1 Overview

Today's lecture will be about one of the most famous classification technique in machine learning: Support Vector Machines (SVMs). The goal is to learn how to apply the theory of Lagrangian duality to a concrete problem.

2 Classification Problems

In a classification problem, the statistician is given a sample of correctly classified data: $\{(x_1, y_1), \ldots, (x_n, y_n)\}$. For each data point i, $x_i \in \mathbb{R}^d$ is the vector of features and $y_i \in K$ is the class of the data point. $\mathbb{R}^d$ is called the feature space and K is the finite set of classes.

When $|K| = 2$, we write $K = \{-1, +1\}$. This is a two-class classification problem, and the goal of the statistician is to learn from the sample a discriminant function $f : \mathbb{R}^d \to \mathbb{R}$ such that:

$$\begin{cases} f(x_i) > 0 & \text{if} \quad y_i = +1 \\ f(x_i) < 0 & \text{if} \quad y_i = -1 \end{cases} \tag{1}$$

Such a function can be used in the following way: given a new feature vector $x \in \mathbb{R}^d$ of unknown class, compute $f(x)$. If it is positive, predict class $+1$, otherwise predict class -1.

2.1 Linear Discrimination

The problem described above is very broad since we don't require anything from the function f. It is common to restrict the function f to a *reasonably*-sized class of functions. One such restriction is affine functions.

If we require f to be an affine function, the problem described in (1) is now equivalent to finding $a \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that:

$$\begin{cases} \langle a, x_i \rangle + b > 0 & \text{if} \quad y_i = +1 \\ \langle a, x_i \rangle + b < 0 & \text{if} \quad y_i = -1 \end{cases} \tag{2}$$

Geometrically, this means that we are looking for a hyperplane $H = \{x \mid \langle a, x \rangle + b = 0\}$ such that the data points in class $+1$ lie in the positive half-space defined by this hyperplane, while the data points in class -1 lie in its negative half-space.

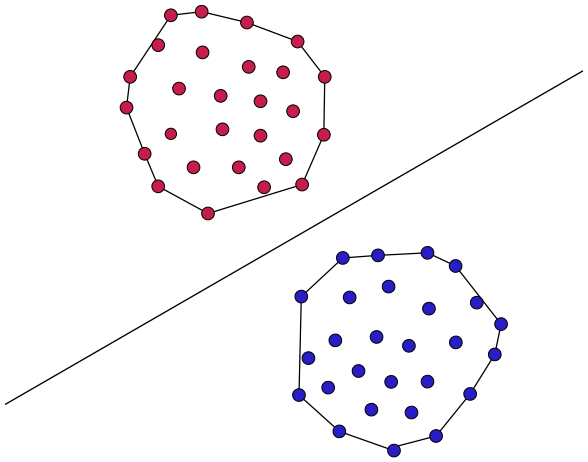

Figure 1: A linear discriminant is a hyperplane separating the feature vectors of class $+1$ (red dots) from the feature vectors in class -1 (blue dots). Proposition 2 states that it is possible to find such a hyperplane if and only if the convex hulls of the vectors of each class do not intersect.

Remark 1. It is useful to note that the set of linear inequalities (2) is equivalent to:

$$\begin{cases} \langle a, x_i \rangle + b \geq 1 & \text{if} \quad y_i = +1 \\ \langle a, x_i \rangle + b \leq -1 & \text{if} \quad y_i = -1 \end{cases} \tag{3}$$

Indeed, any (a, b) which satisfy (3) satisfies (2). Conversely, given (a, b) satisfying (2), we see that if we multiply the inequalities in (2) by some positive constant c large enough, then $(a', b') := (ca, cb)$ will satisfy (3).

3 Support Vector Machines

3.1 Separability

The first question one might ask is whether or not the set of inequalities in (3) is feasible, *i.e.* whether or not there exists a hyperplane separating the points in class -1 from those in class $+1$.

