Assignment no. 1

Machine Learning: Supervised Techniques
365.076 (1UE) WS 2017

Exercise 1 (30 points) Analyze data set 1 (available for download on JKU Moodle). The last column is the label column, the other columns are the input features. Use k-nearest neighbor classifiers with the following k's: 1, 3, 5, . . . , 47, 49, 51. In order to estimate the generalization error, use 10-fold cross validation to determine the average classification error for each k. Try to make a conclusion which k is the best choice. In order to do that, for each data set, visualize the error rates for each k. What do you observe? Repeat your analysis (1) after flipping labels randomly with a probability of 0.2, (2) after adding four noise features with random numbers that are uniformly distributed on the unit interval. What do you observe? Try to explain possible changes of the generalization error and the best k. Submit your program code, along with a short report that summarizes your results, visualizations, and interpretations.

Submission: electronically via Moodle:

`https://moodle.jku.at/jku2015/course/view.php?id=2634`

Please take the submission instructions into account! Deadline: Wednesday, November 8, 2017, 8:00am.