Introduction to Machine Learning **Fall Semester, 2013/14**

Homework 1: Oct 13th, 2013

Due: Oct 27th (See the submission guidelines in the course web site)

Theory Question

Let k-NN(S) be the k Nearest Neighbor classification algorithm on sample S, which takes the majority of the closest k points.

1. Show that if in both 1-NN(S_1) and 1-NN(S_2) the label of point x is positive, then in 1-NN($S_1 \cup S_2$) the label of x is positive.

2. Show an example such that in both 3-NN(S_1) and 3-NN(S_2) the label of x is positive, and in 3-NN($S_1 \cup S_2$) the label of x is negative.

Programming Assignment

Write a simple k Nearest Neighbor implementation (without using Matlab's built-in NN functions!). Run the implementation on the `glass` data set
(from: `http://archive.ics.uci.edu/ml/datasets/Glass+Identification`)

Estimate the performance of the k-NN algorithm with and without normalization[1] and across a range of values for k (from 1 to 25). Plot the accuracy, measured using 10 fold cross validation, as a function of k (with and without normalization of features).

10-fold cross validation means that you split the data into 10 equal size parts. You run 10 times: each time you train on different 9 parts and test on the remaining 10th part.

Briefly explain the results.

[1]Normalization: apply shifting (substruction of a constant) and scaling (multiplication by a constant) such that all features have the same range. The two constants per feature are estimated base on training data alone