Support Vector Machines

Notations

Previously: $w = [w_0, w_1, ..., w_d]$

$x^{(i)} = [1, x_1^{(i)}, ..., x_d^{(i)}]$

For SVM, we separate the intercept term w_0 from the other weights.

$w = [w_1, ..., w_d] \quad w_0$

$x^{(i)} = [x_1^{(i)}, ..., x_d^{(i)}]$

$y \in \{-1, 1\}$

$$g(x) = \begin{cases} 1 & w^T x + w_0 \geq 0 \\ -1 & w^T x + w_0 < 0 \end{cases}$$

Hard Margin SVM

Signed distance from point to hyperplace: $\frac{w^T x^{(i)} + w_0}{||w||_2}$

Geometric margin (>0) of a point: $\gamma^{(i)} = \frac{y^{(i)}(w^T x^{(i)} + w_0)}{||w||_2}$

Functional margin of a point: $\gamma^{(i)} = y^{(i)}(w^T x^{(i)} + w_0)$ = geometric margin if $||w||_2 = 1$

Functional margin of a set: $\gamma = \min\{\gamma^{(1)}, ..., \gamma^{(N)}\}$

Optimization: $\max_{\gamma, w, b} \gamma$ subject to $\frac{y^{(i)}(w^T x^{(i)} + b)}{||w||_2} \geq \gamma$, $i = 1, ..., N$

Constrained optimization: $\max_{\gamma, w, b} \gamma$ subject to $||w||_2 = 1$, $y^{(i)}(w^T x^{(i)} + w_0) \geq \gamma$, $i = 1, ..., N$

$\max_{\gamma, w, b} \frac{\gamma}{||w||_2}$ subject to $y^{(i)}(w^T x^{(i)} + w_0) \geq \gamma$, $i = 1, ..., N$

Rescaling: $\max 1/||w||_2$ subject to $y^{(i)}(w^T x^{(i)} + w_0) \geq 1$, $i = 1, ..., N$

(Canonical weights $w := w_0/\gamma$)

$\min ||w||_2$

$\min ||w||_2^2 = \min(w_1^2 + ... + w_d^2)$

Soft Margin SVM

What if the data isn't linearly separable?

We allow for a few points to be either misclassified or within the margin.

$y^{(i)}(w^T x^{(i)} + w_0) \geq 1 - \xi$

We could incur a cost $\xi^{(i)}$ for how far the $x^{(i)}$ is away from the margin. $\xi = 0$ if on the margin or right side. Large C — lartge penalty for misclassfication.

Objective function: $\min \frac{1}{2}||w||_2^2 + C\Sigma_{i=1}^N \xi^{(i)}$

Setting $\lambda = \frac{1}{C}$, $J(w) = \frac{\lambda}{2} \min ||w||_2^2 + \Sigma_{i=1}^N \max(0, 1 - y^{(i)}(w^T x^{(i)} + w_0))$

$$\text{subgradient}(w) = \begin{cases} \lambda w & y^{(i)}(w^T x^{(i)} + w_0) \geq 0 \\ \lambda w - y^{(i)} x^{(i)} & \text{otherwise} \end{cases}.$$

The Pegasos Algorithm (stochastic regression and adaptive learning rate)

$w = \text{random initialization}$

for t = 1, ..., T:

pick a random example $\{x^{(i)}, y^{(i)}\}$

$\alpha = \frac{1}{\lambda t}$

if $y^{(i)}(w^T x^{(i)}) \geq 1$: // to keep it simple, we will not include a bias unit w_0

$w = w - \alpha \lambda w$

else:

$w = w - \alpha(\lambda w - y^{(i)} x^{(i)})$

Dual Formalization (not required)

$w = \Sigma \alpha^{(i)} y^{(i)} x^{(i)}$

Traning examples where $\alpha^{(i)} \neq 0$ are support vectors.

Kernels (not required)

Replace $\Phi(x^{(i)})^T \Phi(x^{(j)})$ with the kernel $K(x^{(i)}, x^{(j)})$ (see below).

Lagrange Duality solving constrained optimization (not required)

$\min \frac{1}{2}||w||_2^2$ subject to $y^{(i)}(w^T x^{(i)} + w_0) \geq 1$, is equivalent to maximize:

Lagrangian $L(w, w_0, \alpha) = \frac{1}{2}||w||_2^2 - \Sigma_{i=1}^N \alpha^{(i)}(y^{(i)}(w^T x^{(i)} + w_0) - 1)$

