Machine Learning Worksheet 2

Decision Trees and Nearest Neighbours

1 Decision Trees

In the unit square $[0;1] \times [0;1]$ you are given n non-overlapping points. Each point is labelled either a or b. Assume that each feature can be used for splitting the data multiple times in a decision tree. In this section you'll see why simple feature-wise (i.e. coordinate-wise) splitting of the data is not always the best approach to classification.

Problem 1. Show that a decision tree of depth at most $\lceil \log_2 n \rceil$ (which correctly labels all n points) exists. Assume that every point has a *unique* x and y coordinate. At each node the decision tree should only perform a binary split on a single coordinate. For splits you can choose among the standard operators (e.g. $<$, $>$, $=$, $\leq$, $\geq$, $\ldots$). Such a binary decision tree can have as many as n internal nodes (i.e. splits).

Problem 2. Describe in a few sentences a set of n points in $[0;1] \times [0;1]$, along with corresponding a or b labels, so that the smallest decision tree that correctly labels them all has at least $n-1$ *splits.*

Problem 3. Describe (no need to formally show!) n points and corresponding labels that can only be correctly labeled by a tree with at least $n-1$ splits, with the additional condition that the points labeled a and b must be *separable* by a straight line.

2 Nearest Neighbours

Problem 4. Consider the following definition of a distance between two points $\boldsymbol{x}$ and $\boldsymbol{y}$:

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sum_i \sigma_i (\boldsymbol{x}_i - \boldsymbol{y}_i)^2, \qquad \sigma_i > 0$$

Write this in the form of a *Mahalanobis distance*, i.e. use a symmetic matrix $\boldsymbol{\Sigma}$.

Problem 5. You are doing Nearest Neighbour classification with the standard euclidean distance function. Assume that the scale of one feature dominates (e.g. it is 1000 times larger than) the others. What happens? How can you compensate this problem by choosing a different distance metric?

Problem 6. Show that the approximation eq. (2) in the kNN slides holds and that the conclusion with respect to the limit case ($\sigma \to 0$) is correct. Assume that you have only two classes. Let N_0 be the number of exemplars in class 0, N_1 the number of exemplars in class 1 (this implies (see future lectures) that $p(c=0) = N_0/(N_0+N_1)$ and $p(c=1) = N_1/(N_0+N_1)$).

3 Random Projections

Problem 7. Prove that

$$p(h_{\boldsymbol{r}}(\boldsymbol{v}) = h_{\boldsymbol{r}}(\boldsymbol{u})) = 1 - \frac{\theta(\boldsymbol{v}, \boldsymbol{u})}{\pi}$$

where $h_{\boldsymbol{r}}$ and $\theta(\cdot, \cdot)$ are defined as in the lecture.

4 Neighbourhood Component Analysis

Problem 8. Compute the gradient of the NCA objective as given in the slides. *This may be a very difficult excercise if you don't have much practice with math → don't spend too much time on it!* The following matrix identity[1] is helpful:

$$\frac{\partial tr(X^T X A)}{\partial X} = X(A + A^T)$$

tr is the *trace* of a matrix. You may also want to use the shorthand $x_{ij} = (x_i - x_j)$.

[1] for more of these things: K. B. Petersen and M. S. Pedersen. 2008. The Matrix Cookbook. Technical Report.