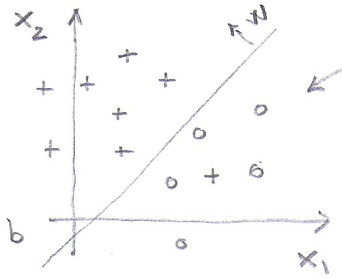
Motivation for SVM

* Two main limitations to statistical learning algorithm (before GPUs):



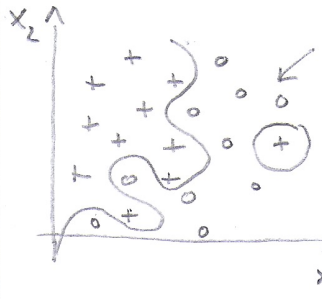
many non separable configurations if the space of decision surfaces is not rich enough

Capacity limitation

↳ idea for solution:

high-dimensional embedding

for a given number of samples configurations become separable in high dimension (assuming general position)



high-capacity empirical risk minimization methods are rich enough to learn noise/outlier features

Overfitting limitation

↳ idea for solution:

Tikhonov regularization

taking irregularity of decision surfaces to be the signature of overfitting, we restrain the space of candidate surfaces by penalizing irregularity.

* Claim: Support vector machines (SVMs) resolve the above tradeoff by implementing a computationally efficient high-dimensional embedding that has good regularity property.

* Theoretical Foundation: Reproducing kernel Hilbert space (RKHS)

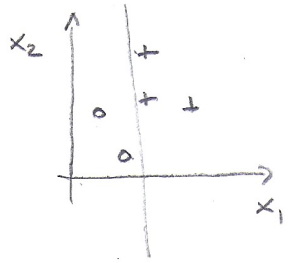
Main idea: choosing a positive definite kernel function $K : X \times X \to \mathbb{R}$ defined on the data space X also defines a high-dimensional (infinite dimensional) embedding in a RKHS H via $\underline{x} \longmapsto (\underline{y} \mapsto K(\underline{x}, \underline{y}))$.

Computationally: The kernel function K measures similarity between data point once embedded and behaves as a dot product: $K(\underline{x}, \underline{y}) = \langle K(\underline{x}, \cdot), K(\underline{y}, \cdot) \rangle_H$ ← Hilbert dot product

Optimal margin classifier

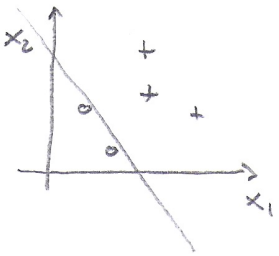
* Algorithms that can exploit the kernel trick need to be formulated in terms of dot product on sample points $\underline{x}_i$, $i = 1, \dots, p$.
* Optimal margin classifiers provide such a formulation for binary linear classification (a.k.a perceptron)

What is the "best" classification?

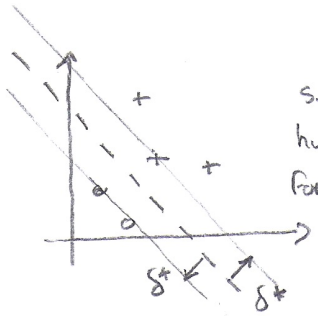


NO

or



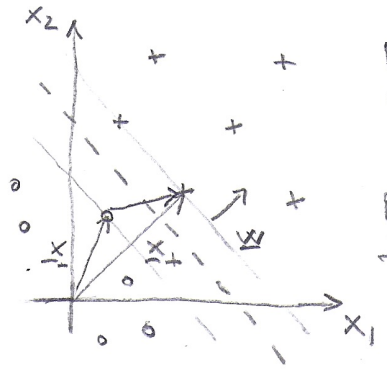
optimal $\Rightarrow$ margin



separating hyperplane for which the margin δ^* is maximal

YES

Constrained optimization problem:



positive points: $\underline{w} \cdot \underline{x}_i \geq b$

negative points: $\underline{w} \cdot \underline{x}_i \leq b$

$\delta^* = \min |\underline{w} \cdot \underline{x} - b| > 0$

$\uparrow$ can be assumed attained for positive and negative point

Rescaling w, b by $1/\delta^*$ is allowed yielding a unit margin.

Thus, introducing the labels $y_i \in \{-1, 1\}^p$, we have:

$$y_i \cdot (\underline{w} \cdot \underline{x}_i - b) \geq 1 \quad \text{with} \quad y_+ \cdot (\underline{w} \cdot \underline{x}_+ - b) = 1$$

$$y_i (\underline{w} \cdot \underline{x}_- - b) = 1$$

if $\underline{x}_+, \underline{x}_-$ are on the positive and negative boundaries respectively.

