IR2 – Data Mining 2002–2003

Exercises IV

Peter Lucas
Institute for Computing and Information Sciences
University of Nijmegen

1 Evaluation

Exercise 1: evaluation of classifiers

Consider the following table of results obtained from running a test set through two different Bayesian classifiers, with underlying probability distribution equal to P_1 and P_2:

Outlook	Temperature	Humidity	Windy	Play	$P_1(play \mid \mathcal{E})$	$P_2(\text{play} \mid \mathcal{E})$
sunny	85	85	false	no	0.2	0.4
sunny	80	90	true	no	0.6	0.7
overcast	83	86	false	yes	0.9	1
rainy	70	96	false	yes	0.5	0.9
rainy	68	80	false	yes	0.6	0.3
rainy	65	70	true	no	0.3	0.2
overcast	64	65	true	yes	0.2	0.1
sunny	72	95	false	no	0.5	0.7

where '*play*' stands for Play = *yes* and $\mathcal{E}$ is evidence taken from the variables in the first 4 columns in the table.

a. Draw ROC curves for cut-off points (thresholds) equal to 0.3, 0.5 and 0.7. Which of the two classifiers performs best?

b. Next, determine the quality of the probability distributions for each of the Bayesian classifiers, and compare their quality.

Exercise 2: finding best models

The two methods normally used to determine how well a model fits given data are maximum likelihood estimates and Bayes' rule.

a. What are the assumptions underlying the two approaches?

b. What are the advantages and disadvantages of either of the two methods?

c. Which of the two methods would you prefer?

2 Unsupervised learning

Exercise 3

Show that weighted Euclidean distance

$$d^w(\mathbf{x}_i, \mathbf{x}_k) = \frac{\sum_{l=1}^{p} w_l (x_{i,l} - x_{k,l})^2}{\sum_{l=1}^{p} w_l}$$

satisfies

$$d^w(\mathbf{x}_i, \mathbf{x}_k) = d(\mathbf{z}_i, \mathbf{z}_k) = \sum_{l=1}^{p} (z_{i,l} - z_{k,l})^2$$

where

$$z_{i,l} = x_{i,l} \left(\frac{w_l}{\sum_{l=1}^{p} w_l} \right)^{1/2}$$

Thus weighted Euclidean distance based on $\mathbf{x}$ is equivalent to unweighted Euclidean distance based on $\mathbf{z}$.

Exercise 4

a. What is the main advantage of hierarchical clustering in comparison to K-means?

b. The cosine measure, $\cos \phi$, where ϕ is the angle between two vectors $\mathbf{x}$ and $\mathbf{x}'$, is a popular similarity measure. The Euclidean distance (either squared or not) is another popular measure, this time for dissimilarity. Make a comparison between these two measures.

Exercise 5

a. Under what conditions is it possible to view the learning of a Bayesian network as a form of unsupervised learning?

b. Assume that two variable A and B are independent. Do you expect that the Apriori-algorithm would put these variables with relevant value into one item set or not? Explain your answer.

c. Why is learning association rules much more difficult than learning classification rules, even when identical syntax is used for the resulting rules?