1 SVM Decision Boundaries

Recall that the soft-margin primal SVM problem is

$$\begin{aligned}\min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}\mathbf{w}\cdot\mathbf{w} + C\sum_{i=1}^{n}\xi_i \\ & \text{s.t.} \quad \xi_i \geq 0 \quad \forall i = 1,\ldots,n \\ & (\mathbf{w}\cdot\mathbf{x}_i + b)y_i \geq (1-\xi_i) \quad \forall i = 1,\ldots,n.\end{aligned}$$

For hard-margin primal SVM, $\xi_i = 0$, $\forall i$. We can get the kernel SVM by taking the dual of the primal problem and then replace the product of $\mathbf{x}_i \cdot \mathbf{x}_j$ by $k(\mathbf{x}_i, \mathbf{x}_j)$, where $k(.,.)$ can be any kernel function:

$$\max_{\alpha} -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j k(\mathbf{x}_i,\mathbf{x}_j) + \sum_{i=1}^{n}\alpha_i$$

$$\text{s.t.} \ \sum_{i=1}^{n}\alpha_i y_i = 0, \forall i = 1,2,...,n$$

$$\alpha_i \geq 0, \forall i = 1,2,...,n$$

Figure 1 plots SVM decision boundaries resulting from using different kernels and/or different slack penalties. In Figure 1, there are two classes of training data, with labels $y_i \in \{-1, 1\}$, represented by circles and squares respectively. The SOLID circles and squares represent the support vectors. Match each plot in Figure 1 with the letter of the optimization problem below.

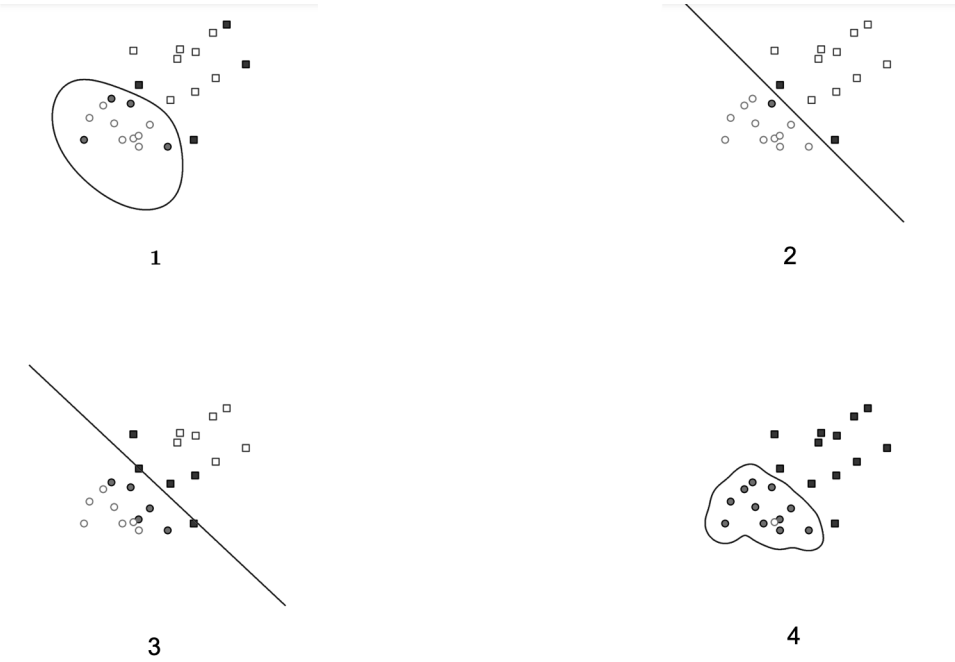


Figure 1: Induced Decision Boundaries

(a) A soft-margin linear SVM with C = 0.1.

(b) A soft-margin linear SVM with C = 10.

