Small Exercises 2

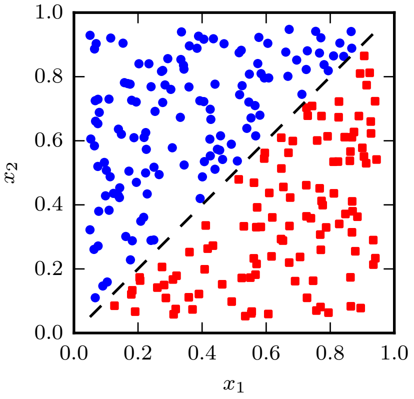
Decision Trees and k-Nearest Neighbors

These exercises are meant to prepare the inverted classroom lecture. Keep your answers short: two or three sentences, sometimes even less, should suffice.

1 Decision Trees

Note: There exist quite a few variants of decision tree algorithms. If not otherwise specified, the exercises refer to decision trees built with CART (Classification and Regression Trees) which feature binary splits.

Problem 1: The plot below shows data of two classes that can easily be separated by a single (diagonal) line. There exists a decision tree of depth 1 that classifies this dataset with 100% accuracy. True or False?



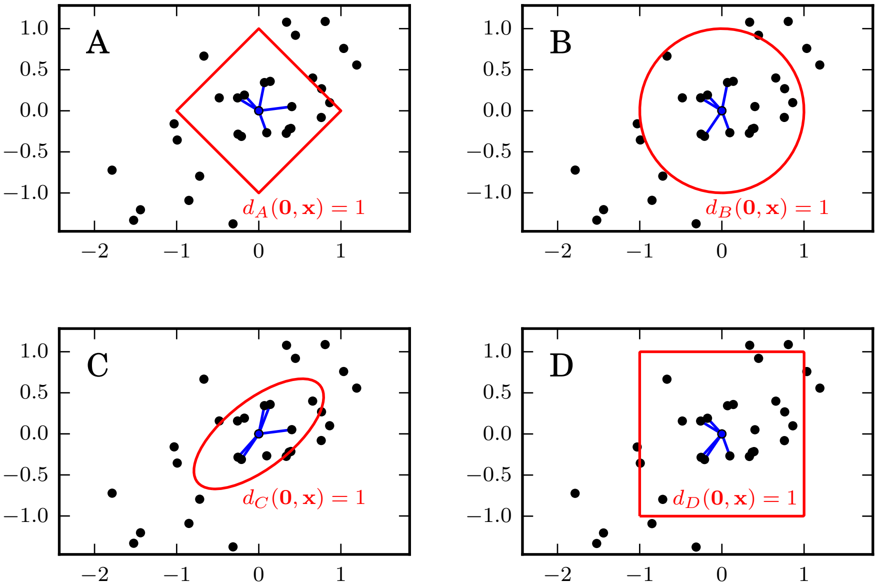
Problem 2: Explain the concept of *overfitting* in 140 characters or less.

Problem 3: What is the maximum value the entropy $H(\mathcal{D})$ can take for a labeled dataset $\mathcal{D}$ containing three classes ($C = \{c_1, c_2, c_3\}$)?

2 k-Nearest Neighbors

Problem 4: Consider a dataset with considerably more examples of one particular class. When k in k-NN is chosen to be a large number, i.e., close to the number of data points, all new points will be classified as that majority class. True or False?

Problem 5: The four plots below show the same data. The origin in each plot is connected to its five nearest neighbors according to some distance metric. The same metric is used to compute the set of points with distance 1 to the origin, which is visualized as a red shape. Assign the correct metric to each plot with 1 = Euclidean distance (L_2-Norm), 2 = Manhattan distance (L_1-Norm), 3 = Chebyshev distance (L_∞-Norm) and 4 = Mahalanobis distance.



Submit to `https://brml.org/ml1/hw1` *by 2016/11/07, 11:59am CET*