Faculdade de Ciências e Tecnologia / UNL, 16 de Junho de 2011 (Época Normal)

Aprendizagem Automática e Data Mining

(consultation allowed; 3 hours)

I

1. Comment the following sentence:

> It is possible to extract association rules only from categorical data. Not from numerical data.

II

1 Consider the SVM cassifier when several classes are linearly separable. Classify using T (true) or F (false) the following sentences:

a) There is a security margin in the vectorial space of the attributes, between classes. That space is not occupied by any training set element.
b) All training set elements are used both in the training phase and test phase.
c) In the test phase, only the training elements that define the several frontiers between classes (the support vectors) are used to decide which classe must be assigned to a new element to be classified.
d) The training of the SVM lies only on the calculation of the weight that must be assigned to each attribute for each classe.

III

1. Consider the *k*-NN Classifier

Classify each sentence using T (true) or F (false)

a) It is a "lazy learning algorithm" type classifier bacause, due to its natture, it can not work with *fast* data structures.
b) The higher the *k* parameter, the higher the precision of the classifier whatever the circumstance.
c) It is a "lazy learning algorithm" type classifier because it does not learn with the classifications made in earlier classifications.
d) Usually, a low *k* value works better when it is frequent that more than one classe exist in small neighborhoods; in other words, when there is a strong class mixing through all the feature space.
e) None of the other sentences are true because *k*-NN is a non-supervised classifier.
f) If data distribuitions of the training set is known, the *k*_NN precision is theoretically unsurpassed.

IV

1. Consider criateria a) and b) for the calculation of the distance used in the *k*-NN classifier, for numerical data.

a) $d(x_i,x_j) = \sqrt{\sum_{a \in Atributos} d_a(x_i,x_j)^2}$ where $d_a(x_i,x_j) = \frac{|x_{i,a} - x_{j,a}|}{range(a)}$

b) $d(x_i,x_j) = \sqrt{\sum_{a \in Atributos} d_a(x_i,x_j)^2}$ where $d_a(x_i,x_j) = |x_{i,a} - x_{j,a}|$

Compare both criteria considering the relative importance of the features used in the classification process.

V

1. Consider that a minimum distance classifier may choose one from the following criteria: Euclidean distance (a), or Mahalanobis distance (b)

a) $distEucl(\vec{y},\vec{\mu}) = [\vec{y} - \vec{\mu}]^T [\vec{y} - \vec{\mu}]$

b) $distMahal(\vec{y},\vec{\mu}) = [\vec{y} - \vec{\mu}]^T \vec{\Sigma}^{-1} [\vec{y} - \vec{\mu}]$

What would be the best option in the following cases (explain why):

a) Data distributions of the training set are completely unknown.
b) Data distributions of the training set form hyper-ellipsoids in a *k*-dimention space, being *k* the number of features.
c) Data distributions of the training set form hyper-ellipsoids having the same shape and orientation, in a *k*-dimention space, being *k* the number of features.
d) The number of features are relatively high and data distributions of the training set form hyper-spheres of the same size in a *k*-dimention space, being *k* the number of features.

VI

1. Consider the Naive Bayes classifier.

a) Can it be used to work with non-categoric numerical data? Why?

b) Suppose you want to classify dogs according to their races. Consider the following sets of features: 1) length of front legs; length of rear legs; weight; length of ears; length of the fur (comprimento do pelo). And 2) length of front legs; weight; length of ears; length of the fur (comprimento do pelo). Which set must provide better precision? Why?

