Support Vector Machines For Classification

9.520 Class 05, 18 February 2004

Ryan Rifkin

Plan

- Regularization derivation of SVMs
- Geometric derivation of SVMs
- Optimality, Duality and Large Scale SVMs
- SVMs and RLSC: Compare and Contrast

The Regularization Setting (Again)

We are given ℓ examples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell)$, with $\mathbf{x}_i \in \mathbb{R}^n$ and $y_i \in \{-1, 1\}$ for all i. As mentioned last class, we can find a classification function by solving a regularized learning problem:

$$\min_{f \in \mathcal{H}} \frac{1}{\ell} \sum_{i=1}^{\ell} V(y_i, f(\mathbf{x}_i)) + \lambda ||f||_K^2.$$

Note that in this class we are specifically consider **binary classification**.

The Hinge Loss

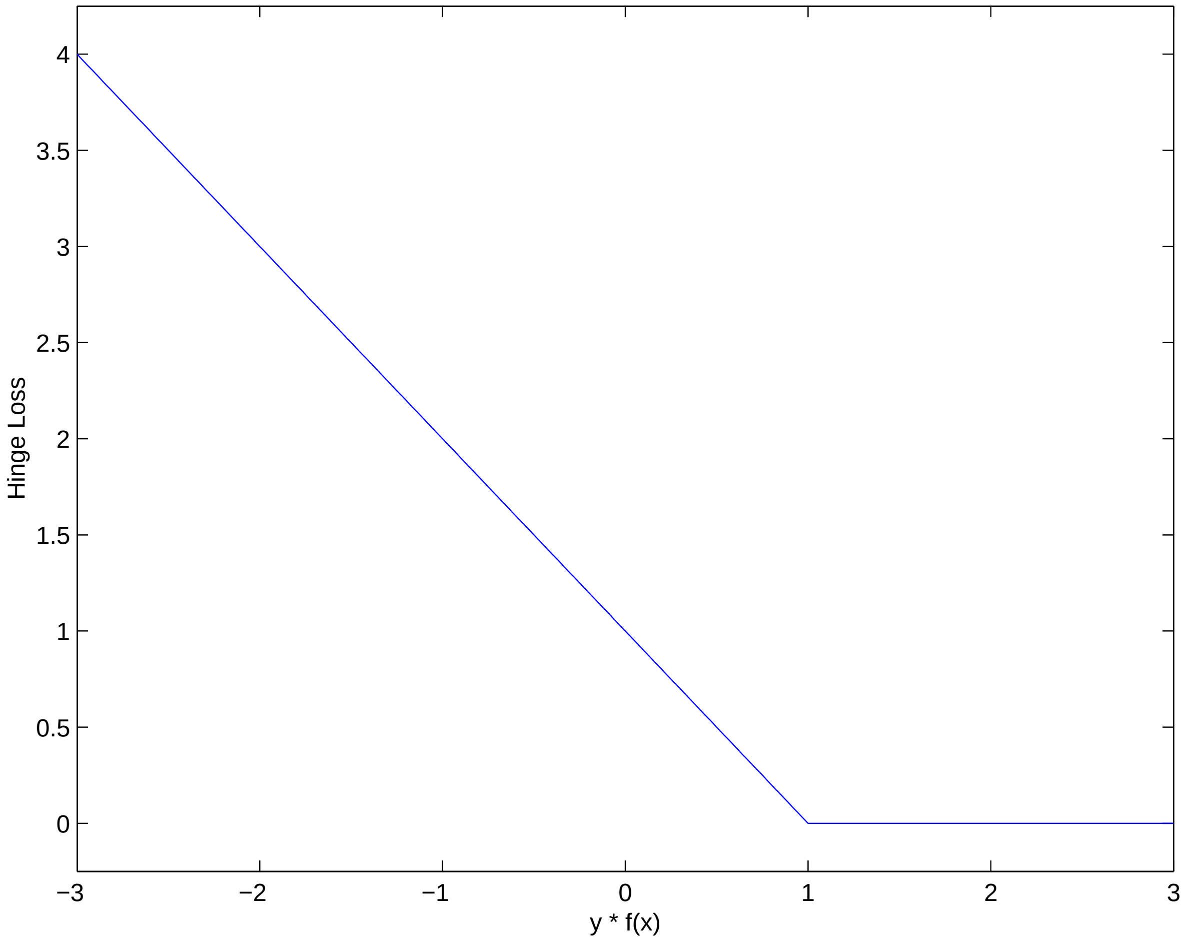
The classical SVM arises by considering the specific loss function

