Mathematical Programming II	Lecture 25
ORIE 6310 Spring 2014	April 29, 2014
Scribe: Jialei Wang	

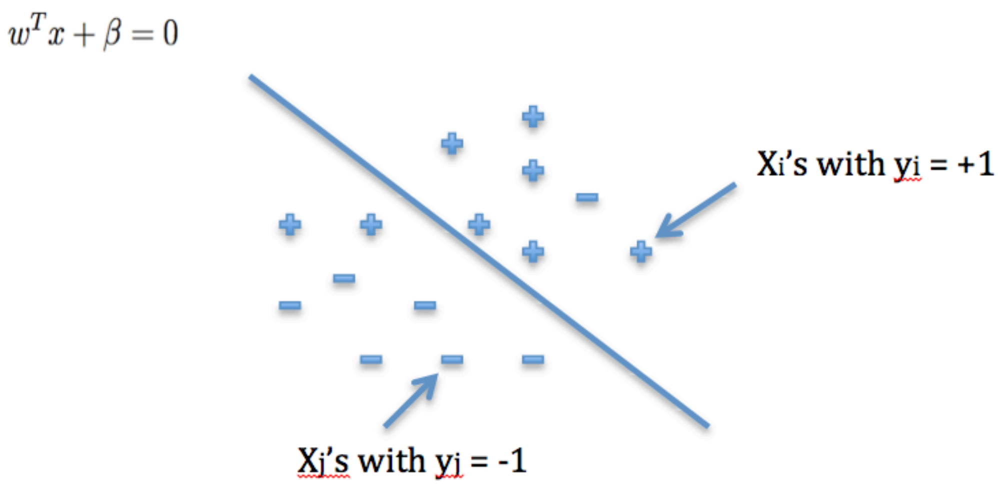
Data Classification, Machine Learning

Suppose we have training data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbf{R}^d$ together with labels $y_1, \ldots, y_n \in \{\pm 1\}$. We want to use this to construct a rule to predict the label of future instances $\boldsymbol{x}$.

We restrict ourselves to linear rules:

$$y = \operatorname{sgn}(w^T x + \beta)$$

for some $w \in \mathbf{R}^d, \beta \in \mathbf{R}$. This is not as restrictive as it looks; e.g. a quadratic rule is a linear rule in $\phi(x) = (x^{(1)}, \ldots, x^{(d)}, (x^{(1)})^2, x^{(1)}x^{(2)}, \ldots, (x^{(d)})^2) \in \mathbf{R}^{d(d+3)/2}$.



Choose w and β so that the rule works well on the training data.We want $x_i^T w + \beta$ to be positive if $y_i = 1$, and negative if $y_i = -1$.

Define $X = [x_1, \ldots, x_n] \in \mathbf{R}^{d \times n}, y \in \mathbf{R}^n, Y = \operatorname{diag}(y_1, \ldots, y_n)$, so we want

$$YX^T w + \beta y$$

to have positive components, and all "big".

We need to normalize: we will choose $||w||_2 \leq 1$.

The data may not be separable, so we allow a perturbation vector $\xi \in \mathbf{R}^n_+$ which we penalize.

So set

$$r = YX^T w + \beta y + \xi,$$

the "signed distance" to the hyperplane $\{x : w^T x + \beta = 0\}$ + perturbations.

We will choose w and β to maximize the smallest r_i, or minimize the largest $\frac{1}{r_i}$, together with a penalty $C > 0$ on each ξ_i. So we get

$$\begin{cases} \min & Ce^T\xi + \frac{1}{p} \\ & YX^Tw + \beta y + \xi - pe \geq 0 \\ & ||w||_2 \leq 1 \\ & w \in \mathbf{R}^d, \beta \in \mathbf{R}, \xi \geq 0 (p > 0). \end{cases}$$

This is roughly equivalent to the Support Vector Machine.

First: replace $\frac{1}{p}$ by q, and add $pq \geq 1$ ($p > 0, q > 0$).

Next: write $p = \rho - \sigma, q = \rho + \sigma$, and we get:

$$\begin{array}{r} pq \geq 1 \\ p > 0, q > 0 \end{array} \Leftrightarrow \begin{array}{c} \rho^2 - \sigma^2 \geq 1 \\ \rho - \sigma > 0 \\ \rho + \sigma > 0 \end{array} \Leftrightarrow \rho \geq \left\| \begin{pmatrix} \sigma \\ 1 \end{pmatrix} \right\|_2.$$

Thus we get the conic programming problem:

$$\begin{cases} \min\limits_{\omega, w, \beta, \xi, \rho, \sigma, \tau, \eta} \quad Ce^T\xi + \rho + \sigma \\ \qquad YX^Tw + \beta y + \xi - \rho e + \sigma e - \eta = 0 \\ \qquad \omega = 1 \\ \qquad \tau = 1 \\ \begin{pmatrix} \omega \\ w \end{pmatrix} \in K_2^{1+d}, \ \beta \in \mathbf{R}, \xi \geq 0, \begin{pmatrix} \rho \\ \sigma \\ \tau \end{pmatrix} \in K_2^{1+2}, \ \eta \geq 0. \end{cases}$$

Its dual is

$$\begin{cases} \max \quad \psi + \theta & = 0 \\ \quad \psi + \upsilon & = 0 \\ XY\alpha + u & = 0 \\ y^T\alpha & = 0 \\ \alpha + \chi & = Ce \\ -e^T\alpha + \lambda & = 1 \\ +e^T\alpha + \mu & = 1 \\ \theta + \nu & = 0 \\ -\alpha + \pi & = 0 \\ \begin{pmatrix} \upsilon \\ u \end{pmatrix} \in K_2^{1+d}, \ \chi \geq 0, \begin{pmatrix} \lambda \\ \mu \\ \nu \end{pmatrix} \in K_2^{1+2}, \ \pi \geq 0. \end{cases}$$