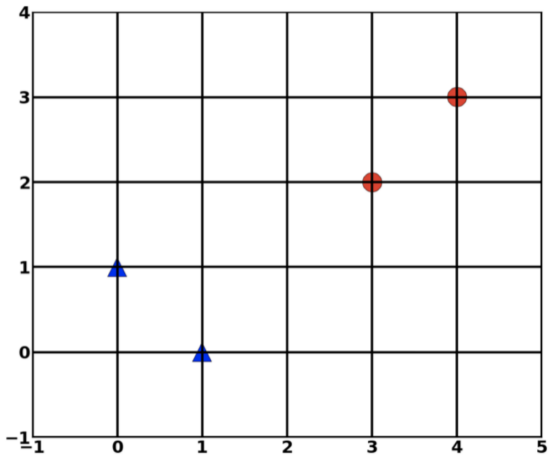
(c) A hard-margin kernel SVM with $K(\mathbf{u},\mathbf{v}) = \exp\left(-\frac{1}{4}\|\mathbf{u}-\mathbf{v}\|^2\right)$

(d) A hard-margin kernel SVM with $K(\mathbf{u},\mathbf{v}) = \exp\left(-4\|\mathbf{u}-\mathbf{v}\|^2\right)$

2 Hard-Margin SVMs

Assume we are given dataset $D = \{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^d$ and $y_i \in \{+1, -1\}$.
In SVM the goal is to find some hyperplane which separates the positive from from negative examples, such that the margin (the minimum distance from the decision boundary to the training points) is maximzed. Let the equation for the hyperplane by $w^T x + b = 0$.

(a) You are presented with the following set of data (triangle = -1, circle = $+1$):



The SVM hyperplane with maximum margin has equation $w^\top x + b = \frac{1}{2}\mathbf{x_1} + \frac{1}{2}\mathbf{x_2} - \frac{3}{2}$.

(i) Draw the decision boundary. Which points are support vectors?

(ii) What is the distance of the hyperplane to these support vectors?

(b) Let's try to measure the width of the SVM slab (we assume that it was fitted to linearly separable data). We can do this by measuring the distance from one of the support vectors, say $\mathbf{x}_+$, to the plane $\mathbf{w}^\mathbf{T}\mathbf{x} + b = 0$. Since the equation of the plane is $\mathbf{w}^\mathbf{T}\mathbf{x} + b = 0$ and since $c\mathbf{w}^\mathbf{T}\mathbf{x} + b = 0$ defines the same plane, we have the freedom to choose the normalization of $\mathbf{w}$. Let us choose normalization such that $\mathbf{w}^\mathbf{T}\mathbf{x}_+ + b = +1$ and $\mathbf{w}^\mathbf{T}\mathbf{x}_- + b = -1$, for the positive and negative support vectors respectively. Show that the width of an SVM slab with linearly separable data is $\frac{2}{\|\mathbf{w}\|}$.

3 Soft-Margin SVMs

For this question, we will be considering the kernelized version of the soft-margin SVM. Recall from class that the primal form is given as:

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}||\mathbf{w}||_2^2 + C\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & y_i\left(\mathbf{w}^T\phi(\mathbf{x}_i) + b\right) \geq 1 - \xi_i, \quad \forall i \in \{1, ..., n\} \\ & \xi_i \geq 0, \quad \forall i \in \{1, ..., n\} \end{aligned}$$

1. Write the Lagrangian for this SVM using $\alpha_i \geq 0$ as the dual variables on the first set of constraints and $\eta_i \geq 0$ as the dual variables on the second set of constraints.

2. Give the partial derivative of the Lagrangian with respect to each primal variable and set each partial derivative equal to zero.

3. Utilizing the expressions derived in the previous part, convert the Lagrangian into an expression for $J(\alpha)$ in terms of just the α_i dual variables, the data y_i and $\mathbf{x}_i$, and the kernel function $k(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x})^T\phi(\mathbf{z})$. Do not include $\phi(\cdot)$ in your final answer.

4. Write down the dual form of the SVM objective for the kernelized version of soft-margin SVM using results from previous parts.

5. Explain how the solution of kernelized soft-margin SVM can be used at test time to make prediction for a test point $\mathbf{x}$?

4 Kernel Trick

1. Suppose we have two feature vectors:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \; y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

If we were using an SVM we would want to

(a) Broadcast the vectors to a higher dimensional space

(b) Take their dot product

Say we want to broadcast the vectors to polynomials of degree 2. What would the resulting x and y vectors look like? How many steps would be involved in computing $x \cdot y$?

2. Recall the formulation for SVM:

$$\max_{\alpha} \sum_i \alpha - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x_i} \cdot \mathbf{x_j}$$

such that $\sum_i \alpha_i y_i = 0$ and $\alpha_i > 0$. 0 How many dot products do we do each time we compute the objective value?

3. Now suppose we have a data set with 10 features, and we wish to use polynomial features of degree 3. How many features are in the new data set? How many multiplication operations does it take to compute the dot product of two data vectors?

4. Now suppose we use the kernel trick. How many multiplications do we do now?