Non-linear discriminant functions.

$$\phi : R^m \to R^{m'}$$

New training set,

$$\{(z_i, y_i), i = 1, \ldots, n\}, \quad z_i = \phi(x_i).$$

New Dual:

$$\max_{\mu} \; q(\mu) = \sum_{i=1}^{n} \mu_i - \frac{1}{2} \sum_{i,j=1}^{n} \mu_i \mu_j y_i y_j \phi(x_i)^T \phi(x_j)$$

$$\text{s.t} \quad 0 \le \mu_i \le C, \; i = 1, \ldots, n, \; \sum_{i=1}^{n} y_i \mu_i = 0$$

The problem is still a QP problem over R^n irrespective of ϕ & m'.

But we still want to compute $\phi(x)$.

Scanned by CamScanner

Kernel idea.

suppose $\exists$ a fn. $K: R^m \times R^m \rightarrow R$ s.t

$$K(x_i, x_j) = \phi(x_i)\phi(x_j)$$

& computing $K(x_i, x_j)$ is as expensive as $x_i^T x_j$

then, Dual can be solved by replacing

$z_i^T z_j$ by $K(x_i, x_j)$.

what happens during testing?

we have, $W^* = \sum \mu_i^* y_i \, \phi(x_i)$

& $b^* = y_j - \phi(x_j)^T w^* = y_j - \sum_i \mu_i^* y_i \phi(x_i)^T \phi(x_j)$

$\forall$ test pattern x, we need to compute.

$$f(x) = \cancel{\phi}\, \phi(x)^T W^* + b.$$

$$= \sum_i \mu_i^* y_i \phi(x_i)^T \phi(x) + b^*$$

$$= \sum_i \mu_i^* y_i \, K(x_i, x) + \left(y_j - \sum_i \mu_i^* y_i K(x_i, x_j)\right)$$

Scanned by CamScanner

$\therefore$ Never do we need to compute ϕ

In theory, ϕ can even by ∞ dim.

$\therefore$ For an SVM, all that is needed to be stored is

μ_i^* & x_i^* for $i \in S$.

$\Downarrow$

support vectors.

Eg of a Kernel fn in R^2: $K(x_i, x_j) = (1 + x_i^T x_j)^2$

Let $x \in R^2$, $x_i = (x_{i1}, x_{i2})^T$

$$K(x_i, x_j) = (1 + x_{i1}x_{j1} + x_{i2}x_{j2})^2$$

To show, $\exists\ \phi$ in $m' > m$ s.t $\phi(x_i)^T \phi(x_j) = K(x_i, x_j)$.

Consider $\phi : R^2 \rightarrow R^6$

$$\phi(x) = [1\ \ \sqrt{2}x_1\ \ \sqrt{2}x_2\ \ x_1^2\ \ x_2^2\ \ \sqrt{2}x_1x_2]$$

One can show that $\phi(x_i)^T \phi(x_j) = K(x_i, x_j)$

Note: ϕ is non-unique.

Scanned by CamScanner

Kernels in general.

Mercer theorem: Given a symmetric fn

$K : R^m \times R^m \to R$, $\exists$ an inner product

space H & mapping $\phi : R^m \to H$, so that

$K(x_1, x_2) = \phi(x_1)^T \phi(x_2)$ if for all sq. integrable

fns g,

$$\int K(x_1, x_2)\, g(x_1)\, g(x_2)\, dx_1\, dx_2 \geq 0$$

In other words, if $\bar{K}$ a $n \times n$ matrix

with $\bar{K}_{ij} = K(x_i, x_j)$. If $\bar{K}_{nn}$ is PSD for all

n data points, then K is a valid kernel.
(set)

$$\sum_{i,j=1}^{n} c_i c_j K(x_i, x_j) \geq 0$$

Scanned by CamScanner

One can show that

1) Polynomial kernel:

$$K_p(x_1, x_2) = (1 + x_1^T x_2)^p$$

2) Gaussian kernel:

$$K_G(x_1, x_2) = e^{-\frac{\|x_1 - x_2\|^2}{\sigma^2}}$$

3) Sigmoidal kernel:

$$K_S(x_1, x_2) = \tanh(a x_1^T x_2 + \theta)$$

~~Gr~~ all the above satisfy Mercer's theorem.

SVM with K_G:

$$f(x) = \sum_{i \in S} \alpha_i^* y_i K(x_i, x) + b^*$$

$$= \sum_{i \in S} \alpha_i^* y_i e^{-\frac{\|x_i - x\|^2}{2\sigma^2}} + b^*$$

[RBF NN]

Scanned by CamScanner

SVM with Ks

$$f(x) = \sum_{i \in S} \alpha_i^* y_i \tanh(a x^T x_i + \theta) + b^*$$

[NN with one hidden layer with tanh activation
# of nodes in hidden determined by α_i^*]

Why do SVMs perform well:

$$E\,P_{err} \leq \min\left(\frac{s}{n}, \frac{R^2 \|W\|^2}{n}, \frac{m}{n}\right).$$

$s \to$ no of support vectors

$R \to$ radius of smallest sphere enclosing all example.

$\|W\|^{-2}$ – margin of the hyperplane

$m \to$ feature dim

$n \to$ no of example.

1) ~~G~~ good data compression

2) Large margin

3) dim of feature space is small.

