SVM

Hard SVM $\qquad \gamma = \frac{1}{\|w_0\|}$

$$w_0, b_0 = \underset{w, b_0}{\operatorname{argmin}} \|w\|^2 \quad \text{s.t. } y_i(\langle w, x_i\rangle + b) \geq 1 \qquad w = \frac{w_0}{\|w_0\|},\ b = \frac{b_0}{\|w_0\|}$$

def:

A distribution D is (γ, ρ)-separable $\iff$ IT IS SEPARABLE WITH MARGIN γ FOR EVERY x S.T $\|x\| < \rho$, $\|w\| = 1$

Theorem: Let D be a distribution $\mathbb{R}^d \times \{-1, 1\}$ separable with a homogeneous separator ($b = 0$) and a (γ, ρ)-margin. The choice of a training set drawn i.i.d from D with probability at least $1 - \delta$ will have an 1-0 ERROR bounded by

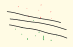
$$\sqrt{\frac{4(\rho/\gamma)^2}{m}} + \sqrt{\frac{2\log(2/\delta)}{m}}$$

SOFT SVM

$$\min_{w, b, \xi} \lambda \|w\| + \frac{1}{m}\sum_{i=1}^{m} \xi_i$$

s.t.

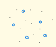
$$\forall i \quad y_i(\langle w, x_i\rangle + b) \geq 1 - \xi_i$$



$k(x, x') = \langle \phi(x), \phi(x')\rangle$

$\hat{\phi}(x) \approx \phi(x)$

$\subset \mathbb{R}^d \qquad$ RFF



Kernel Methods

§1. The Kernel Trick

Def. A map $K: X \times X \to \mathbb{R}$ is called a kernel if there exists $\psi: X \to \mathcal{H}$, $\mathcal{H}$ = Hilbert space such that $K(x, y) = \langle \psi(x), \psi(y)\rangle_{\mathcal{H}}$.

Consider the following optimization problem

$$(*) \quad \min_w \Big(f(\langle w, \psi(x_1)\rangle, \dots, \langle w, \psi(x_m)\rangle) + R(\|w\|) \Big)$$

where $f: \mathbb{R}^m \to \mathbb{R}$, smooth, $R: \mathbb{R}_+ \to \mathbb{R}$, ψ feature map.

$\Rightarrow$ Soft-SVM:

- $R(a) = \lambda a^2$, $\lambda > 0$
- $f(a_1, \dots, a_m) = \frac{1}{m}\sum_{i=1}^{m} \max\{0, 1 - y_i a_i\}$.

$\Rightarrow$ Hard-SVM:

- $R(a) = a^2$
- $f(a_1, \dots, a_m) = \begin{cases} 0 & \exists b : y_i(a_i + b) \geq 1 \\ \infty & \text{ow.} \end{cases}$

Representer Theorem: Assume $\psi: X \to \mathcal{H}$ is a feature map. Then the optimal solution w^* of $(*)$ lies in $\mathrm{lin}_{\mathbb{R}}\{\psi(x_1), \dots, \psi(x_m)\}$.

$$\Rightarrow \exists \alpha_1, \dots, \alpha_m \in \mathbb{R} : w^* = \sum_{i=1}^{m} \alpha_i \psi(x_i)$$

Proof: Omitted.

Corollary: $(*)$ is equivalent to the optimization problem

$$\min_\alpha \Big(f\Big(\sum_j \alpha_j K(x_j, x_1), \dots, \sum_j \alpha_j K(x_j, x_m)\Big) + R\Big(\sqrt{\textstyle\sum_{i,j} \alpha_i \alpha_j K(x_i, x_j)}\Big)\Big)$$

with output $w^* = \sum_{i=1}^{m} \alpha_i^* \psi(x_i)$. This only relies on kernel evaluations $K(x, y) = \langle \psi(x), \psi(y)\rangle_{\mathcal{H}}$.

Proof: Let $w = \sum_{j=1}^{m} \alpha_j \psi(x_j)$. Then

$$\langle w, \psi(x_i)\rangle = \sum_j \alpha_j \langle \psi(x_i), \psi(x_j)\rangle = \sum_j \alpha_j K(x_i, x_j)$$

and

$$\|w\|^2 = \langle w, w\rangle = \Big\langle \sum_j \alpha_j \psi(x_j), \sum_j \alpha_j \psi(x_j)\Big\rangle$$

$$= \sum_{i,j} \alpha_i \alpha_j \langle \psi(x_i), \psi(x_j)\rangle = \sum_{i,j} \alpha_i \alpha_j K(x_i, x_j). \quad \square$$

§2. Implementing Soft-SVM with kernels

We tackle the optimization problem in the feature space

$$\min_w \Big(\frac{\lambda}{2}\|w\|^2 + \frac{1}{m}\sum_{i=1}^{m} \max\{0, 1 - y_i \langle w, \psi(x_i)\rangle\}\Big)$$

for the Sample $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$.

Algorithm: SGD for solving Soft-SVM

Input: $T \leftarrow$ # iterations

$\beta^1 = 0$

for $t = 1 \dots T$

- $\alpha^t \leftarrow \frac{1}{\lambda t}\beta^t$
- $i \leftarrow \{0, \dots, m-1\}$ unif @ random
- for $j \neq i$
  - $\beta_j^{t+1} = \beta_j^t$
- If $\big(y_i \sum_{j=1}^{m} \alpha_j^t K(x_j, x_i) < 1\big)$:
  - $\beta_i^{t+1} = \beta_i^t + y_i$
- Else:
  - $\beta_i^{t+1} = \beta_i^t$

Output: $\bar{w} = \sum_{j=1}^{m} \bar{\alpha}_j \psi(x_j)$ where $\bar{\alpha} = \frac{1}{T}\sum_{t=1}^{T} \alpha^t$.

General paradigm:

1) Given domain X, Choose $\psi: X \to \mathcal{H}$ feature map
2) Given labelled training examples $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$
   Compute $\hat{S} = \{(\psi(x_1), y_1), \dots, (\psi(x_m), y_m)\}$
3) Train a (linear) predictor h over $\hat{S}$
4) Predict $\ell(\hat{x}) = h(\psi(\hat{x}))$ for a training pt $\hat{x}$.

Expressive power: Let $p: \mathbb{R}^n \to \mathbb{R}$ be a degree k multivariate polynomial.

$$\Rightarrow p(x) = \sum_{\substack{J \in [n]^r \\ r \leq k}} w_J \prod_{i=1}^{r} x_{J_i}$$

Then p can be expressed with the hyperplane

$$p(x) = \langle w, \psi(x)\rangle$$

where $\psi(x)_J = \prod_{i=1}^{|J|} x_{J_i}$ and $w = (w_J)_J$.

E_1

$e_1 : \dots$

$e_2 : \dots$

$\mu_1 : \dots$

E_2

$e_1 : \dots$

$\mu_1 : \dots$

$\mu_2 : \dots$

$$k\Big(\sum_{\substack{e \in E_1 \\ e' \in E_2}} k(e, e') + \sum_{\substack{\mu \in E_1 \\ \mu' \in E_2}} k(\mu, \mu')\Big)$$