We have $\psi = -\upsilon \leq -||XY\alpha||_2$, so eliminate ψ and put $-||XY\alpha||_2$ in the objective. Next,

$$\begin{aligned} & 0 \leq \alpha \leq Ce \\ \lambda = 1 + e^T\alpha, \mu = 1 - e^T\alpha, \nu = -\theta & \Rightarrow (1 + e^T\alpha)^2 \geq (1 - e^T\alpha)^2 + \theta^2 \\ & \Rightarrow 4e^T\alpha \geq \theta^2. \end{aligned}$$

So we arrive at the simplified dual

$$\begin{cases} \max\limits_{\alpha} & -||XY\alpha||_2 + 2\sqrt{e^T\alpha} \\ & y^T\alpha = 0 \\ & 0 \leq \alpha \leq Ce. \end{cases}$$

Let's assume $y = (+1; +1; \ldots; +1; -1; \ldots; -1)$ and similarly $X = [X_+, X_-]$, $\alpha = (\alpha_+; \alpha_-)$, and $e = (e_+; e_-)$. Then we get

$$\begin{cases} \max\limits_{\alpha_+, \alpha_-} & -||X_+\alpha_+ - X_-\alpha_-||_2 + 2\sqrt{e_+^T\alpha_+ + e_-^T\alpha_-} \\ & e_+^T\alpha_+ = e_-^T\alpha_- \\ & 0 \leq \alpha_+ \leq Ce_+, 0 \leq \alpha_- \leq Ce_-. \end{cases}$$

To simplify further, assume the data is separable, and we eliminate ξ (or make C so large that $\xi = 0$ is optimal).

Then we write $\begin{cases} \alpha_+ = \gamma\overline{\alpha}_+, \text{ with } \gamma = \frac{1}{e_+^T\alpha_+} \\ \alpha_- = \gamma\overline{\alpha}_-, \text{ with } \gamma = \frac{1}{e_-^T\alpha_-} \end{cases}$, (if $e_+^T\alpha_+ = e_-^T\alpha_- = 0$, set $\gamma = 0, \overline{\alpha}_+, \overline{\alpha}_-$ whatever with sums 1).

Then we can rewrite the dual as

$$\max_{\substack{\overline{\alpha}_+ \geq 0 \\ e_+^T\overline{\alpha}_+ = 1}} \max_{\substack{\overline{\alpha}_- \geq 0 \\ e_-^T\overline{\alpha}_- = 1}} \max_{\gamma \geq 0} -\gamma||X_+\overline{\alpha}_+ - X_-\overline{\alpha}_-||_2 + 2\sqrt{2}\sqrt{\gamma}$$

The inner maximization is solved by $\gamma = \frac{2}{||X_+\overline{\alpha}_+ - X_-\overline{\alpha}_-||_2^2}$.

Then we get

$$\begin{cases} \max\limits_{\overline{\alpha}_+, \overline{\alpha}_-} & \dfrac{2}{||X_+\overline{\alpha}_+ - X_-\overline{\alpha}_-||_2} \\ & e_+^T\overline{\alpha}_+ = e_-^T\overline{\alpha}_- = 1 \\ & \overline{\alpha}_+, \overline{\alpha}_- \geq 0. \end{cases}$$

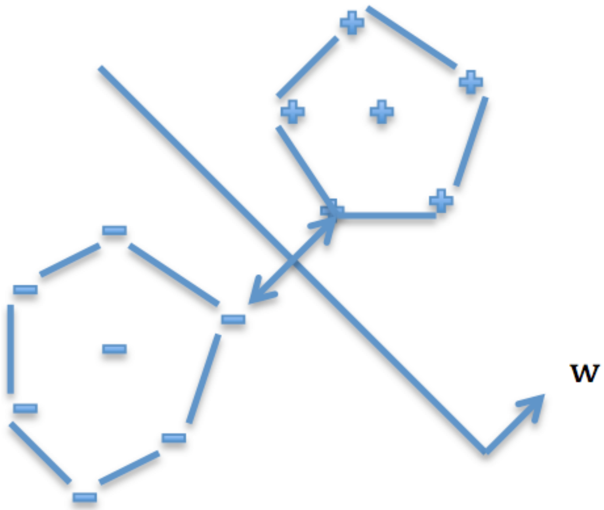


Figure 1: Find the closest points in the convex hulls of the positive points and the negative points

Key: in either the original or the simplified form, strong duality holds, and so "$s^T x = 0$."
In particular, $\begin{pmatrix} \omega \\ w \end{pmatrix} \begin{pmatrix} v \\ u \end{pmatrix} = \begin{pmatrix} 1 \\ w \end{pmatrix}^T \begin{pmatrix} ||XY\alpha|| \\ -XY\alpha \end{pmatrix} = 0.$
As long as there are both positive and negative instances in the data, $\alpha \neq 0$ in the dual opt solution (because of the square root term). So $XY\alpha \neq 0$ in the separable case, and usually otherwise; then w is proportional to $XY\alpha$, so $w = \frac{XY\alpha}{||XY\alpha||_2}$. Hence we can recover the direction of the normal to the hyperplane as the difference of two convex combinations coming from the solution to the dual problem.