Let us define $I = \{i \mid y_i = +1\}$ the set of data points in class $+1$ and similarly $J = \{j \mid y_j = -1\}$ the data points in class -1. The linear inequalities in (3) can be rewritten in matrix form as $XA \geq U$, where X is the $n \times (d+1)$ matrix whose i-th row is $[x_i^T \ 1]$ if $i \in I$ and $[-x_i^T \ -1]$ if $i \in J$, A is the concatenation of a and b, and U is the vector whose components are all one.

Then, by Farkas' lemma, if (3) is not feasible, *i.e.* if there does not exist A such that $XA \geq U$, then there exists Y such that $Y^T X = 0$, $Y^T U > 0$ and $Y \geq 0$. Let us write $Y_i = \lambda_i$ when $i \in I$ and $Y_j = \mu_j$ when $j \in J$ (note that (I, J) is a partition of $\{1, \ldots, n\}$). Then $Y \geq 0$ implies that $\lambda \geq 0$

and $\mu \geq 0$. And the conditions $Y^T U = 0$ and $Y^T U > 0$ imply:

$$\sum_{i \in I} \lambda_i x_i = \sum_{j \in J} \mu_j x_j \tag{4}$$

$$\sum_{i \in I} \lambda_i = \sum_{j \in J} \mu_j \tag{5}$$

$$\sum_{i \in I} \lambda_i + \sum_{j \in J} \mu_j > 0 \tag{6}$$

Equation (6) as well as $\lambda \geq 0$ and $\mu \geq 0$ imply that $\lambda \neq 0$ and $\mu \neq 0$. Hence we can define:

$$\lambda'_i = \frac{\lambda_i}{\sum_{i \in I} \lambda_i} \quad \text{and} \quad \mu'_j = \frac{\mu_j}{\sum_{j \in J} \mu_j}$$

Finally, dividing (4) both sides by $\sum_{i \in I} \lambda_i$ and using (5), we obtain:

$$\sum_{i \in I} \lambda'_i x_i = \sum_{j \in J} \mu'_j x_j \tag{7}$$

Note that $\sum_{i \in I} \lambda'_i = \sum_{j \in J} \mu'_j = 1$ and recall that the convex hull C of k vectors $z_1, \ldots, z_k$ of $\mathbb{R}^d$, the smallest convex set containing the k vectors can be written:

$$C = \left\{ \sum_{i=1}^{k} \lambda_i x_i \mid \lambda \geq 0 \text{ and } \sum_{i=1}^{k} \lambda_i = 1 \right\}$$

As a consequence, the existence of λ' and μ' satisfying (7) is exactly equivalent to saying that the convex hulls of $\{x_i, i \in I\}$ and $\{x_j, j \in J\}$ intersect. We have now proved:

Proposition 2. *The set of linear inequalities in (3) is feasible if and only if the convex hulls of $\{x_i, i \in I\}$ and $\{x_j, j \in J\}$ do not intersect.*

3.2 Robust Separation

From now on, we will assume that the condition of Proposition 2 is satisfied, so that we can find a hyperplane separating our data points. We note that the existence of one such hyperplane in fact implies the existence of infinitely many such hyperplanes and we need to decide on a criterion to select one which will be a "good" discriminant for our classification task.

Note that a separating hyperplane satisfying (3) defines a region of the space which contains no data points: $M = \{x \mid -1 \leq \langle a, x \rangle + b \leq 1\}$. This region is contained between the two parallel hyperplanes $\{x \mid \langle a, x \rangle + b = -1\}$ and $\{x \mid \langle a, x \rangle + b = +1\}$ and is called the classification margin. We define its width to be the distance between the two boundary hyperplanes. You showed in your second assignment that this distance is equal to $\frac{2}{\|a\|}$.