$$V(f(\mathbf{x}), y) \equiv (1 - yf(\mathbf{x}))_+,$$

where

$$(k)_+ \equiv \mathrm{max}(k, 0).$$

The Hinge Loss



Substituting In The Hinge Loss

With the hinge loss, our regularization problem becomes

$$\min_{f \in \mathcal{H}} \frac{1}{\ell} \sum_{i=1}^{\ell} (1 - y_i f(\mathbf{x}_i))_+ + \lambda ||f||_K^2.$$

Slack Variables

This problem is non-differentiable (because of the “kink” in V), so we introduce slack variables ξ_i, to make the problem easier to work with:

$$\begin{array}{rll} \min\limits_{f \in \mathcal{H}} & \frac{1}{\ell}\sum_{i=1}^{\ell} \xi_i + \lambda ||f||_K^2 & \\ \text{subject to :} & y_i f(\mathbf{x}_i) \geq 1 - \xi_i & i = 1, \ldots, \ell \\ & \xi_i \geq 0 & i = 1, \ldots, \ell \end{array}$$

Applying The Representer Theorem

Substituting in:

$$f^*(\mathbf{x}) = \sum_{i=1}^{\ell} c_i K(\mathbf{x}, \mathbf{x}_i),$$

we arrive at a constrained quadratic programming problem:

$$\begin{array}{lll} \underset{\mathbf{c} \in \mathbb{R}^\ell}{\text{min}} & \frac{1}{\ell}\sum_{i=1}^{\ell} \xi_i + \lambda \mathbf{c}^T K \mathbf{c} & \\ \text{subject to}: & y_i \sum_{j=1}^{\ell} c_j K(x_i, x_j) \geq 1 - \xi_i & i = 1, \ldots, \ell \\ & \xi_i \geq 0 & i = 1, \ldots, \ell \end{array}$$

Adding A Bias Term

If we add an unregularized bias term b, which presents some theoretical difficulties to be discussed later, we arrive at the "primal" SVM:

$$\begin{array}{lll}
\min\limits_{\mathbf{c}\in\mathbb{R}^\ell,\xi\in\mathbb{R}^\ell} & \frac{1}{\ell}\sum_{i=1}^{\ell}\xi_i + \lambda\mathbf{c}^T K\mathbf{c} & \\
\text{subject to :} & y_i(\sum_{j=1}^{\ell} c_j K(x_i,x_j) + b) \geq 1 - \xi_i & i = 1,\dots,\ell \\
 & \xi_i \geq 0 & i = 1,\dots,\ell
\end{array}$$

Forming the Lagrangian

For reasons that will be clear in a few slides, we derive the Wolfe dual quadratic program using Lagrange multiplier techniques:

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

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.

Eliminating b and ξ

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

We write a reduced Lagrangian in terms of the remaining variables:

$$L^R(\mathbf{c}, \alpha) = \lambda \mathbf{c}^T K \mathbf{c} - \sum_{i=1}^{\ell} \alpha_i (y_i \sum_{j=1}^{\ell} c_j K(x_i, x_j) - 1)$$

Eliminating $\mathbf{c}$

Assuming the K matrix is invertible,

$$\begin{aligned}\frac{\partial L^R}{\partial \mathbf{c}} = 0 &\implies 2\lambda K\mathbf{c} - KY\alpha = 0 \\ &\implies c_i = \frac{\alpha_i y_i}{2\lambda}\end{aligned}$$

Where Y is a diagonal matrix whose i'th diagonal element is y_i; $Y\alpha$ is a vector whose i'th element is $\alpha_i y_i$.

