LECTURE 7
Support vector machines

SVMs have been used in a multitude of applications and are one of the most popular machine learning algorithms. We will derive the SVM algorithm from two perspectives: Tikhonov regularization, and the more common geometric perspective. We will focus on the linear SVM.

7.1. SVMs from Tikhonov regularization

We start with Tikhonov regularization

$$\min_{\beta \in \mathbb{R}^p} \left[n^{-1} \sum_{i=1}^{n} V(y_i, \beta^T x_i) + \lambda \|\beta\|^2 \right]$$

and use the hinge loss functional

$$n^{-1} \sum_{i=1}^{n} V(f, z_i) := n^{-1} \sum_{i=1}^{n} (1 - y_i \beta^T x_i)_+,$$

where $(k)_+ := \max(k, 0)$.

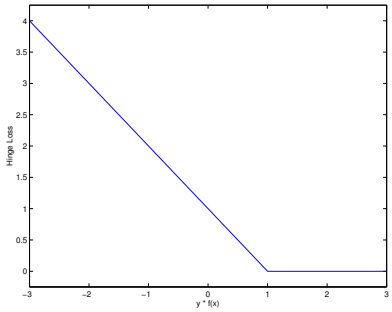

Figure 1. Hinge loss.

The resulting optimization problem is

$$\min_{\beta \in \mathbb{R}^p} \left[n^{-1} \sum_{i=1}^{n} (1 - y_i \beta^T x_i)_+ + \lambda \|\beta\|^2 \right], \tag{7.1}$$

which is non-differentiable at $(1 - y_i f(x_i)) = 0$ so we introduce slack variables and write the following constrained optimization problem:

$$\begin{aligned} \min_{\beta \in \mathbb{R}^p} \quad & n^{-1} \textstyle\sum_{i=1}^n \xi_i + \lambda ||\beta||^2 \\ \text{subject to :} \quad & y_i \beta^T x_i \geq 1 - \xi_i \qquad i = 1, \ldots, n \\ & \xi_i \geq 0 \qquad\qquad i = 1, \ldots, n. \end{aligned}$$

The SVM contains an unregularized bias term b so the separating hyperplane need not go through the origin. Plugging this form into the above constrained quadratic problem results in the "primal" SVM

$$\begin{aligned} \min_{\beta \in \mathbb{R}^p, \xi \in \mathbb{R}^n, b \in \mathbb{R}} \quad & n^{-1} \textstyle\sum_{i=1}^n \xi_i + \lambda \|\beta\|^2 \\ \text{subject to :} \quad & y_i \left(\beta^T x_i + b \right) \geq 1 - \xi_i \quad i = 1, \ldots, n \\ & \xi_i \geq 0 \qquad\qquad i = 1, \ldots, n. \end{aligned}$$

Note the following trick, one can reparameterize β as

$$\beta = \sum_{j=1}^n c_j x_j,$$

this is an advantageous representation since one now only needs n variables to parameterize β rather than p. So we now rewrite the optimization problem as

$$\begin{aligned} \min_{c \in \mathbb{R}^n, \xi \in \mathbb{R}^n, b \in \mathbb{R}} \quad & n^{-1} \textstyle\sum_{i=1}^n \xi_i + \lambda \sum_{ij} c_i c_j x_i^T x_j \\ \text{subject to :} \quad & y_i \left(\sum_j c_j x_j^T x_i + b \right) \geq 1 - \xi_i \quad i = 1, \ldots, n \\ & \xi_i \geq 0 \qquad\qquad i = 1, \ldots, n. \end{aligned}$$

We now derive the Wolfe dual quadratic program using Lagrange multiplier techniques:

$$\begin{aligned} L(c, \xi, b, \alpha, \zeta) \quad = \quad & n^{-1} \sum_{i=1}^n \xi_i + \lambda \sum_{ij} c_i c_j x_i^T x_j \\ & - \sum_{i=1}^n \alpha_i \left(y_i \left\{ \sum_j c_j x_j^T x_i + b \right\} - 1 + \xi_i \right) \\ & - \sum_{i=1}^n \zeta_i \xi_i. \end{aligned}$$