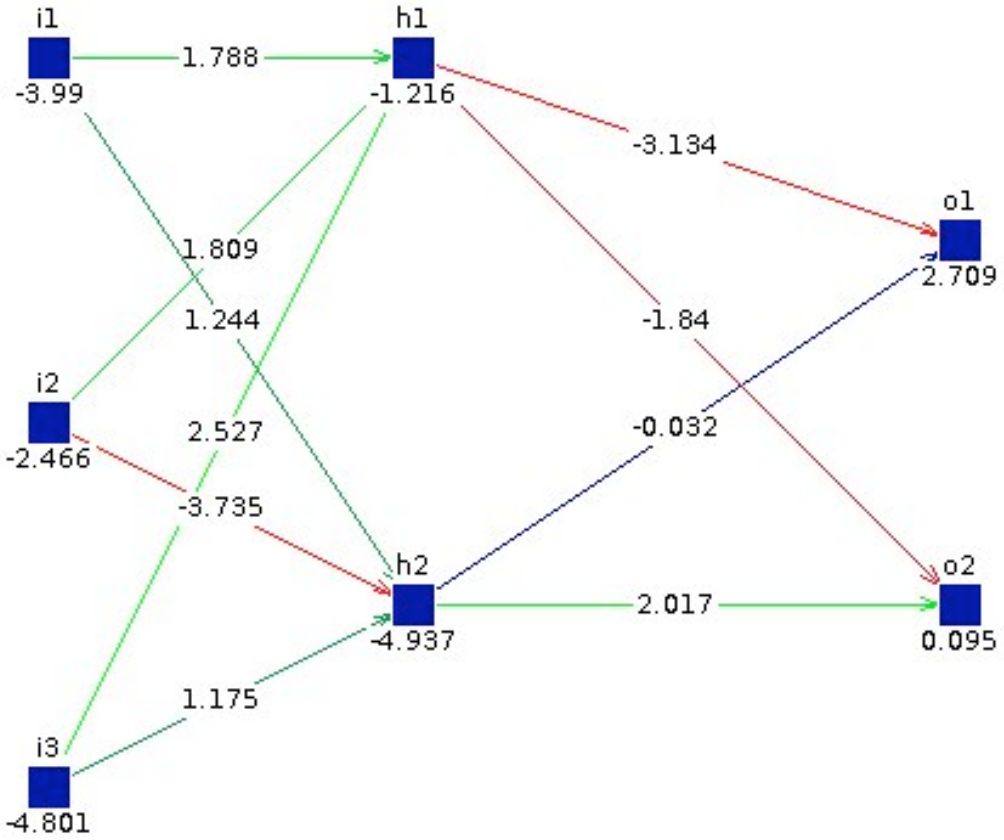
c) Why is it called “Naive”?.

VII

1. Suppose you know that a classifier has a precision of, at least, 90% when working with a certain application. What should be the size of a random set in order to form a test set, such that, with a 0.90 probability, the precision given by the classifier lies in an +/- 2% interval?

VIII

1. Please consider the following artificial neural network:



and input pattern **x**=(.3, -0.2, 0.4); **y**=(0, 1).

Using the usual neural network parameters for this type of network, please answer the following questions (including justification):

a) What is the *output* (o1, o2) for this MLP?
b) What is the value for updating the connection between i2 and h1.
c) If neurons h1 and h2 belong to the output layer (map) of a self-organizing feature map (SOM) with 5x8 neurons. Among h1 and h2 what would be the BMU and the update of the weight from i3 to that neuron if both neurons have map distance 2. Consider a fine-tuning train phase.

2. Please comment: "Assuming boolean (0 or 1) input, *for implementing a logical not(i1) AND i2 AND i3 in h1 we must use connections with weight 5 for positive literals (ie. I2 and i3) and weight --5 for negated literal (i1). Bias should be always set to -10*.".

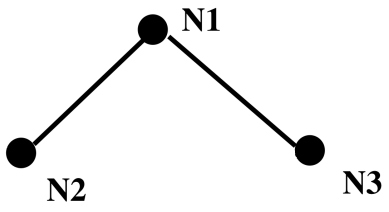
IX

1. Considering the SLIQ algorithm for inducing decision trees, please take into account the following information while inducing some decision tree:

(3)

Income	Class List Index
12	2
13	2
14	4
20	1
23	3
25	3
26	1
27	3
28	3
40	4
45	4
50	4

	Class	Leaf
1	A	N2
2	B	N2
3	A	N3
4	B	N3



a) Please present internal SLIQ algorithm histograms before and after presenting pattern (3).
b) Present calculated values and division criteria for node **N2**.

X

Classify using T (true) or F (false) each of the following sentences:

a) When using the single-link criterion, hierarchical clusters show greater detail, in comparison with complete-link.

b) *k*-means and Model-Based Cluster Analysis work assuming that clusters are hyper-spheres.

c) Model-Based Cluster Analysis works assuming that data distributions are multi-guassian.

d) It is arguable to consider that, approaches using *k*-means clustering are not completely unsupervised.

XI

1. Sort by descending order the following Multi-word Units / Relevant Expressions, considering how informative they are for an unsupervised classification by topics of documents.

Agricultura Biológica

technical data

in case of

Human Rights

Singing in the Rain

bank robbery

2. Comment the following sentence:

"economic crisis in" must not be considered a Multi-word Unit.