The Dual Program

Substituting in our expression for $\mathbf{c}$, we are left with the following "dual" program:

$$\begin{array}{rl} \max\limits_{\alpha \in \mathbb{R}^{\ell}} & \sum_{i=1}^{\ell} \alpha_i - \frac{1}{4\lambda} \alpha^T Q \alpha \\ \text{subject to}: & \sum_{i=1}^{\ell} y_i \alpha_i = 0 \\ & 0 \le \alpha_i \le \frac{1}{\ell} \qquad i = 1, \ldots, \ell \end{array}$$

Here, Q is the matrix defined by

$$Q = \mathbf{y} K \mathbf{y}^{\mathbf{T}} \Longleftrightarrow Q_{ij} = y_i y_j K(x_i, x_j).$$

Standard Notation

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

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

Using this definition (after multiplying our objective function by the constant$\frac{1}{2\lambda}$, the basic regularization problem becomes

$$\min_{f \in \mathcal{H}} \ C \sum_{i=1}^{\ell} V(y_i, f(\mathrm{x}_i)) + \frac{1}{2}||f||_K^2.$$

Like λ, the parameter C also controls the tradeoff between classification accuracy and the norm of the function. The primal and dual problems become...

The Reparametrized Problems

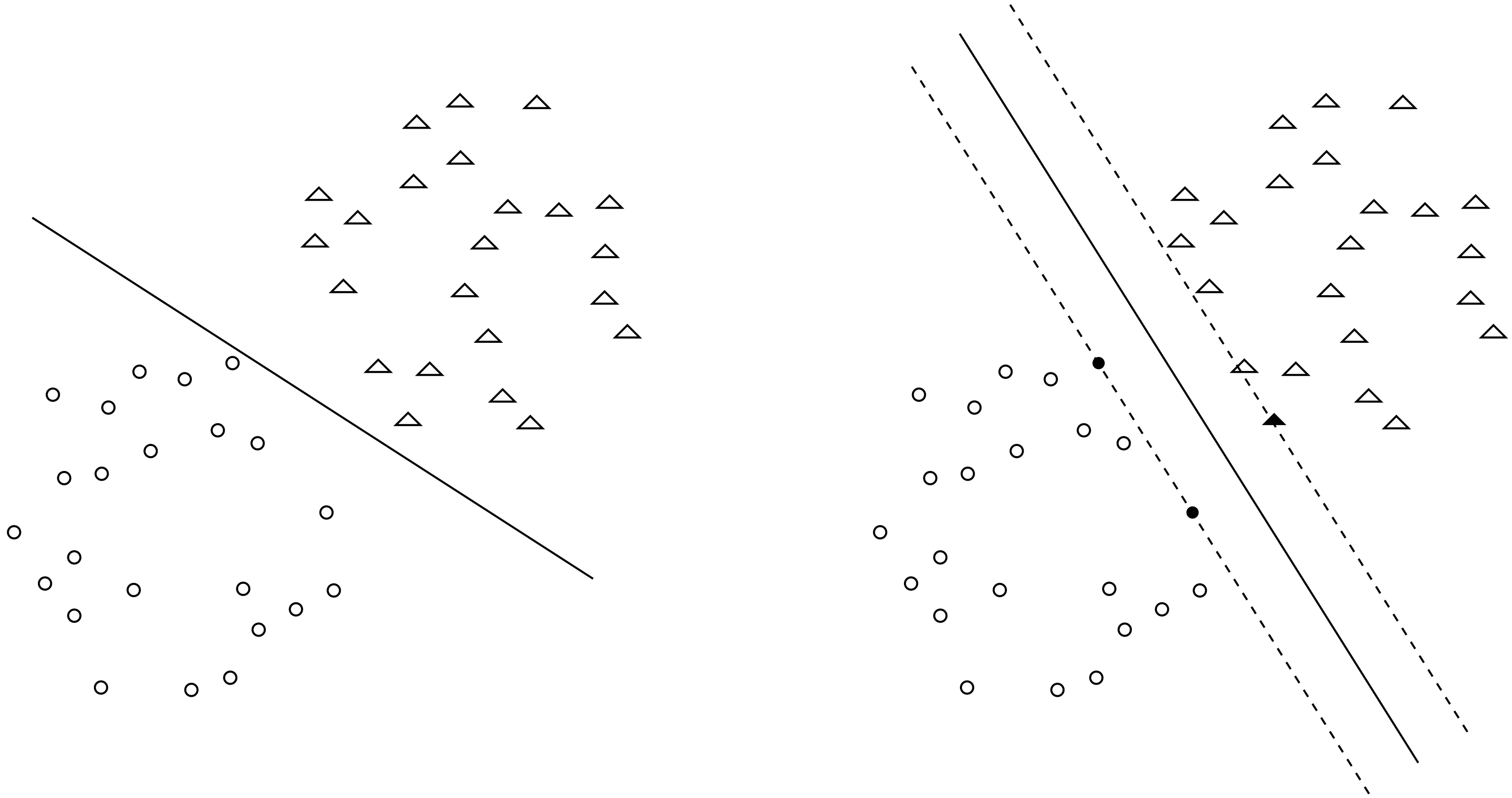
$$\begin{array}{rll}\min\limits_{\mathbf{c}\in\mathbb{R}^\ell,\xi\in\mathbb{R}^\ell} & C\sum_{i=1}^{\ell}\xi_i+\frac{1}{2}\mathbf{c}^TK\mathbf{c} & \\ \text{subject to}: & y_i(\sum_{j=1}^{\ell}c_jK(x_i,x_j)+b)\geq 1-\xi_i & i=1,\ldots,\ell \\ & \xi_i\geq 0 & i=1,\ldots,\ell\end{array}$$