We want to minimize L with respect to $\mathbf{c}$, b, and ξ, and maximize L with respect to α and ζ, subject to the constraints of the primal problem and nonnegativity constraints on α and ζ. We first eliminate b and ξ by taking partial derivatives:

$$\begin{aligned} \frac{\partial L}{\partial b} = 0 \quad & \Longrightarrow \quad \sum_{i=1}^n \alpha_i y_i = 0 \\ \frac{\partial L}{\partial \xi_i} = 0 \quad & \Longrightarrow \quad \frac{1}{n} - \alpha_i - \zeta_i = 0 \Longrightarrow 0 \leq \alpha_i \leq \frac{1}{n}. \end{aligned}$$

The above two conditions will be constraints that will have to be satisfied at optimality. This results in a reduced Lagrangian:

$$L^R(c,\alpha) = \lambda \sum_{ij} c_i c_j x_i^T x_j - \sum_{i=1}^n \alpha_i \left(y_i \sum_j c_j x_j^T x_i - 1 \right).$$

We now eliminate c

$$\frac{\partial L^R}{\partial c} = 0 \quad \Longrightarrow c_i = \tfrac{\alpha_i y_i}{2\lambda},$$

Substituting the above expression for $\mathbf{c}$ into the reduced Lagrangian we are left with the following "dual" program:

$$\begin{aligned} \max_{\alpha \in \mathbb{R}^n} \quad & \textstyle\sum_{i=1}^n \alpha_i - \frac{1}{4\lambda} \alpha^T Q \alpha \\ \text{subject to :} \quad & \textstyle\sum_{i=1}^n y_i \alpha_i = 0 \\ & 0 \le \alpha_i \le \tfrac{1}{n} \qquad i = 1, \ldots, n, \end{aligned}$$

where Q is the matrix defined by

$$Q_{ij} = y_i y_j x_i^T x_j.$$

In most of the SVM literature, instead of the regularization parameter λ, regularization is controlled via a parameter C, defined using the relationship

$$C = \frac{1}{2\lambda n}.$$

Like λ, the parameter C also controls the trade-off between classification accuracy and the norm of the function. The primal and dual problems become respectively:

$$\begin{aligned} \min_{\mathbf{c} \in \mathbb{R}^n, \xi \in \mathbb{R}^n} \quad & C \textstyle\sum_{i=1}^n \xi_i + \frac{1}{2} \sum_{ij} c_i c_j x_i^T x_j \\ \text{subject to :} \quad & y_i \left(\textstyle\sum_{j=1}^n c_j x_i^T x_j + b \right) \ge 1 - \xi_i \quad i = 1, \ldots, n \\ & \xi_i \ge 0 \qquad\qquad\qquad\qquad i = 1, \ldots, n \end{aligned}$$

$$\begin{aligned} \max_{\alpha \in \mathbb{R}^n} \quad & \textstyle\sum_{i=1}^n \alpha_i - \frac{1}{2} \alpha^T Q \alpha \\ \text{subject to :} \quad & \textstyle\sum_{i=1}^n y_i \alpha_i = 0 \\ & 0 \le \alpha_i \le C \qquad i = 1, \ldots, n. \end{aligned}$$

7.2. SVMs from a geometric perspective

The "traditional" approach to developing the mathematics of SVM is to start with the concepts of separating hyperplanes and margin. The theory is usually developed in a linear space, beginning with the idea of a perceptron, a linear hyperplane that separates the positive and the negative examples. Defining the margin as the distance from the hyperplane to the nearest example, the basic observation is that intuitively, we expect a hyperplane with larger margin to generalize better than one with smaller margin.

We denote our hyperplane by $\mathbf{w}$, and we will classify a new point $\mathbf{x}$ via the function

$$f(x) = \operatorname{sign} \left[\langle \mathbf{w}, \mathbf{x} \rangle \right]. \tag{7.2}$$

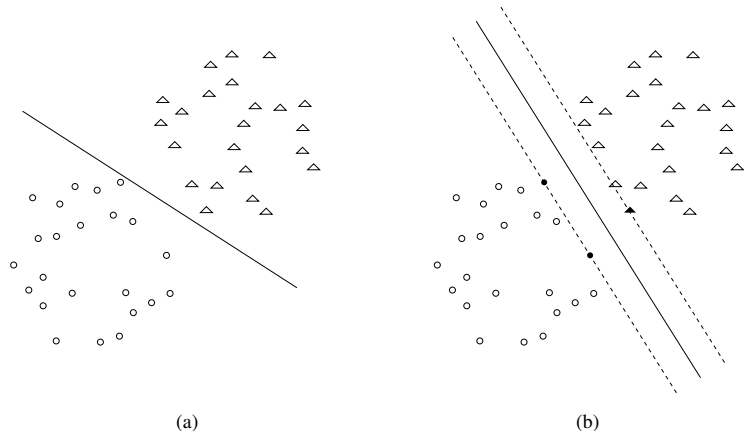

Figure 2. Two hyperplanes (a) and (b) perfectly separate the data. However, hyperplane (b) has a larger margin and intuitively would be expected to be more accurate on new observations.