This provides a criterion to select a "good" separating hyperplane: select a hyperplane which separates the data points the most, *i.e.* for which the width of the margin is maximal. Such a hyperplane is called a *maximum margin hyperplane*. Formally the optimization problem can be written:

$$\max_{a,b} \frac{2}{\|a\|} \tag{8}$$

$$\text{s.t. } \langle a, x_i \rangle + b \geq 1, \ i \in I \tag{9}$$

$$\langle a, x_j \rangle + b \leq -1, \ j \in J \tag{10}$$

Equivalently we could minimize the quantity $\frac{\|a\|}{2}$ or its squared $\frac{\|a\|^2}{4}$. This leads to the following convex optimization problem in standard form:

$$\min_{a,b} \frac{\|a\|^2}{4} \tag{11}$$

$$\text{s.t. } \langle a, x_i \rangle + b \geq 1, \ i \in I \tag{12}$$

$$\langle a, x_j \rangle + b \leq -1, \ j \in J \tag{13}$$

3.3 Computing the Dual

We will now see how to compute the dual of Problem (11). We start by forming the Lagrangian of the problem:

$$L(a, b, \lambda, \mu) = \frac{\|a\|^2}{4} + \sum_{i \in I} \lambda_i \big(1 - \langle a, x_i \rangle - b\big) + \sum_{j \in J} \mu_j \big(\langle a, x_j \rangle + b + 1\big)$$

where λ and μ are the dual variables and are positive vectors of $\mathbb{R}^{|I|}$ and $\mathbb{R}^{|J|}$ respectively. By reordering the terms we get:

$$L(a, b, \lambda, \mu) = \frac{\|a\|^2}{4} + \left\langle a, \sum_{j \in J} \mu_j x_j - \sum_{i \in I} \lambda_i x_i \right\rangle + \left\langle b, \sum_{j \in J} \mu_j - \sum_{i \in I} \lambda_i \right\rangle + \sum_{i \in I} \lambda_i + \sum_{j \in J} \mu_j$$

We can now compute the dual function $F(\lambda, \mu) = \min_{a,b} L(a, b, \lambda, \mu)$. First, taking the derivate of L with respect to a and b, we see that the critical points of L as a function of its first two variables verify:

$$\frac{a}{2} + \sum_{j \in J} \mu_j x_j - \sum_{i \in I} \lambda_i x_i = 0 \tag{14}$$

$$\sum_{j \in J} \mu_j - \sum_{i \in I} \lambda_i = 0 \tag{15}$$

As a consequence, if (15) is satisfied, then a critical point exists, and since $L(a, b, \lambda, \mu)$ is convex in (a, b) the critical point is a global minimum. The value of a at the critical point is given by (14) and plugging it in the definition of the Lagrangian, we obtain:

$$F(\lambda, \mu) = -\left\| \sum_{i \in I} \lambda_i x_i - \sum_{j \in J} \mu_j x_j \right\|^2 + \sum_{i \in I} \lambda_i + \sum_{j \in J} \mu_j$$

If (15) is not satisfied, then it is easy to see that the minimum of $F(\lambda, \mu) = -\infty$. Indeed, in this case one can simply let $a = 0$ and observe that $\left\langle b, \sum_{j\in J} \mu_j - \sum_{i\in I} \lambda_i \right\rangle$ converges to $-\infty$ as b converges to ∞ or $-\infty$.

The dual of Problem (11) is simply maximizing the dual function F over $\lambda \geq 0$ and $\mu \geq 0$. It is sufficient to restrict the domain to λ and μ satisfying (15) (otherwise the dual function is equal to $-\infty$ and clearly not maximum). Hence, the dual problem can be written:

$$\max_{\lambda,\mu} \; -\left\| \sum_{i\in I} \lambda_i x_i - \sum_{j\in J} \mu_j x_j \right\|^2 + \sum_{i\in I} \lambda_i + \sum_{j\in J} \mu_j \tag{16}$$

$$\text{s.t.} \sum_{i\in I} \lambda_i = \sum_{j\in J} \mu_j \tag{17}$$

$$\lambda \geq 0, \; \mu \geq 0 \tag{18}$$

It is possible to further simplify this problem by parametrizing it in terms of $s = \sum_{i\in I} \lambda_i = \sum_{j\in J} \mu_j$, $\lambda' = \frac{\lambda}{s}$ and $\mu' = \frac{\mu}{s}$:

$$\max_{\lambda',\mu',s} \; -s^2 \left\| \sum_{i\in I} \lambda'_i x_i - \sum_{j\in J} \mu'_j x_j \right\|^2 + 2s \tag{19}$$