$$\begin{array}{rll}\max\limits_{\alpha\in\mathbb{R}^\ell} & \sum_{i=1}^{\ell}\alpha_i-\frac{1}{2}\alpha^TQ\alpha & \\ \text{subject to}: & \sum_{i=1}^{\ell}y_i\alpha_i=0 & \\ & 0\leq\alpha_i\leq C & i=1,\ldots,\ell\end{array}$$

The Geometric Approach

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.

Large and Small Margin Hyperplanes



(a) (b)

Classification With Hyperplanes

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

$$f(x) = \text{sign }(\mathbf{w} \cdot \mathbf{x}). \tag{1}$$

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 $||\mathbf{x} - \mathbf{x}^{\mathbf{w}}||\ldots$

Deriving the Maximal Margin, I

For some k (assume $k > 0$ for convenience),

$$\begin{aligned} \mathbf{w} \cdot \mathbf{x} &= k \\ \mathbf{w} \cdot \mathbf{x}^{\mathbf{w}} &= 0 \\ \Longrightarrow \quad \mathbf{w} \cdot (\mathbf{x} - \mathbf{x}^{\mathbf{w}}) &= k \end{aligned}$$

Deriving the Maximal Margin, II

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

$$\begin{aligned}
\mathbf{w}\cdot(\mathbf{x}-\mathbf{x}^{\mathbf{w}}) &= \mathbf{w}\cdot\left(\frac{||\mathbf{x}-\mathbf{x}^{\mathbf{w}}||}{||\mathbf{w}||}\mathbf{w}\right)\\
&= ||\mathbf{w}||^{2}\frac{||\mathbf{x}-\mathbf{x}^{\mathbf{w}}||}{||\mathbf{w}||}\\
&= ||\mathbf{w}||\ ||\mathbf{x}-\mathbf{x}^{\mathbf{w}}||\\
&\implies ||\mathbf{w}||\ ||(\mathbf{x}-\mathbf{x}^{\mathbf{w}})|| = k\\
&\implies ||\mathbf{x}-\mathbf{x}^{\mathbf{w}}|| = \frac{k}{||\mathbf{w}||}
\end{aligned}$$

Deriving the Maximal Margin, III

k is a "nuisance paramter". WLOG, 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 $\mathbf{w} \cdot \mathbf{x} = 0$ and $\mathbf{w} \cdot \mathbf{x} = 1$.

The Linear, Homogeneous, Separable SVM

$$\begin{array}{rl} \min\limits_{\mathbf{w} \in \mathbb{R}^n} & ||\mathbf{w}||^2 \\ \text{subject to}: & y_i(\mathbf{w} \cdot \mathbf{x}) \geq 1 \quad i = 1, \ldots, \ell \end{array}$$

Bias and Slack

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

$$f(x) = \text{sign }(\mathbf{w} \cdot \mathbf{x} + b).$$

In practice, we want to work with datasets that are not linearly separable, so we introduce slacks ξ_i, just as before. We can still define the margin as the distance between the hyperplanes $\mathbf{w} \cdot \mathbf{x} = 0$ and $\mathbf{w} \cdot \mathbf{x} = 1$, but this is no longer particularly geometrically satisfying.