Margin magnitude

$$\frac{\underline{w}}{\|\underline{w}\|} \cdot (\underline{x}_+ - \underline{x}_-) = \frac{b + 1 - (b-1)}{\|\underline{w}\|} = \frac{2}{\|\underline{w}\|}$$

Optimal margin:

$$\underline{w}^* = \arg\max_{\underline{w}} \frac{2}{\|\underline{w}\|} \quad \text{such that} \quad y_i \cdot (\underline{w} \cdot \underline{x}_i) \geq 1$$

$$= \arg\min_{\underline{w}} \frac{\|\underline{w}\|^2}{2} \quad \text{such that} \quad y_i (\underline{w} \cdot \underline{x}_i) \geq 1$$

This is a quadratic optimization problem subjected to linear constraints.

Lagrangian formulation

To include the constraints in the optimization of the margin we consider the following Lagrangian function

$$L(\underline{w}, \underline{\alpha}) = \frac{1}{2}\|w\|^2 - \sum_i \alpha_i \left[y_i \underbrace{(\underline{w}.\underline{x}_i - b)}_{w_0 \overset{\text{def}}{=}} - 1 \right], \quad \alpha_i \geq 0$$

↑ Lagrange multiplier

We have: $\max_{\underline{\alpha} \geq 0} L(\underline{w}, \underline{\alpha}) = \begin{cases} +\infty & \text{if } y_i(\underline{w}.\underline{x}_i - b) < 1 \text{ for some } i \\ \frac{1}{2}\|w\|^2 & \text{otherwise} \end{cases}$

Thus we seek to find: $\underline{w}^* = \arg \underbrace{\left(\min_{\underline{w}} \max_{\underline{\alpha} \geq 0} L(\underline{w}, \underline{\alpha}) \right)}_{= \overline{L}^*}$

This is the primal problem. Let us consider the dual problem obtained by interverting min and max:

$$\underline{L}^* = \max_{\underline{\alpha} \geq 0} \min_{\underline{w}} L(\underline{w}, \underline{\alpha})$$

In general $\underline{L}^* \leq \overline{L}^*$. Indeed, posing $\begin{cases} \underline{L}^* = \min_{\underline{w}} L(\underline{w}, \underline{\alpha}^*) \\ \overline{L}^* = \max_{\underline{\alpha}} L(\underline{w}^*, \underline{\alpha}) \end{cases}$

we have: $\underline{L}^* = \min_{\underline{w}} L(\underline{w}, \alpha^*) \leq L(\underline{w}^*, \underline{\alpha}^*) \leq \max_{\underline{\alpha}} L(\underline{w}^*, \underline{\alpha}) = \underline{L}^*$

The duality gap $\overline{L}^* - \underline{L}^*$ is zero under certain convexity conditions that are not in our situation.

Thus solving the dual problem (easier) is equivalent to solving the primal problem.

SVM method

Writing the KKT conditions for the dual problem yields

$$\frac{\partial L}{\partial \underline{w}} = 0 \Rightarrow * \ \underline{w} - \sum_i \alpha_i y_i \underline{x}_i = 0 \Leftrightarrow \underline{w} = \sum_i \alpha_i y_i \underline{x}_i$$

$$* \ \sum_i \alpha_i y_i = 0$$

↑ The optimal weight vector is a linear combination of some positions vectors ($\alpha_i \neq 0$) : the support vectors.

The dual problem becomes finding the α_i's, i.e., solving

$$\arg \max_{\substack{\underline{\alpha} \geq 0 \\ \sum_i \alpha_i y_i = 0}} \left\{ \frac{1}{2} \left(\sum_i \alpha_i y_i \underline{x}_i \right) \left(\sum_j \alpha_j y_j \underline{x}_j \right) - \sum_i \alpha_i \left[y_i \left(\sum_j \alpha_j y_j \underline{x}_j \right) \underline{x}_i - 1 \right] \right\}$$

$$= \arg \max_{\substack{\underline{\alpha} \geq 0 \\ \sum_i \alpha_i y_i = 0}} \left\{ -\frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j (\underline{x}_i \cdot \underline{x}_j) \right\}$$

↑ dot product on the data vectors ⇒ kernel trick possible !