Scanned by CamScanner

SVM from a risk min view.

we have $\min_{w,b,\xi_i} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i$

s.t $y_i(w^T x_i + b) \geq 1 - \xi_i, \; i=1,\dots,n$

$\xi_i \geq 0, \; i=1,\dots,n$

Now given any w, b, ξ_i has to satisfy the following.

$$\xi_i \geq \max(0, 1 - y_i(w^T x_i + b))$$

$\therefore$ The above problem can be effectively written as

$$\min_{w, b} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \max(0, 1 - y_i(w^T x_i + b))$$

& final classifier, $f(x) = w^T x + b$.

Scanned by CamScanner

We know that 0-1 loss is non-differentiable.

$L(h(x), y)$ can be made into a fn of single variable $yf(x)$, if $y \in \{-1, 1\}$.

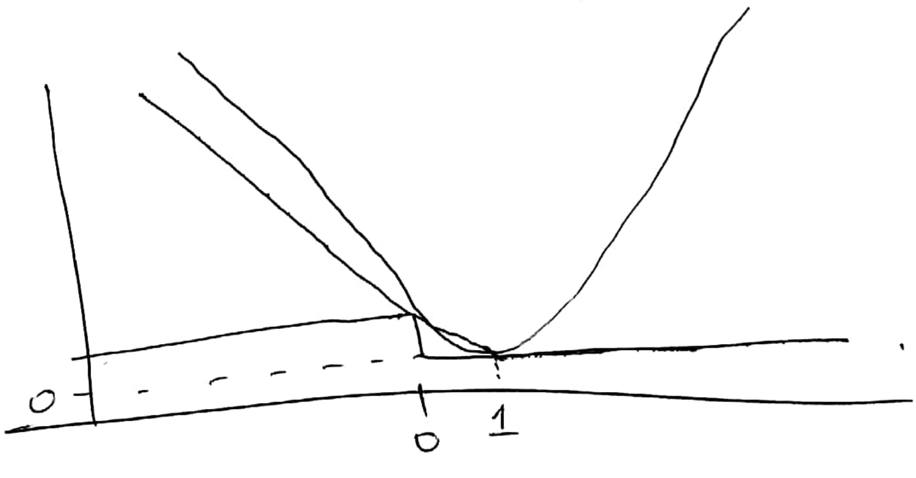
For 0-1 loss, ~~$yh(x)$~~ is ~~< 0~~

$L_{01} = 1$ if $yh(x) < 0$

0 otherwise

$L_{sq\text{-}error} = (1 - yh(x))^2$

$L_{hinge} = \max(0, 1 - yh(x))$



$yh(x)$.

$\therefore$ under this formulation, all losses are Convex approximat. for 0-1 loss.

1194.

Scanned by CamScanner

$\therefore$ SVM optimization problem can be written as follows:

$$\min_{w, b} \frac{1}{n} \sum_{i=1}^{n} L\left(y_i, f(x_i)\right) + C' \frac{1}{2} W^T W$$

$f(x_i) = W^T x_i + b$.

$\therefore$ SVM is empirical risk minimization under hinge-loss with L_2 regularizer.
[soft-margin loss].

How would kernels fit in this framework?

Scanned by CamScanner

Representer theorem:

For any positive definite kernel, $\exists$ a vector space with an inner product $\mathcal{H}$, s.t kernel is the innerproduct in that space.

Mercer thorem says that $\exists$ ϕ from $\mathcal{X}$ to $\mathcal{H}$.

With these, representer theorem says

Let $\Omega : [0, \infty) \rightarrow R^+$ be a strictly monotically increasing function. Then any minimizer g over $\mathcal{H}$ of the regularized risk

$$c\left(\left(x_i, y_i, g(x_i)\right), i=1, \dots n\right) + \Omega\left(\|g\|^2\right).$$

admits a representation

$$g(x) = \underline{\sum_{i=1}^{n} \alpha_i k(x_i, x)}.$$

Scanned by CamScanner

This is a very powerful theorem because it says that the minimizer of the empirical risk is a linear combination of kernels centered around data points alone!!

$\therefore$ Even though $\mathcal{H}$ may be very high dim. One can design an opti problem for risk minimization by searching for n real no α_i

This is precisely what SVM does!!

The idea of kernel is generalizable.

Eg: suppose we are doing knn in $\phi(x)$ dim.

Let $C_+ = \frac{1}{n_+} \sum_{i:y_i=+1} \phi(x_i)$, $C_- = \frac{1}{n_-} \sum_{i:y_i=-1} \phi(x_i)$

Scanned by CamScanner

Now, KNN would put a new pattern in class +1, if

$$\| \phi(x) - C_+ \|^2 < \| \phi(x) - C_- \|^2$$

$$\| \phi(x) - C_+ \|^2 = \phi(x)^T \phi(x) - 2\phi(x)^T C_+ + C_+^T C_+$$

$\Rightarrow$ we put x in class +1 if

$$\phi(x)^T C_+ - \phi(x)^T C_- + \frac{1}{2}\left(C_-^T C_- - C_+^T C_+\right) > 0$$

$$\phi(x)^T C_+ = \phi(x^T) \left(\frac{1}{n_+} \sum_{i : y_i = +1} \phi(x_i) \right)$$

$$= \frac{1}{n_+} \sum K(x_i, x)$$

$$C_+^T C_+ = \frac{1}{n_+^2} \sum_{i,j : y_i = y_j = 1} K(x_i, x_j)$$

Very much related to kernel density estimates.

Scanned by CamScanner