Ludwig-Maximilians-Universität München
Institut für Informatik
Dr. Eirini Ntoutsi
Erich Schubert

Knowledge Discovery in Databases

SS 2012

Übungsblatt 6: Kernel Functions and Linear Regression

Aufgabe 6-1 *Support Vector Machines*

If learning a Support Vector Machine would solely try to minimize the number of misclassified training examples, what problem could potentially arise from that? How can this problem be resolved?

Aufgabe 6-2 *Kernel Functions*

As explained in the lecture, a kernel function ("kernel") is positive (semi-) definite. A matrix A is positive definite, if all the eigenvalues are non-negative, or in other words if for every $x \in \mathbb{R}^d$: $x^\top \cdot A \cdot x \geq 0$

Show that the following functions are kernel functions if x and $\hat{x}$ are vectors in $\mathbb{R}^d$:

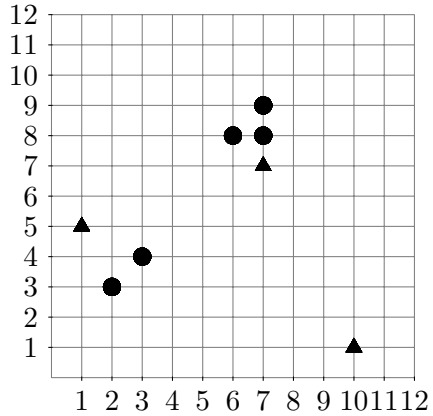
(a) $k_1(x, \hat{x}) = 1$

(b) $k_2(x, \hat{x}) = 3 \cdot x^\top \cdot \hat{x}$

(c) $k_3(x, \hat{x}) = 3 \cdot x^\top \cdot \hat{x} + 5$

Aufgabe 6-3 *k-means clustering*

Given the following data set with 8 objects (in $\mathbb{R}^2$):



In the following, compute complete partitionings of the data set into $k = 2$ clusters. As distance function use the Manhattan distance (L_1 norm), which is defined on two objects x, y as $L_1(x, y) = \sum_{i=1}^{d} |x_i - y_i|$

(a) Compute a partitioning into $k = 2$ clusters using the k-means algorithm *as introduced in the lecture* (which is the Lloyd variant). The initial assignment of objects is given using the triangle and circle markers. Start with computing the initial centroids, and draw the cluster assignments after each step and explain the step. Remember to use the L_1 norm for computing distances!

(b) Compute a partitioning into $k = 2$ clusters using the original MacQueen k-means. In contrast to the Lloyd version (in the lecture), objects are processed one at a time, assigned to the nearest mean, and then the mean is updated right away. The initial assignment of objects is again given using the triangle and circle markers. Start with computing the initial centroids, and draw the cluster assignments after each step and explain the step. Remember to use the L_1 norm for computing distances!

You can copy the last page of this exercise sheet multiple times if you need more space for sketching.

(c) Explain shortly why this variant of k-means depends on the ordering of the data points.

(d) Optional: also try k-medoids instead.