EXERCISE SESSION III: MACHINE LEARNING COURSE – EDOC - 2012

Question 1: SVM - Concepts

i) Consider the two-class distribution given in Figure 1. Illustrate the result of performing classification when using **RBF** to separate the two classes. In particular:
   a. Draw the resulting hyperplanes in the original space and the support vectors.
   b. Discuss the number and relative values of non-zero Lagrange multipliers and their role in determining the separating line.

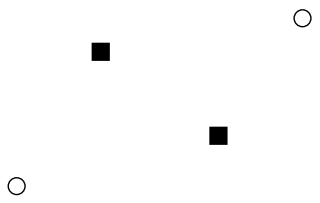


Figure 1: Points in class y=-1 and class y=+1 are denoted with circles and square respectively.

ii) Which choice of kernel in SVM would fail to separate the group of points in Figure 1?

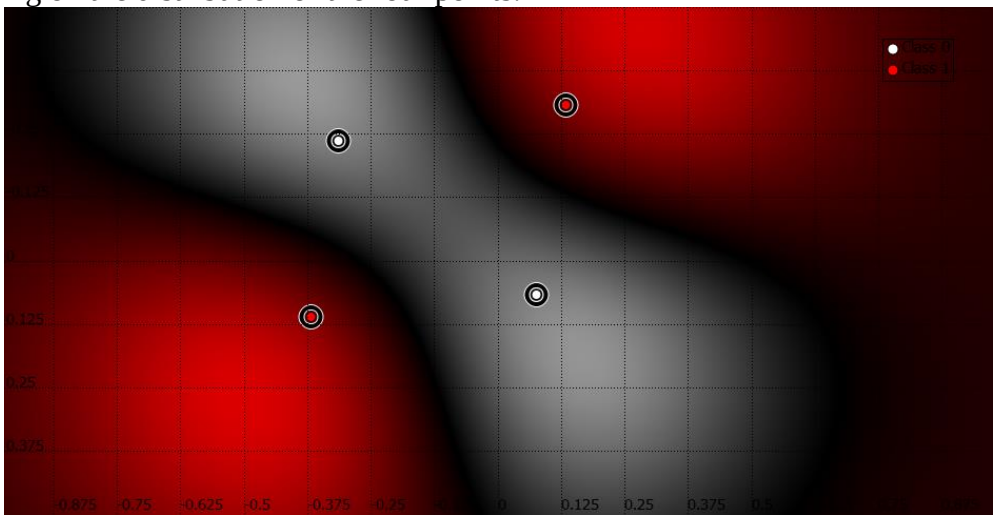
iii) **Give an example** of dataset where SVM with RBF kernel would not be able to separate perfectly the two classes (no matter which kernel width is used). Here, again we consider perfect classification with no slack variables.

iv) Discuss question i) with a polynomial kernel. Pay attention to where the origin is.

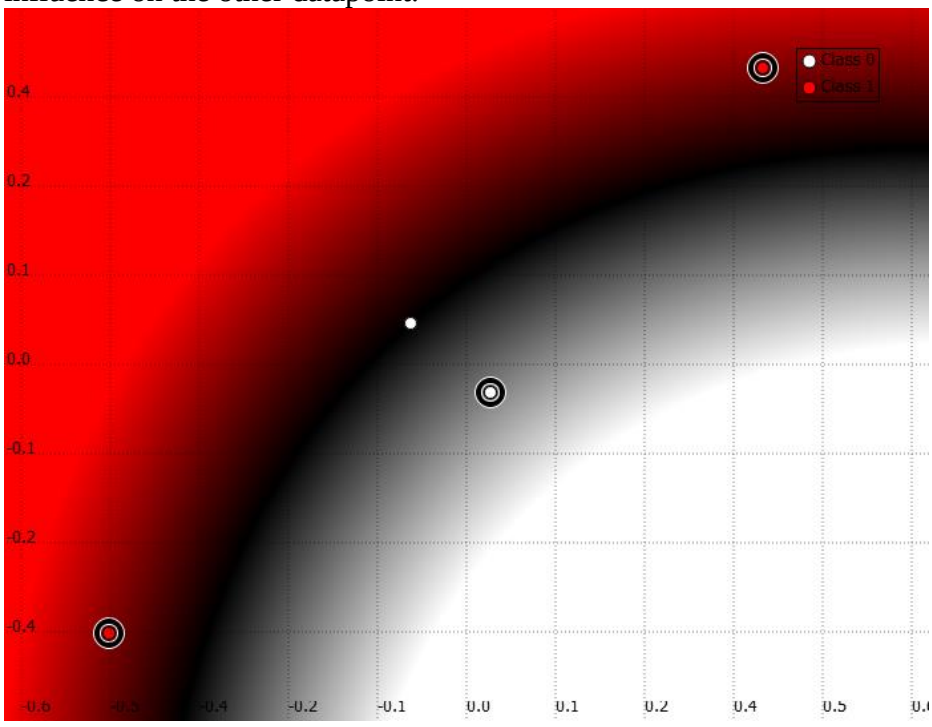
v) Construct an example of dataset where SVM with polynomial kernel with p=2 would not be able to separate the two classes. Here, we consider perfect classification with no slack variables.