$$\text{s.t.} \sum_{i\in I} \lambda'_i = \sum_{j\in J} \mu'_j = 1 \tag{20}$$

$$\lambda' \geq 0, \; \mu' \geq 0, \; s \geq 0 \tag{21}$$

The maximum in s is clearly attained when:

$$s = \frac{1}{\left\| \sum_{i\in I} \lambda'_i x_i - \sum_{j\in J} \mu'_j x_j \right\|^2}$$

and one can finally write the dual problem of (11):

$$\max_{\lambda',\mu'} \; \frac{1}{\left\| \sum_{i\in I} \lambda'_i x_i - \sum_{j\in J} \mu'_j x_j \right\|^2} \tag{22}$$

$$\text{s.t.} \sum_{i\in I} \lambda'_i = \sum_{j\in J} \mu'_j = 1 \tag{23}$$

$$\lambda' \geq 0, \; \mu' \geq 0 \tag{24}$$

3.4 Strong Duality

It is not hard to see that whenever the condition of Proposition 2 holds, then the set of inequalities in (3) are also strictly feasible: that is, they are still feasible by replacing the inequalities by strict inequalities. In other words, Slater's condition holds and we have strong duality: the optimum value of (22) is equal to the optimum value of (11). By taking the inverse square root both sides, we can rewrite this in terms of the original problem (8):

$$\begin{aligned} \max_{a,b} \; & \frac{2}{\|a\|} \\ \text{s.t.} \; & \langle a, x_i \rangle + b \geq 1, \; i \in I \\ & \langle a, x_j \rangle + b \leq -1, \; j \in J \end{aligned} \quad = \quad \begin{aligned} \min_{\lambda',\mu'} \; & \left\| \sum_{i\in I} \lambda'_i x_i - \sum_{j\in J} \mu'_j x_j \right\| \\ \text{s.t.} \; & \sum_{i\in I} \lambda'_i = \sum_{j\in J} \mu'_j = 1 \\ & \lambda' \geq 0, \; \mu' \geq 0 \end{aligned} \tag{25}$$

The geometric interpretation of equation (25) is the following: the problem on the right-hand side consists in finding the minimum distance between one point in the convex hull of $(x_i)_{i \in I}$ and one point in the convex hull of $(x_j)_{j \in J}$. Strong duality implies that this minimum distance is equal to the margin of a maximum-margin hyperplane.

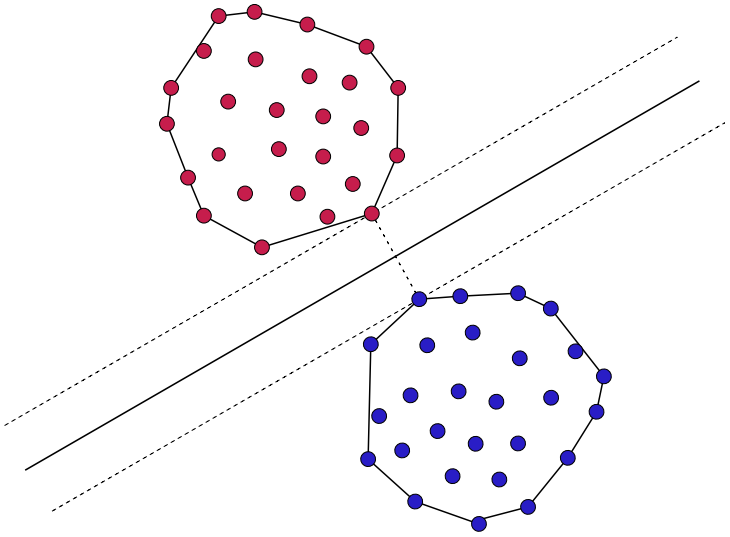

Figure 2: The maximum margin hyperplane (solid) is orthogonal to the vector joining the two closest vectors in the convex hulls of the red dots and the blue dots. These two vectors lie on the hyperplanes at the boundary of the margin (dashed lines).