The New Primal

s

With slack variables, the primal SVM problem becomes

$$\begin{array}{rll} \min\limits_{\mathbf{w}\in\mathbb{R}^n, b\in\mathbb{R}} & C\sum_{i=1}^{\ell}\xi_i + \frac{1}{2}||\mathbf{w}||^2 & \\ \text{subject to}: & y_i(\mathbf{w}\cdot\mathbf{x}+b) \geq 1-\xi_i & i=1,\ldots,\ell \\ & \xi_i \geq 0 & i=1,\ldots,\ell \end{array}$$

Historical Perspective

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.

More Historical Perspective

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.

The Primal and Dual Problems Again

$$\begin{array}{rcr}\min\limits_{\mathbf{c}\in\mathbb{R}^\ell,\xi\in\mathbb{R}^\ell} & C\sum_{i=1}^\ell \xi_i + \frac{1}{2}\mathbf{c}^T K\mathbf{c} & \\ \text{subject to :} & y_i(\sum_{j=1}^\ell c_j K(x_i,x_j)+b) \geq 1-\xi_i & i=1,\ldots,\ell \\ & \xi_i \geq 0 & i=1,\ldots,\ell \end{array}$$

$$\begin{array}{rcr}\max\limits_{\alpha\in\mathbb{R}^\ell} & \sum_{i=1}^\ell \alpha_i - \frac{1}{2}\alpha^T Q\alpha & \\ \text{subject to :} & \sum_{i=1}^\ell y_i\alpha_i = 0 & \\ & 0 \leq \alpha_i \leq C & i=1,\ldots,\ell \end{array}$$

Optimal Solutions

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

The Reparametrized Lagrangian

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

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

Complementary Slackness

Consider the dual variables are associated with the primal constraints as follows:

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

Complementary slackness tells us that at optimality, either the primal inequality is satisfied with 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

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

Optimality Conditions

All optimal solutions must satisfy:

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

Optimality Conditions, II

The 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) conditons.

Toward Simpler Optimality Conditions — Determining b

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

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

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

Towards Simpler Optimality Conditions, I

Defining our classification function $f(x)$ as

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

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

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

Towards Simpler Optimality Conditions, II

Conversely, suppose $\alpha_i = C$:

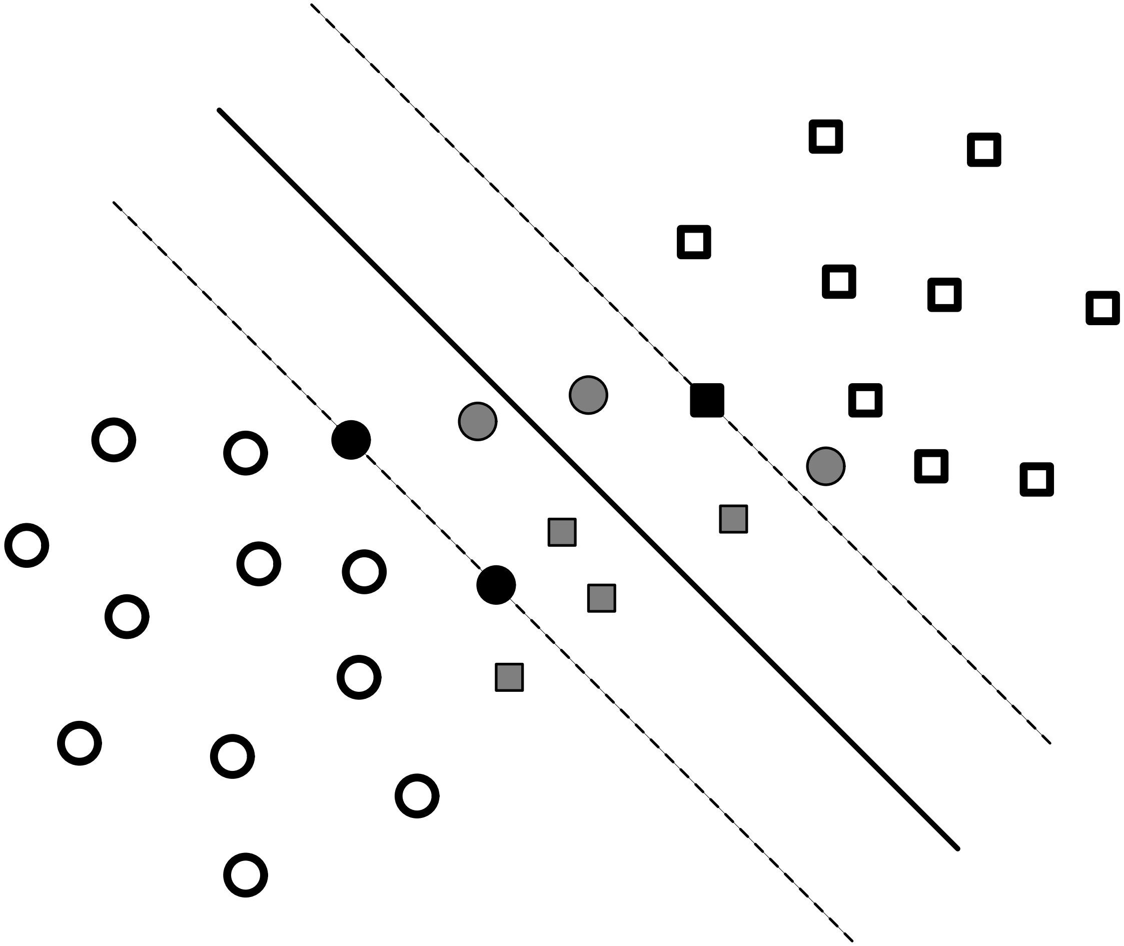
$$\begin{aligned} \alpha_i = C \quad &\Longrightarrow \quad y_i f(x_i) - 1 + \xi_i = 0 \\ &\Longrightarrow \quad y_i f(x_i) \leq 1 \end{aligned}$$

