Uppsala universitet, 2020
1TD184 Optimisation

Assignment 2 – intermediate option

# Assignment 2: Breast cancer diagnosis using a Support Vector Machine

*This assignment introduces a concept from automatic pattern classification, an area in rapid development. Here is presented a useful standard technique for "pass–or–fail" discrimination based on previous knowledge: the Support Vector Machine. The method is applied on real medical data made available by the University of Wisconsin.*

## Background

Suppose that you have a set of $m$ data points representing objects that you have classified into two distinct categories. It can be measurements of radius, area, smoothness etc. for tumor cells that have been classified as either malignant or benign. If $n$ properties are measured for each cell, plotting the data points in $\Re^n$ would form two clusters representing either benign or malignant cells. Pattern classification deals with the problem of setting up a rule by which you can judge to which class you should attribute a new data point so that information from a medical examination can be used to judge whether a patient has the disease or not. In its simplest form, a linear hyper-plane is used to separate one class from the other. The coefficients of the hyper-plane constitute the rule from which automatic classification is done; Which side of the plane a new data point corresponds to determines its classification. The range of applications is enormous: in-line quality control in manufacturing processes, face recognition and combustion engine knock detection are only a few scattered examples.

## The Support Vector Machine (SVM)

A two dimensional illustration of the clustering problem is presented in Fig. 1. The black and the white circles represent known data points classified as either having the property $y_i = -1$ or $y_i = 1$, respectively. These points are called the *training set.* Suppose that it is possible to find a line[1] (a hyper-plane in the general case) $\boldsymbol{w}^T\boldsymbol{x} + b = 0$ separating the two clusters, like the one indicated in the right figure. A new data point would then be classified as either "black" or "white" depending on the sign of $\boldsymbol{w}^T\boldsymbol{x} + b$. The problem is to find good values of $\boldsymbol{w}$ and $b$. Ideally, we would like to have a total separation between the points, so that for all points in the training set

$$\begin{aligned} \boldsymbol{w}^T\boldsymbol{x}_i + b &\geq 1 \quad \text{for} \quad y_i = 1, \text{ (i.e. the white circles)} \qquad (1)\\ \boldsymbol{w}^T\boldsymbol{x}_i + b &\leq -1 \quad \text{for} \quad y_i = -1 \text{ (i.e. the black circles).} \qquad (2) \end{aligned}$$

[1] There are also nonlinear versions of the SVM that can be better suited for some problems.

Then every point on or north-east of the line $\boldsymbol{w}^T\boldsymbol{x} + b = 1$ belongs to the class of white circles, and all points on or south-west of $\boldsymbol{w}^T\boldsymbol{x} + b = -1$ belong to the class of black points. The coefficients of the right hand side are not special in any way. They merely represent a scaling of the problem, but are conventionally chosen to the values ±1.

A reasonable requirement on $\boldsymbol{w}$ and $b$ is that the line $\boldsymbol{w}^T\boldsymbol{x} + b = 0$ they define should separate the two classes as distinctly as possible. This occurs when the distance between the lines $\boldsymbol{w}^T\boldsymbol{x} + b = 1$ and $\boldsymbol{w}^T\boldsymbol{x} + b = -1$, the *separation margin*, is large as possible. From basic linear algebra it can be shown that this distance is $2/|\boldsymbol{w}|$. Our requirements can now be formulated as the optimisation problem

$$\begin{aligned} \text{minimise } f(\boldsymbol{w}, b) &= \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} \qquad (3) \\ \text{subject to } y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) &\geq 1, \;\; i = 1 \ldots m. \qquad (4) \end{aligned}$$

Some of the points will lie on the dashed lines/hyper-planes. These are the so-called *support vectors*, and the *support vector machine* is nothing but the mathematical rule by which new data points can be classified.