Given a separating hyperplane $\mathbf{w}$ we let $\mathbf{x}$ be a datapoint closest to $\mathbf{w}$, and we let $\mathbf{x}^{\mathbf{w}}$ be the unique point on $\mathbf{w}$ that is closest to x. Obviously, finding a maximum margin $\mathbf{w}$ is equivalent to maximizing $||\mathbf{x} - \mathbf{x}^{\mathbf{w}}||$. So for some k (assume $k > 0$ for convenience),

$$\begin{aligned} \langle \mathbf{w}, \mathbf{x} \rangle &= k \\ \langle \mathbf{w}, \mathbf{x}^{\mathbf{w}} \rangle &= 0 \\ \langle \mathbf{w}, (\mathbf{x} - \mathbf{x}^{\mathbf{w}}) \rangle &= k. \end{aligned}$$

Noting that the vector $\mathbf{x} - \mathbf{x}^{\mathbf{w}}$ is parallel to the normal vector w,

$$\begin{aligned} \langle \mathbf{w}, (\mathbf{x} - \mathbf{x}^{\mathbf{w}}) \rangle &= \left\langle \mathbf{w}, \left(\frac{||\mathbf{x} - \mathbf{x}^{\mathbf{w}}||}{||\mathbf{w}||} \mathbf{w} \right) \right\rangle \\ &= ||\mathbf{w}||^2 \frac{||\mathbf{x} - \mathbf{x}^{\mathbf{w}}||}{||\mathbf{w}||} \\ &= ||\mathbf{w}||\ ||\mathbf{x} - \mathbf{x}^{\mathbf{w}}|| \\ k &= ||\mathbf{w}||\ ||(\mathbf{x} - \mathbf{x}^{\mathbf{w}})|| \\ \frac{k}{||\mathbf{w}||} &= ||\mathbf{x} - \mathbf{x}^{\mathbf{w}}||. \end{aligned}$$

k is a "nuisance parameter" and without any loss of generality, we fix k to 1, and see that maximizing $||\mathbf{x} - \mathbf{x}^{\mathbf{w}}||$ is equivalent to maximizing $\frac{1}{||w||}$, which in turn is equivalent to minimizing $||w||$ or $||w||^2$. We can now define the margin as the distance between the hyperplanes $\langle \mathbf{w}, \mathbf{x} \rangle = 0$ and $\langle \mathbf{w}, \mathbf{x} \rangle = 1$.

So if the data is linear separable case and the hyperplanes run through the origin the maximum margin hyperplane is the one for which

$$\begin{aligned} \min_{\mathbf{w} \in \mathbb{R}^n} \quad & ||\mathbf{w}||^2 \\ \text{subject to :} \quad & y_i \langle \mathbf{w}, \mathbf{x}_i \rangle \geq 1 \quad i = 1, \ldots, n. \end{aligned}$$

The SVM introduced by Vapnik includes an unregularized bias term b, leading to classification via a function of the form:

$$f(x) = \text{sign}[\langle \mathbf{w}, \mathbf{x} \rangle + b].$$

In addition, we need to work with datasets that are not linearly separable, so we introduce slack variables ξ_i, just as before. We can still define the margin as the distance between the hyperplanes $\langle \mathbf{w}, \mathbf{x} \rangle = 0$ and $\langle \mathbf{w}, \mathbf{x} \rangle = 1$, but the geometric intuition is no longer as clear or compelling.

With the bias term and slack variables the primal SVM problem becomes

$$\begin{array}{ll} \min\limits_{\mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R}} & C \sum_{i=1}^n \xi_i + \frac{1}{2}||\mathbf{w}||^2 \\ \text{subject to :} & y_i \left(\langle \mathbf{w}, \mathbf{x} \rangle + b \right) \geq 1 - \xi_i \quad i = 1, \ldots, n \\ & \xi_i \geq 0 \qquad\qquad\quad i = 1, \ldots, n. \end{array}$$

Using Lagrange multipliers we can derive the same dual from in the previous section.

