Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Thomas Seidl
Julian Busch, Evgeniy Faerman,
Florian Richter, Klaus Schmid

Knowledge Discovery in Databases

SS 2016

Exercise 9: Classification

Regarding tutorials on 22.06.-24.06.2016.

Exercise 9-1 Naive Bayes

The skiing season is open. To reliably decide when to go skiing and when not, you could use a classifier such as Naive Bayes. The classifier will be trained with your observations from the last year. Your notes include the following attributes:
The weather: The attribute `weather` can have the following three values: `sunny`, `rainy` and `snow`.
The snow level: The attribute `snow level` can have the following two values: ≥ 50 (There are at least 50 cm of snow) and < 50 (There are less than 50 cm of snow).

Assume you wanted to go skiing 8 times during the previous year. Here is the table with your decisions:

weather	**snow level**	**ski ?**
sunny	< 50	no
rainy	< 50	no
rainy	≥ 50	no
snow	≥ 50	yes
snow	< 50	no
sunny	≥ 50	yes
snow	≥ 50	yes
rainy	< 50	yes

(a) Compute the *a priori* probabilities for both classes `ski = yes` and `ski = no` (on the training set)!

(b) Compute the conditional distributions for the two classes for each attribute.

(c) Decide for the following weather and snow conditions, whether to go skiing or not! Use the Naive Bayes classificator for finding the decision.

	weather	**snow level**
day A	sunny	≥ 50
day B	rainy	< 50
day C	snow	< 50

Exercise 9-2 Linear Discriminant Classifier (SSE)

(a) Download the `trainingData.csv` file from the course website. This training data includes 2-dimensional points with two possible class labels 1 and 2. Implement a linear classifier and use it to compute the parameters w and w_0 w.r.t the training data.

(b) Download the `testData.csv` file from the course website. This data includes 2-dimensional points without class labels. Apply your linear classifier to the data with the parameters you computed in (a). Visualize your results and report the accuracy, precision, recall and F1 score achieved by your classifier.

Exercise 9-3 m-fold Cross Validation

Suppose, you have a 2-dimensional dataset consisting of 5 classes with 90 objects each, arranged as follows

x	**y**	**class_label**
x_0	y_0	0
$\vdots$	$\vdots$	$\vdots$
x_{89}	y_{89}	0
x_{90}	y_{90}	1
$\vdots$	$\vdots$	$\vdots$
x_{179}	y_{179}	1
$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$
x_{360}	y_{360}	4
$\vdots$	$\vdots$	$\vdots$
x_{449}	y_{449}	4

and that the classes are linearly separable. Suppose further, that someone has produced a poor implementation of the m-fold cross validation procedure and applied it in combination with a multiclass linear classifier to obtain the following results:

m	**accuracy**
2	20 %
3	40 %
5	0 %
6	100 %
10	100 %

What is the problem with the implementation of the m-fold cross validation? Describe and explain the result for each value of m in short and precise sentences. How could the implementation be improved?