

Submission sheet
Assignment IV

TASK 0: warm up

Number of instances:

Number of attributes:

Number of instances in each class: B M

TASK 1: basic k-NN classification

Accuracies (with confidence, where available):

1) Holdout

2) 10-fold cross-validation

3) Leave-one-out

4) 10-fold cross-validation, K=10

TASK 2: data scaling

Accuracy after scaling:

TASK 3: Feature selection

Accuracy after feature selection:

List of relevant attributes:

TASK 4: Combined approaches

Accuracy after rescaling and feature selection:

TASK 5: PCA

How many components are needed to explain 50% of the variance in the data?

Accuracy, varying the number of components:

TASK 6: optimizing the parameters – K

What is the best value for K among the ones you tested?

TASK 7: Decision tree

What is the number of nodes in the tree (min 20 instances per leaf)?

TASK 8: text data

What accuracy do you obtain with a 10-fold cross validation?

Have you found any text preprocessing operators or settings of the classifier leading to a better accuracy? Which ones?

Have you found any text preprocessing operators or settings of the classifier that you would have expected leading to a better accuracy, but in practice did not help? If yes, can you explain why?