Historically, most developments begin with the geometric form, derived a dual program which was identical to the dual we derived above, and only then observed that the dual program required only dot products and that these dot products could be replaced with a kernel function. In the linearly separable case, we can also derive the separating hyperplane as a vector parallel to the vector connecting the closest two points in the positive and negative classes, passing through the perpendicular bisector of this vector. This was the "Method of Portraits", derived by Vapnik in the 1970's, and recently rediscovered (with non-separable extensions) by Keerthi.

7.3. Optimality conditions

The primal and the dual are both feasible convex quadratic programs. Therefore, they both have optimal solutions, and optimal solutions to the primal and the dual have the same objective value.

We derived the dual from the primal using the (now reparameterized) Lagrangian:

$$\begin{aligned} L(\mathbf{c}, \xi, b, \alpha, \zeta) \quad = \quad & C \sum_{i=1}^n \xi_i + \sum_{ij} c_i c_j x_i^T x_j \\ & - \sum_{i=1}^n \alpha_i \left(y_i \left\{ \sum_{j=1}^n c_j x_i^T x_j + b \right\} - 1 + \xi_i \right) \\ & - \sum_{i=1}^n \zeta_i \xi_i. \end{aligned}$$

We now consider the dual variables associated with the primal constraints:

$$\begin{aligned} \alpha_i \quad &\Longrightarrow \quad y_i \left\{ \sum_{j=1}^n c_j x_i^T x_j + b \right\} - 1 + \xi_i \\ \zeta_i \quad &\Longrightarrow \quad \xi_i \geq 0. \end{aligned}$$

Complementary slackness tells us that at optimality, either the primal inequality is satisfied at equality or the dual variable is zero. In other words, if $\mathbf{c}$, ξ, b, α and ζ

are optimal solutions to the primal and dual, then

$$\begin{aligned} \alpha_i \left(y_i \left\{ \sum_{j=1}^{n} c_j x_i^T x_j + b \right\} - 1 + \xi_i \right) &= 0 \\ \zeta_i \xi_i &= 0 \end{aligned}$$

All optimal solutions must satisfy:

$$\begin{aligned} \sum_{j=1}^{n} c_j x_i^T x_j - \sum_{j=1}^{n} y_i \alpha_j x_i^T x_j = 0 \qquad & i = 1, \ldots, n \\ \sum_{i=1}^{n} \alpha_i y_i = 0 & \\ C - \alpha_i - \zeta_i = 0 \qquad & i = 1, \ldots, n \\ y_i \left(\sum_{j=1}^{n} y_j \alpha_j x_i^T x_j + b \right) - 1 + \xi_i \geq 0 \qquad & i = 1, \ldots, n \\ \alpha_i \left[y_i \left(\sum_{j=1}^{n} y_j \alpha_j x_i^T x_j + b \right) - 1 + \xi_i \right] = 0 \qquad & i = 1, \ldots, n \\ \zeta_i \xi_i = 0 \qquad & i = 1, \ldots, n \\ \xi_i, \alpha_i, \zeta_i \geq 0 \qquad & i = 1, \ldots, n \end{aligned}$$

The above optimality conditions are both necessary and sufficient. If we have $\mathbf{c}$, ξ, b, α and ζ satisfying the above conditions, we know that they represent optimal solutions to the primal and dual problems. These optimality conditions are also known as the Karush-Kuhn-Tucker (KKT) conditions.

Suppose we have the optimal α_i's. Also suppose (this "always" happens in practice") that there exists an i satisfying $0 < \alpha_i < C$. Then

$$\begin{aligned} \alpha_i < C &\implies \zeta_i > 0 \\ &\implies \xi_i = 0 \\ &\implies y_i \left(\sum_{j=1}^{n} y_j \alpha_j x_i^T x_j + b \right) - 1 = 0 \\ &\implies b = y_i - \sum_{j=1}^{n} y_j \alpha_j x_i^T x_j \end{aligned}$$

So if we know the optimal α's, we can determine b.