Solutions:

i. a/b. With RBF kernel, one needs either 3 or all the 4 datapoints for support vector, depending on the distribution of the four points.



When two points of one class are close, one support vector can have enough influence on the other datapoint.



The Lagrange multipliers balance the influence of each support vector. When taking pairs of support vectors with local influence, the alpha-s determine the distance of the contour lines to the two support vectors.

ii. Polynomial kernel with p=1 would fail to separate the two classes as it would yield a single line. A polynomial kernel of order p=2 would be able to separate the datapoints at the condition that the points are centered or at least that the origin is near their center. Assuming small enough kernel width, an rbf would be successful at separating the datapoints.

iii. SVM will fail to separate data only when two datapoints overlap perfectly. A RBF kernel can allow arbitrary classification, as long as you give enough penalties with C. At worse, all datapoints become support vectors and one picks a tiny kernel width.

iv. We try to see what type of separating hyperplane can be obtained from two support vectors $v_1 = \begin{bmatrix} a_1 \\ b_1 \end{bmatrix}$ and $v_2 = \begin{bmatrix} a_2 \\ b_2 \end{bmatrix}$. Since there are only two support vectors, we have:

$\alpha_1 = \alpha_2 = \alpha$.

The classifier function is:

$$\alpha_1 \langle X, v_1 \rangle^2 - \alpha_2 \langle X, v_2 \rangle^2 + b = 0$$

$$\Rightarrow \alpha_1 \left(X^T v_1 \right)^2 - \alpha_2 \left(X^T v_2 \right)^2 + b = 0$$

$$\Rightarrow \left(x a_1 + y b_1 \right)^2 - \left(x a_2 + y b_2 \right)^2 + \frac{b}{\alpha} = 0 \qquad (1.1)$$

$$\Rightarrow x^2 (a_1^2 - a_2^2) + y^2 (b_1^2 - b_2^2) + 2xy(a_1 b_1 - a_2 b_2) - \frac{b}{\alpha} = 0$$

This equation is a hyperbola if the determinant D is negative:

$$D = \det \begin{vmatrix} (a_1^2 - a_2^2) & 2(a_1 b_1 - a_2 b_2) \\ 2(a_1 b_1 - a_2 b_2) & (b_1^2 - b_2^2) \end{vmatrix}$$

$$D = (a_1^2 - a_2^2)(b_1^2 - b_2^2) - 4(a_1 b_1 - a_2 b_2)$$

$$= -a_1^2 b_2^2 - a_2^2 b_1^2 - 2a_1 b_1 a_2 b_2$$

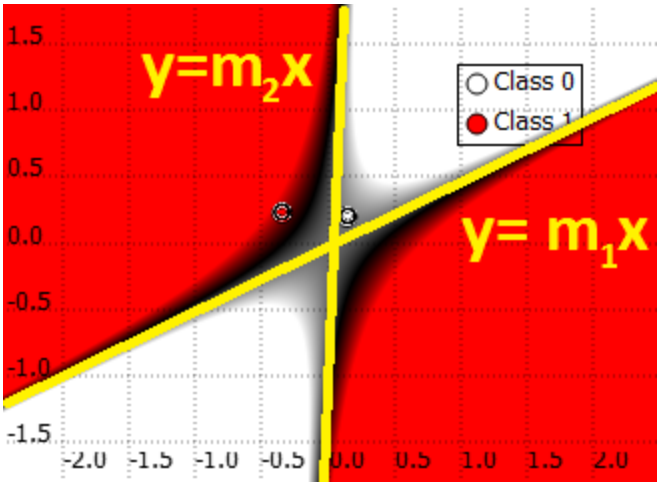
$$= -(a_1 b_2 + a_2 b_1)^2 < 0$$

Calculate the slope of the asymptote: from the equation(1.1), we can obtain:

