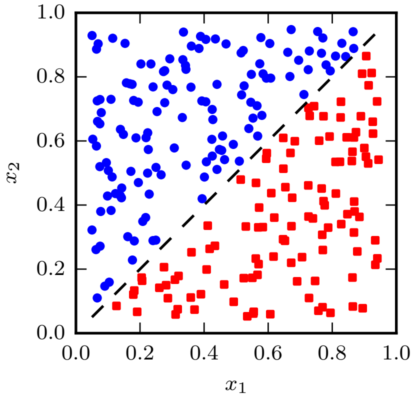
Small Exercises 2

Decision Trees and k-Nearest Neighbors

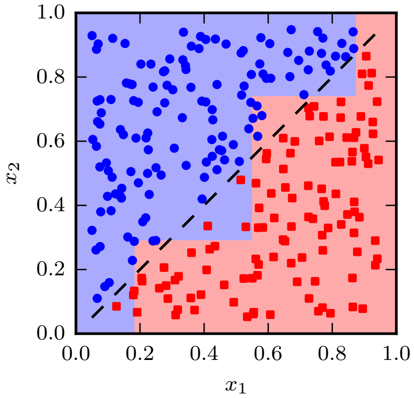
1 Decision Trees

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?



False, the feature test in a node can only use a single feature to split the training data. This leads to axis-parallel decision boundaries. Below you see the decision boundaries for a tree of depth 3. It classifies the dataset with 92.8% accuracy.



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

Overfitting occurs when a often excessively complex model closely fits the training data, but does not generalize to new unseen data.

(This explanation has 132 characters.)

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

Entropy assumes its maximum when each class occurs with the same frequency:

$$p(c_1) = p(c_2) = p(c_3) = \frac{1}{3}$$

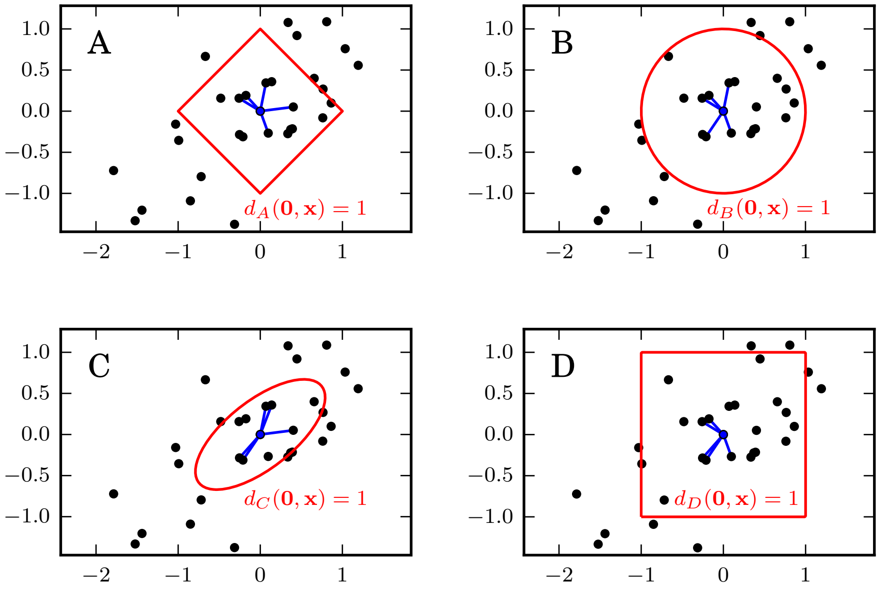
$$\Rightarrow H(\mathcal{D}) = -\sum_{c_i \in C} p(c_i) \log_2 p(c_i) = -3 \times \frac{1}{3} \log_2 \frac{1}{3} = -\log_2 \frac{1}{3} \approx 1.58496 \text{ bits.}$$

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?

True. When k is close to the number of data points, the neighborhood of a new point contains almost all points in the training set regardless of their distance. The majority class in the neighborhood is thus equal to the majority class in the dataset.

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.



The metric can be determined by the shape of set of points within distance 1 of the origin. The correct assignments are:

- $A = 2$ Manhattan distance, L_1-Norm:

  $$d_A(\boldsymbol{u}, \boldsymbol{v}) = \textstyle\sum_i |u_i - v_i|$$

  We know that all blocks a taxi driver can reach with the same number of steps lie on a diagonal, leading to the rotated square.

- $B = 1$ Euclidean distance, L_2-Norm:

  $$d_B(\boldsymbol{u}, \boldsymbol{v}) = \sqrt{\textstyle\sum_i (u_i - v_i)^2} = \sqrt{(\boldsymbol{u} - \boldsymbol{v})^T (\boldsymbol{u} - \boldsymbol{v})}$$

  All dimensions of the data are treated equally. In 2D this leads to a circle for the set of points with the same distance.

- $C = 4$ Mahalanobis distance:

  $$d_A(\boldsymbol{u}, \boldsymbol{v}) = \sqrt{(\boldsymbol{u} - \boldsymbol{v})^T \Sigma^{-1} (\boldsymbol{u} - \boldsymbol{v})}$$

  Mahalanobis distance consists of a transformation into a different space and then computation of Euclidian distance in that transformed space. Here, the covariance matrix of the dataset is used for Σ. Distances along the axis with highest variance are treated as less significant.

- $D = 3$ Chebyshev distance, L_∞-Norm:

  $$d_D(\boldsymbol{u}, \boldsymbol{v}) = \max_i |u_i - v_i|$$

  Only the dimension for which the difference is largest is relevant for this metric.