described by feature 1 (f_1) and by feature 2 (f_2), and with class labels defined by Class (c), where "1" and "2" represents the positive and negative class, respectively.

Feature 1 (f_1)	2.5	0.5	2.2	1.9	3.1	2.3	2	1	1.5	1.1
Feature 2 (f_2)	2.4	0.7	2.9	2.2	3.0	2.7	1.6	1.1	1.6	0.9
Class (c)	1	2	1	2	1	1	2	1	2	2

1. In the context of PCA is there any unnecessary information in the table? Justify?

2. Compute the variance explained by the principal components, as well as the direction of projection. Represent them in the space spanned by f_1 and f_2.
   (Note: The eigenvalues of a given matrix $\mathbf{M}$ *can be found by solving* $|\lambda\mathbf{I} - \mathbf{M}| = 0$. *The eigenvectors are the vectors* $\mathbf{V}$ *that satisfy* $\mathbf{MV} = \lambda\mathbf{V}$. *)*

Your answer to 1):

The class information. PCA translates the system's axis to the directions that maximize variance not class distinction (like LDA). It's a method of unsupervised learning.

For PC1: $2\times 2 \quad \times 2\times 1$

$$\begin{bmatrix} -0.6134 & 0.6154 \\ 0.6154 & -0.5134 \end{bmatrix}\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = 0$$

$-0.6134x + 0.6154y = 0 \Rightarrow x = 1{,}003$

$0.6154x - 0.5134y = 0 \Rightarrow y = 1{,}2$

Direction of projection in the PC given by:

$$\begin{bmatrix} 1{,}003 \\ 1{,}2 \end{bmatrix}$$

1 2

Your answer to 2):

$$C = \begin{matrix} 1 \\ 2 \end{matrix}\begin{bmatrix} Var(1) & Cov(1,2) \\ Cov(2,1) & Var(2) \end{bmatrix} = \begin{bmatrix} 0.6166 & 0.6154 \\ 0.6154 & 0.7166 \end{bmatrix}$$

$$Var(1) = \frac{1}{9}\sum_{i=1}^{10}(x_i - \bar{x})^2 = 0.6166$$

($\bar{x} \downarrow 1.81$)

$$Var(2) = \frac{1}{9}\sum_{i=1}^{10}(x_i - 1.91)^2 = 0.7166$$

($1.91 \downarrow$, $1.81 \downarrow$)

$$c_{2,1} = \frac{1}{9}\sum_{k=1}^{10}(x_{k,i} - m_i)(x_{k,j} - m_j)$$

$$= \frac{1}{9}[0.3381 + 1.5851 + 0.3861 + 0.0261 + 1.4061 + 0.3871 - 0.0589 + 0.6561 + 0.0961 + 0.7171]$$

$$= 0{,}6154$$

$$c_{1,2} = 0{,}6154$$

PC1: 95.7 %

PC2: 4.3 %

$$|\lambda I - C| = 0 \Rightarrow \left|\begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} - \begin{bmatrix} 0.6166 & 0.6154 \\ 0.6154 & 0.7166 \end{bmatrix}\right| = 0$$

$$\Rightarrow \left|\begin{bmatrix} \lambda - 0.6166 & -0.6154 \\ -0.6154 & \lambda - 0.7166 \end{bmatrix}\right| = 0$$

$$\Rightarrow (\lambda - 0.6166)(\lambda - 0.7166) - 0.3787 = 0$$

$$\Rightarrow \lambda^2 - 0.7166\lambda - 0.6166\lambda + 0.442 - 0.3787 = 0$$

$$\Rightarrow \lambda^2 - 1.3332\lambda + 0.0682 = 0$$

$$\Rightarrow \lambda = \frac{1{,}3332 \pm \sqrt{1{,}777 - 0{,}273}}{2} \quad (\approx 1{,}224)$$

$\Rightarrow \lambda = 1{,}23$

$\lambda = 0.0546$

Name: **Number:** **Practical Class:**

Question 2 - Minimum distance classifiers

□ **20 pts**

Find the generic discriminant function of a minimum distance classifier when the Standardized Euclidean distance is considered. Consider as class prototypes the mean and that the squared Standardized Euclidean distance is given by:

$$d^2 = (\mathbf{x} - \mathbf{m}_k)^T \mathbf{D}^{-1} (\mathbf{x} - \mathbf{m}_k), \tag{1}$$

where D is the diagonal matrix with diagonal elements given by v_j^2, which denotes the variance of the j-th feature.

Your answer:

$$\hat{d}(x) = (x - m_k)^T D^{-1} (x - m_k)$$

$$= x^T D^{-1} x - x^T D^{-1} m_k - m_k^T D^{-1} x + m_k^T D^{-1} m_k$$

$$= x^T D^{-1} x - 2 m_k^T D^{-1} x + m_k^T D^{-1} m_k$$

$x^T D^{-1} m_k = m_k^T D^{-1} x$

$$= x^T D^{-1} x - 2 m_k^T D^{-1} \left(x + \frac{m_k}{2}\right)$$

$x^T D^{-1} x$ — Does not depend on the class

$g_k(x)$

Its the same as minimize : $-2 m_k^T D^{-1} \left(x + \frac{1}{2} m_k\right)$

Which is the same as maximize $g_k(x)$, the class discriminant

Erfior no eu

Name: **Number:** **Practical Class:**

Question 3 - kNN

☐ **10pts**

1. If a data set has significant class noise, do we want to use a smaller or larger value of k? Why? Explain graphically your answer.
2. Does kNN tend to overfit with smaller values of k or with larger values of k? Why? Explain graphically your answer.

Your answer to 1):

If the dataset has much noise, we typically want to use a larger value of k because if we don't, those outsiders will have great impact in the cluster calculation.

Your answer to 2):

Tends to overfit with smaller values of k because variance tends to increase, bias decreases.
High values of k are more Robust to noise.

Name: **Number:** **Practical Class:**

Question 4 - SVM

☐ **20pts**

1. What equations are used for classification in a SVM?
2. In a linear SVM describe the influence of the free parameter? Explain graphically your answer.

Your answer to 1):

Decision hyperplane : $w^T x_i + b = 0$

A test sample is classified according to the decision function's signal:

Linear SVM

$y_i(w^T x_i + b) \geq 1 - \boxed{\epsilon_i}$

Slack Variables (Allow violations of the margin constrains)

Kernel SVM

$$w = \sum_{i=1}^{N} a_i y_i x_i \qquad K(x_i, x) = x_i^T x$$

$$w^T x + b = \sum_{i=1}^{N} \alpha_i y_i (x_i^T x + b$$

$$y(x)\left[\sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b\right] \geq 1 - \epsilon_i$$

Your answer to 2):

A small C from $\min\left\{\frac{1}{2}\|w\|^2 + C\sum \varepsilon_i\right\}$ allows a larger margin which allows for more misclassifications and vice versa

$C = 100$

$C = 10$