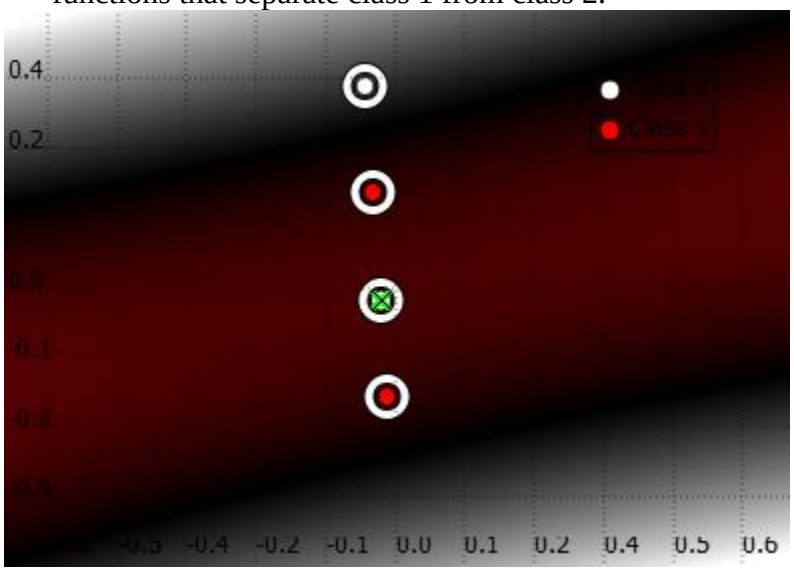
$\left(x a_1 + y b_1 \right) = \pm \left(x a_2 + y b_2 \right)$ for b=0.

$$\Rightarrow \qquad y = \frac{a_1 - a_2}{b_2 - b_1} x = \mathrm{m}_1 x$$

and $$y = \frac{a_1 + a_2}{-b_2 - b_1} x = \mathrm{m}_2 x$$



v. Below is such an example. One cannot construct a combination of hyperbolic functions that separate class 1 from class 2.



Question 2: SVM – error bound

Recall that the inequality constraints for the non-separable case in SVM are given by:

$$x_i \cdot w + b \geq +1 - \xi_i \quad \text{for } y_i = +1$$
$$x_i \cdot w + b \leq -1 + \xi_i \quad \text{for } y_i = -1$$
$$\xi_i \geq 0 \quad \forall \text{i}$$

where $x_i, i = 1,..M$ are the M data points with associated labels $y_i \in \{-1;+1\}$ and associated margins ξ_i. Show that $\sum_{i=1}^{M} \xi_i$ is an upper bound on the error of the classifier (The error is 1 if a point is misclassified, 0 otherwise).

Solution:

$$E_i = \frac{1}{2}\left|y_i - sign\left(x_i w + b\right)\right| \neq 0 \text{ iff } \xi_i > 1$$
$$\Rightarrow E = \sum_{i=1}^{M} E_i = \sum_{i=1,\xi_i>1}^{M} E_i \leq \sum_{i=1,\xi_i>1}^{M} \xi_i$$
$$\xi_i \geq 0, \forall i = 1,...,M \Rightarrow E \leq \sum_{i=1,\xi_i>1}^{M} \xi_i \leq \sum_{i=1}^{M} \xi_i$$

Question 3: ν-SVM (to be done at home)

Explain why ν is a lower bound on the fraction of support vectors in support vector classification. Recall that the dual of ν-SVM is given by:

$$\text{minimize } \hat{L}(\alpha_i) = -\frac{1}{2}\sum_{i,j=1}^{M} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j)$$

subject to

$$\sum_{i=1}^{M} \alpha_i y_i = 0 \quad ; \quad 0 \leq \alpha_i \leq C/M \quad ; \quad \sum_{i=1}^{M} \alpha_i \geq C\nu$$

Solution:

The left term in the constraint: $\sum_{i=1}^{M} \alpha_i \geq C\nu$ can be as high as: $\sum_{i=1,\alpha_i>0}^{M} \frac{C}{M}$ in the case where all the support vectors ($\alpha_i > 0$) have their maximum value $\frac{C}{M}$ (see second constraint). In that case, if p is the number of support vectors, the constraint gives:

$$p\frac{C}{M} \geq C\nu$$

And we have:

$$\frac{p}{M} \geq \nu$$

Which means that ν is a lower bound on the fraction of points that are support vectors.