Q1: Semisupervised clustering F1-measure

The figures above show a two-class dataset. The classes are shown by the shapes. This dataset was subject to a clustering algorithm whose result is shown using the dotted ellipses. In each case, the algorithm returned two clusters as shown.

a) Compute the semi-supervised clustering F1-Measure in each case. The labeled data points for computing the F1-Measure are shown in color. To recall, the semi-supervised clustering F1-Measure is given by

\[ F(C, K) = \frac{\sum_{c_i \in C} \left( \frac{|c_i|}{M} \max_k \frac{1}{|k|} \right)}{\sum_{c_i \in C} \left( \frac{|c_i|}{M} \max_k \frac{1}{|k|} \right)} \]

\[ F(c_i, k) = \frac{2R(c_i, k)P(c_i, k)}{R(c_i, k) + P(c_i, k)} \]

\[ R(c_i, k) = \frac{n_{ik}}{|c_i|} \]

\[ P(c_i, k) = \frac{n_{ik}}{|k|} \]

Where M is the total number of labeled datapoints, k indexes the cluster number, C = \{c_i\} is the set of classes, n_{ik} is the number of datapoints of class c_i in cluster k, |c_i| is the number of datapoints in class c_i and |k| is the number of datapoints in cluster k.

b) Compare result of 1a with 1b and discuss the effect of the percentage of labeled data. Compare 2a with 2b and discuss the effect of unbalanced classes.
Q2: Mixture of Gaussians and Baye’s classification

1) Consider a two class classification problem where each class is modeled with a single Gauss function aligned with the X-Y axes. Assume that the two covariance matrices (for the two classes) are the same. Draw the classification boundary and show how the boundary moves as a function of the number of datapoints in each class. Consider a ratio of 1:1, 1:2 and 1:4, respectively, for the number of datapoints in class 1 versus the number of datapoints in class 2. Assume that the datapoints in each case have been generated following a Gauss distribution in each class.

2) Consider the example below of a two class dataset where each class is represented by a 2D Gaussian function. To which class will points 1, 2, 3 and 4 (drawn in the figure) be assigned?

3) Draw a 2D example of a distribution of datapoints fit with a particular choice of GMM parameters that leads to overfitting. Discuss overfitting versus generalization. Discuss the consequences of overfitting when doing classification for a 2-class problem.