Reduced Optimality Conditions

Proceeding similarly, we can write the following “reduced” optimality conditions (full proof: homework):

$$\begin{aligned} \alpha_i = 0 \quad &\Longleftrightarrow \quad y_i f(x_i) \geq 1 \\ 0 < \alpha_i < C \quad &\Longleftrightarrow \quad y_i f(x_i) = 1 \\ \alpha_i = C \quad &\Longleftrightarrow \quad y_i f(x_i) \leq 1 \end{aligned}$$

Geometric Interpretation of Reduced Optimality Conditions



SVM Training

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.

Off-the-shelf QP software

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

Decomposition, I

Partition the dataset into a *working set* W and the remainig points R. We can rewrite the dual problem as:

$$\max_{\alpha_W \in \mathbb{R}^{|W|},\ \alpha_R \in \mathbb{R}^{|R|}} \quad \sum_{\substack{i=1\\ i\in W}}^{\ell} \alpha_i + \sum_{\substack{i=1\\ i\in R}} \alpha_i$$

$$-\frac{1}{2}[\alpha_{\mathbf{W}}\ \alpha_{\mathbf{R}}] \begin{bmatrix} Q_{WW} & Q_{WR} \\ Q_{RW} & Q_{RR} \end{bmatrix} \begin{bmatrix} \alpha_{\mathbf{W}} \\ \alpha_{\mathbf{R}} \end{bmatrix}$$

$$\text{subject to :} \quad \sum_{i\in W} y_i\alpha_i + \sum_{i\in R} y_i\alpha_i = 0$$

$$0 \le \alpha_i \le C,\ \forall i$$

Decomposition, II

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:

$$\begin{array}{rc} \max\limits_{\alpha_W \in \mathbb{R}^{|W|}} & (\mathbf{1} - Q_{WR}\alpha_{\mathbf{R}})\alpha_W - \frac{1}{2}\alpha_{\mathbf{W}} Q_{WW} \alpha_{\mathbf{W}} \\ \text{subject to :} & \sum_{i \in W} y_i \alpha_i = -\sum_{i \in R} y_i \alpha_i \\ & 0 \leq \alpha_i \leq C, \ \forall i \in W \end{array}$$

Decomposition, III

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.

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.

Good Large-Scale Solvers

- SVM Light: `http://svmlight.joachims.org`
- SVM Torch: `http://www.idiap.ch/learning/SVMTorch.html`. Does regression.
- SVM Fu: `http://fpn.mit.edu/SvmFu`