Submission sheet
Assignment IV

TASK 0: warm up

Number of instances: 569

Number of attributes: 34

Number of instances in each class: B 212 M 357

TASK 1: basic k-NN classification

Accuracies (with confidence, where available):

1) Holdout 0.6199

2) 10-fold cross-validation 0.5369

3) Leave-one-out 0.5965

4) 10-fold cross-validation, K=10 0.6020

TASK 2: data scaling

Accuracy after scaling: 0.9532

TASK 3: Feature selection

Accuracy after feature selection: 0.9124

List of relevant attributes:

LARGE_compactness, LARGE_concave_points, LARGE_fractal_dimension, LARGE_radius, LARGE_smoothness, SD_concave_po

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?

2

Accuracy, varying the number of components:

1: 0.8768; 5: 0.9411; 10: 0.9352; 15: 0.9411; 20: 0.9411

TASK 6: optimizing the parameters – K

What is the best value for K among the ones you tested? 1

TASK 7: Decision tree

What is the number of nodes in the tree (min 20 instances per leaf)? 11

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?