Support Vector Machine

Problem 1. Consider a binary classification problem in one-dimensional space where the sample contains four data points $S = \{(1, -1), (-1, -1), (2, 1), (-2,1)\}$ as shown in Fig. 1.

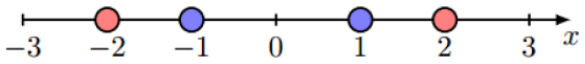


Figure 1: Red points represent instances from class +1 and blue points represent instances from class -1.

a) Define $H_t = [t, \infty)$. Consider a class of linear separators $\mathrm{H} = \{H_t : t \in \mathbb{R}\}$, i.e., for $\forall H_t \in \mathrm{H}, H_t(x) = 1 \; if \; x \geq t$ otherwise -1. Is there any linear separator $H_t \in \mathrm{H}$ that achieves 0 classification error on this sample? If yes, show one of the linear separators that achieves 0 classification error on this example. If not, briefly explain why there cannot be such linear separator. (10 points)

b) Now consider a feature map $\phi : \mathbb{R} \to \mathbb{R}^2$ where $\phi(x) = (x, x^2)$. Apply the feature map to all the instances in sample S to generate a transformed sample $\acute{S} = \{(\phi(x), y) : (x, y) \in S\}$. Let $\acute{H} = \{ax_1 + ax_2 + c \geq 0 : a^2 + b^2 \neq 0\}$ be a collection of half-spaces in R^2. More specifically, $H_{a,b,c}\big((x_1, x_2)\big) = 1$ if $ax_1 + ax_2 + c \geq 0$ otherwise -1. Is there any half-space $\acute{H} \in \acute{\mathrm{H}}$ that achieves 0 classification error on the transformed sample $\acute{S}$? If yes, give the equation of the maxmargin linear separator and compute the corresponding margin. (10 points)

c) What is the kernel corresponding to the feature map $\phi(.)$ in the last question, i.e., give the kernel function $K(x, z) : \mathbb{R} \times \mathbb{R} \to R$. (10 points)

Problem 2. In question 1, we explicitly constructed the feature map and find the corresponding kernel to help classify the instances using linear separator in the feature space. However, in most cases it is hard to manually construct the desired feature map, and the dimensionality of the feature space can be very high, even infinity, which makes explicit computation in the feature space infeasible in practice. In this question we will develop the dual of the primal optimization problem to avoid working in the feature space explicitly. Suppose we have a sample set $S = (x_1, y_1), \ldots, (x_n, y_n)$ of labeled examples in $\mathbb{R}^d$ with label set $\{+1, -1\}$. $Let \; \emptyset : \mathbb{R}^d \to R^D$ be a feature map that transform each input example to a feature vector in $\mathbb{R}^D$. Recall from the lecture notes that the primal optimization of SVM is given by

$$\begin{aligned} \underset{w,\varepsilon_i}{minimize} \quad & \frac{1}{2}||\boldsymbol{W}||_2^2 + C\sum_{i=1}^{n} \varepsilon_i \\ Subject\ to \quad & y_i\big(W^T \emptyset(x_i)\big) \geq 1 - \varepsilon_i \qquad \forall_i = 1, \ldots, n \\ & \varepsilon_i \geq 0, \qquad\qquad\qquad\quad \forall_i = 1, \ldots, n \end{aligned}$$

which is equivalent to the following dual optimization

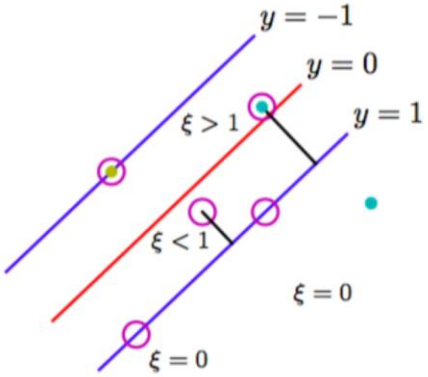
$$\underset{\alpha_i}{\text{minimize}} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \emptyset(X_i)^T \emptyset(X_j)$$

$$\text{Subject to} \quad 0 \le \alpha_i \le C \qquad \forall_i = 1, \dots, n$$

$$\sum_{i=1}^{n} \alpha_i \, y_i = 0 \qquad \forall_i = 1, \dots, n$$

Recall from the lecture notes $\varepsilon_1, \dots, \varepsilon_n$ are called slack variables. The optimal slack variables have intuitive geometric interpretation as shown in Fig. 3. Basically, when $\varepsilon_i = 0$, the corresponding feature vector $\emptyset(x_i)$ is correctly classified and it will either lie on the margin of the separator or on the correct side of the margin. Feature vector with $0 < \varepsilon_i \le 1$ lies within the margin but is still be correctly classified. When $\varepsilon_i > 1$, the corresponding feature vector is misclassified. Support vectors correspond to the instances with $\varepsilon_i > 0$ or instances that lie on the margin. The optimal vector **w** can be represented in terms of $\alpha_i, i = 1, \dots, n$ as $\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \phi(X_i)$.

a) Suppose the optimal $\varepsilon_1, \dots, \varepsilon_n$ have been computed. Use the ε_i to obtain an upper bound



on the number of misclassified instances. (10 points)

b) In the primal optimization of SVM, what's the role of the coefficient C? Briefly explain your answer by considering two extreme cases, i.e., C→ 0 and C → ∞. (10 points)

c) Explain how to use the kernel trick to avoid the explicit computation of the feature vector $\emptyset(x_i)$? Also, given a new instance x, how to make prediction on the instance without explicitly computing the feature vector $\emptyset(x)$? (10 points)

Kernels

Problem 3. In this question you will be asked to construct new kernels from existing kernels. Suppose $K_1(x,z)$: $\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and $K_2(x,z)$: $\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ are both kernels, show the following functions are also kernels: (10 points)

a) $K(x,z) = c_1 K_1(x,z) + c_2 K_2(x,z)\ with\ c_1, c_2 \geq 0.$

b) $K(x,z) = K_1(x,z).K_2(x,z)$

Generalized Lagrangian Function. (15 points)

Problem 4. Consider the optimization problem

$$\min_w f(w)$$
$$s.t. \quad g_j(w) \leq 0, \qquad \forall_j = 1, \dots, m$$
$$h_j(w) = 0, \forall_j = 1, \dots, p$$

Show that for the generalized Lagrangian function, defined by

$$L(w,\alpha,\beta) \triangleq f(w) + \sum_{j=1}^{n} \alpha_j g_j(w) + \sum_{j=1}^{P} \beta_j h_j(w)$$

the following always holds

$$\max_{\alpha \geq 0, \beta} \min_w L(w,\alpha,\beta) \leq \min_w \max_{\alpha \geq 0, \beta} L(w,\alpha,\beta)$$

Principal Components analysis (PCA) (15 points)

Problem 5. Consider 3 data points in the 2-d space: (-1, -1), (0,0), (1,1).

a) What is the first principal component (write down the actual vector)?
b) If we project the original data points into the 1-d subspace by the principal component you choose, what are their coordinates in the 1-d subspace? And what is the variance of the projected data?

c) For the projected data you just obtained above, now if we represent them in the original 2-d space and consider them as the reconstruction of the original data points, what is the reconstruction error?