The presented approach works as long as the data clusters are separable by a hyperplane. In reality, this is not always the case. Some measurement points will be "outliers", there can be misclassifications etc. bringing about a situation where there is no clear demarcation line between the classes. In order to handle also this situation, we need to allow for some of the points to "trespass" the separating plane and violate Eq. (1)–(2), but of course we would like to find the separating plane that requires as little violation as possible. One way of doing this is to relax the requirement of Eq. (1)–(2) by changing them to

$$\begin{aligned} \boldsymbol{w}^T\boldsymbol{x}_i + b &\geq 1 - \xi_i \quad \text{for} \quad y_i = 1, \text{ (i.e. the white circles)} \qquad (5) \\ \boldsymbol{w}^T\boldsymbol{x}_i + b &\leq -1 + \xi_i \quad \text{for} \quad y_i = -1 \text{ (i.e. the black circles)}, \qquad (6) \end{aligned}$$

where $\xi_i \geq 0$, which effectively moves the two dashed lines against and past each other. In addition, we add a term in the objective function of Eq. (3) that penalises the violation, so that we get the new optimisation problem

$$\begin{aligned} \text{minimise } f(\boldsymbol{w}, b) &= \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{m}\xi_i \\ \text{subject to } y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) &\geq 1 - \xi_i, \;\; i = 1 \ldots m, \\ & \quad \xi_i \geq 0. \end{aligned}$$

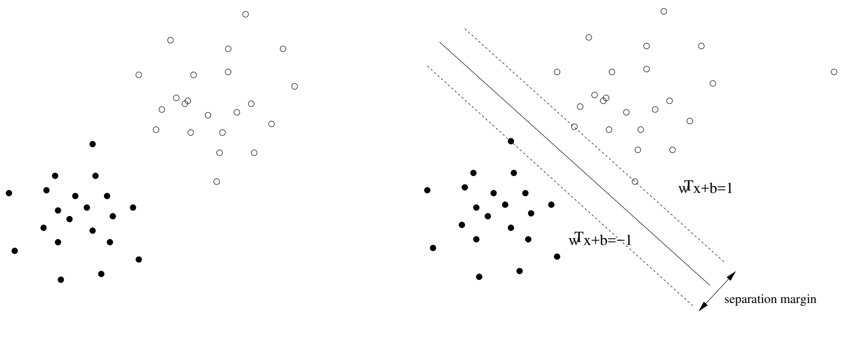

Figure 1: a) Two clusters of data points. b) Clusters and separating plane.

The value of the constant $C$ determines how much emphasis we should put on violation of separation and has to be chosen empirically.

## Tasks

Before you even read through the tasks, think the problem through for yourself and try to figure out how to interpret the problem in the language and variable names of a constrained optimisation problem from the course.

1. Set up the optimisation problem for the general, non-separable case. You will need to stack the unknown parameters into one vector of unknowns and to construct a few auxiliary matrices in order to formulate the problem on matrix form.

2. Download the Wisconsin Breast Cancer Database from http://www.siam.org/books/ot108. There are two files. One (wdbc.dat) is the data and the other one (wdcb.names) contains explaining information. There are $m = 569$ data points with $n = 30$ measured properties. Each data point is represented by a row in the data file. The first two columns present the patient number, and the classification "M" or "B" depending on whether the cells were malignant or benign.

3. Convert the data file to a .dat file that you can load in Matlab (e.g. using a text editor). You may change "M" and "B" to either +1 or -1. This column will contain the $y_i$, while columns 3–32 contains each $x_i$.

4. Write a Matlab programme that sets up the appropriate matrices and solves the problem using `quadprog`. Use the first 500 patients, or better yet an arbitrary subset, in the data base as your training set. Try initially with $C = 1000$.

5. Use your SVM to predict whether the remaining 69 patients (the *test set*) have malignant or benign cells. Compare the results with the actual diagnosis for each patient.

6. Evaluate the accuracy (proportion of correct predictions), the sensitivity (proportion of positive diagnoses for patients with malignant cells) and the specificity (proportion of negative diagnoses for patients without the disease). Try to find a good value of $C$.