Defining our classification function $f(x)$ as

$$f(x) = \sum_{i=1}^{\ell} y_i \alpha_i x^T x_i + b,$$

we can derive "reduced" optimality conditions. For example, consider an i such that $y_i f(x_i) < 1$:

$$\begin{aligned} y_i f(x_i) < 1 &\implies \xi_i > 0 \\ &\implies \zeta_i = 0 \\ &\implies \alpha_i = C. \end{aligned}$$

Conversely, suppose $\alpha_i = C$:

$$\begin{aligned} \alpha_i = C &\implies y_i f(x_i) - 1 + \xi_i = 0 \\ &\implies y_i f(x_i) \leq 1. \end{aligned}$$

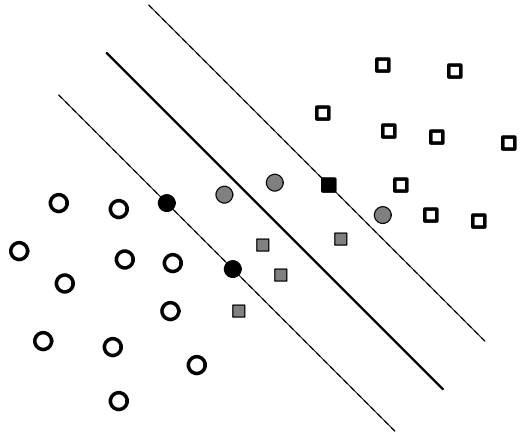

Figure 3. A geometric interpretation of the reduced optimality conditions. The open squares and circles correspond to cases where $\alpha_i = 0$. The dark circles and squares correspond to cases where $y_i f(x_i) = 1$ and $\alpha_i \leq C$, these are samples at the margin. The grey circles and squares correspond to cases where $y_i f(x_i) < 1$ and $\alpha_i = C$.

7.4. Solving the SVM optimization problem

Our plan will be to solve the dual problem to find the α's, and use that to find b and our function f. The dual problem is easier to solve the primal problem. It has simple box constraints and a single inequality constraint, even better, we will see that the problem can be decomposed into a sequence of smaller problems.

We can solve QPs using standard software. Many codes are available. Main problem — the Q matrix is dense, and is $n \times n$, so we cannot write it down. Standard QP software requires the Q matrix, so is not suitable for large problems.

To get around this memory issue we partition the dataset into a working set W and the remaining points R. We can rewrite the dual problem as:

$$\begin{aligned} \max_{\alpha_W \in \mathbb{R}^{|W|},\ \alpha_R \in \mathbb{R}^{|R|}} \quad & \textstyle\sum_{\substack{i=1 \\ i \in W}}^{n} \alpha_i + \sum_{\substack{i=1 \\ i \in R}} \alpha_i \\ & -\tfrac{1}{2} [\alpha_{\mathbf{W}}\ \alpha_{\mathbf{R}}] \begin{bmatrix} Q_{WW} & Q_{WR} \\ Q_{RW} & Q_{RR} \end{bmatrix} \begin{bmatrix} \alpha_{\mathbf{W}} \\ \alpha_{\mathbf{R}} \end{bmatrix} \\ \text{subject to :} \quad & \textstyle\sum_{i \in W} y_i \alpha_i + \sum_{i \in R} y_i \alpha_i = 0 \\ & 0 \leq \alpha_i \leq C,\ \forall i. \end{aligned}$$

Suppose we have a feasible solution α. We can get a better solution by treating the $\alpha_{\mathbf{W}}$ as variable and the $\alpha_{\mathbf{R}}$ as constant. We can solve the reduced dual problem:

$$\begin{aligned} \max_{\alpha_W \in \mathbb{R}^{|W|}} \quad & (\mathbf{1} - Q_{WR}\alpha_{\mathbf{R}})\alpha_W - \tfrac{1}{2}\alpha_{\mathbf{W}} Q_{WW}\alpha_{\mathbf{W}} \\ \text{subject to:} \quad & \textstyle\sum_{i\in W} y_i\alpha_i = -\sum_{i\in R} y_i\alpha_i \\ & 0 \le \alpha_i \le C, \ \forall i \in W. \end{aligned}$$

The reduced problems are fixed size, and can be solved using a standard QP code. Convergence proofs are difficult, but this approach seems to always converge to an optimal solution in practice.

An important issue in the decomposition is selecting the working set. There are many different approaches. The basic idea is to examine points not in the working set, find points which violate the reduced optimality conditions, and add them to the working set. Remove points which are in the working set but are far from violating the optimality conditions.