$\nabla L(w, w_0, \alpha) = w - \Sigma \alpha^{(i)} y^{(i)} x^{(i)} = 0$

$w = \Sigma \alpha^{(i)} y^{(i)} x^{(i)}$

Plugging them back $L(w, w_0, \alpha) = -\frac{1}{2} \Sigma_{i,j=1}^N \alpha^{(j)} \alpha^{(i)} y^{(j)} y^{(i)} x^{(j)T} x^{(i)} + \Sigma_{i=1}^N \alpha^{(i)}$

CS229

Functional margin of hyperplane defined by (w,b) with repect to $\{x^{(i)}, y^{(i)}\}$:

$\gamma^{(i)} = y^{(i)}(w^T x^{(i)} + b)$

If $y^{(i)} = 1$, want $w^T x^{(i)} + b >> 0$

If $y^{(i)} = -1$, want $w^T x^{(i)} + b << 0$

Want $j^{(i)} >> 0$

If $\gamma^{(i)} > 0$, means $h(x^{(i)}) = y^{(i)}$

Functional margin of hyperplane with repect to trainng set: $\hat{\gamma} = \min \gamma^{(i)}$

Geometric margin of hyperplane defined by (w, b) with repect to $\{x^{(i)}, y^{(i)}\}$:

$\gamma^{(i)} = \frac{w^T x^{(i)} + b}{||w||}$

Geometric margin of hyperplane with repect to trainng set: $\hat{\gamma} = \min \gamma^{(i)}$

Optimal margin classifier:

Choose w, b to maximize the $\hat{\gamma}$

$\max_{\gamma, w, b} \gamma$ subject to $\frac{y^{(i)}(w^T x^{(i)} + b)}{||w||} \geq \gamma$

choose $||w|| = \frac{1}{\gamma}$, becomes $\max \frac{1}{\gamma}$ subject to $y^{(i)}(w^T x^{(i)} + b) \geq 1$

equivalent to $\min \frac{1}{2}||w||^2$ subject to $y^{(i)}(w^T x^{(i)} + b) \geq 1$

Suppose w is a linear combination $w = \Sigma_{i=1}^m \alpha_i y^{(i)} x^{(i)} (y^{(i)} = \pm 1)$ (Representer theorem)

$\min \frac{1}{2}||w||^2$
$= \min \frac{1}{2}(\Sigma_{i=1}^m \alpha_i y^{(i)} x^{(i)})^T (\Sigma_{i=1}^m \alpha_i y^{(i)} x^{(i)})$
$= \min \frac{1}{2}\Sigma_i \Sigma_j \alpha_i \alpha_j y^{(i)} y^{(j)} x^{(i)T} x^{(j)} = \min \frac{1}{2}\Sigma_i \Sigma_j \alpha_i \alpha_j y^{(i)} y^{(j)} < x^{(i)}, x^{(j)} >$

subject to $y^{(i)}((\Sigma_j \alpha_j y^{(j)} x^{(j)})^T x^{(i)} + b) \geq 1$ → $y^{(i)}(\Sigma_j \alpha_j y^{(j)} < x^{(j)}, x^{(i)} > + b) \geq 1$

equivalent to $\max \Sigma_i \alpha_i - \frac{1}{2}\Sigma_i \Sigma_j \alpha_i \alpha_j y^{(i)} y^{(j)} < x^{(i)}, x^{(j)} >$ subject to $\alpha_i \geq 0$, $\Sigma y^{(i)} \alpha_i = 0$ ("Dual Optimization Problem")

$h(x) = g(w^T x + b) = g(\Sigma_i \alpha_i y^{(i)} x^{(i)})^T x + b) = g((\Sigma_i \alpha_i y^{(i)} < x^{(i)}, x > + b)$

Kernel trick

1. Write our algorithm in terms of $< x_i, x_j > or < x, z >$
2. Let there be matching from x to $\Theta(x)$ 20 → 100000
3. Find a way to compute $K(x, z) = \Theta(x)^T \Theta(z)$
4. Replace $< x, z >$ in algorithm with $K(x, z)$

Polynomial kernel: $K(x, z) = (x^T z + c)^d$ — contains all features of polynomials up to d; O(n) linear run time

How to make kernels?

If x, z are similar, $K(x, z)$ is large.

If x, z are dissimilar, $K(x, z)$ is small.

Gaussian kernel: $K(x, z) = exp(-\frac{||x - z||^2}{2\sigma^2})$

L_1 norm soft margin SVM

$\min_{w,n,\xi} \frac{1}{2}||w||^2 + C\Sigma_{i=1}^{m}\xi_i$ subject to $y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi$, $\xi_i \geq 0$

equivalent to $\max \Sigma_i \alpha_i - \frac{1}{2}\Sigma_i \Sigma_j \alpha_i \alpha_j y^{(i)} y^{(j)} < x^{(i)}, x^{(j)} >$ subject to $0 \leq \alpha_i \geq C$, $\Sigma y^{(i)} \alpha_i = 0$