$\underline{x}_i \cdot \underline{x}_j \longleftrightarrow k(\underline{x}_i, \underline{x}_j)$ where k is some well-chosen positive definite function

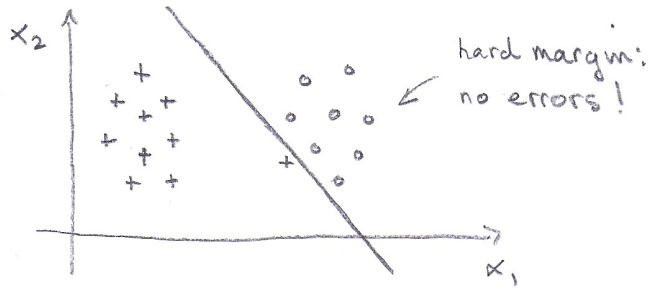
The problem becomes a problem for numerical analysts
The existence and uniqueness of a solution is guaranteed by convexity under assumption of separability.

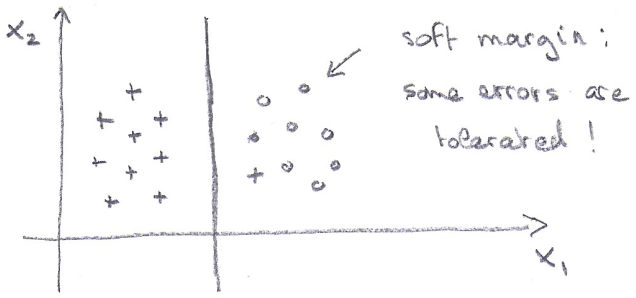
Popular kernel choice : $k(x,y) = (1 + xy)^d$ ← polynomials of degree up to d.

$k(x,y) = e^{-\frac{\|x-y\|}{\sigma}}$ ← radial basis function

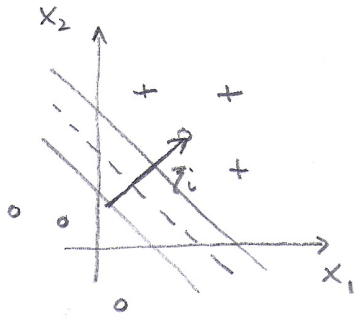
if $\sigma \ll 1$ → tend to overfit. tuning is involved.

Link to regularization





Idea: assigning a cost to errors via new optimization variables called slackness variables ξ_i, $i = 1, \dots, p$.



$$\xi_i = \left[1 - y_i(\underline{w} \cdot \underline{x}_i - b)\right]_+ \leftarrow \text{positive part}$$

Thus $\xi_i \geq 0$ ← constraints

Thus suggests assigning the cost $\sum_i \xi_i$

Soft margin optimization:

trade off tuning variable ↓

Primal problem:

$$\arg\min_{\underline{w}, \xi \geq 0} \left(\frac{1}{2} \|\underline{w}\|^2 + C \sum_i \xi_i \right)$$

such that: $y_i(\underline{w} \cdot \underline{x}_i - b) \geq 1 - \xi_i$

Dual problem:

$$\arg\min_{0 \leq \alpha_i \leq C, \ \sum_i \alpha_i y_i = 0} \left(\frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j (\underline{x}_i \cdot \underline{x}_j) - \sum_i \alpha_i \right)$$

Formulation in term of regularization

$$w^* = \arg\min_w \left(C(\underline{w}, \underline{x}_i, y_i) + \frac{1}{2C} \|w\|^2 \right)$$

↑ empirical risk ↑ norm on the separating vectors

$$F^* = \arg\min_{F \in H} \left[C(F, \underline{x}_i, y_i) + \lambda \Omega\left(\|F\|_H^2\right) \right]$$

- $F \in H$: ↑ RKHS space obtained via the choice of a kernel
- λ: tradeoff parameter
- Ω: strictly increasing function
- $\|F\|_H^2$: ↑ RKHS norm penalizing irregular decision surfaces

Representer Theorem

A priori, regularization problem are infinite dimensional.

The representer theorem state that it is not the case in RKHS.

$$\arg\min_{f\in H}\left[C(f,x_i,y_i)+\lambda\Omega(\|f\|_H^2)\right]\in\left\{\sum_{i=1}^{p}\alpha_i K(x_i,\cdot)\,|\,\alpha_i\in\mathbb{R}\right\}=F$$

Σ finite dimensional.

Sketch of the proof

For f in H, $f=f_{//}+f_{\perp}$, $f_{//}\in F$ and $f_{\perp}\in F^{\perp}$: orthogonal space to the finite dimensional space F.

For all i, $\langle f_{\perp}, K(x_i,\cdot)\rangle = f_{\perp}(x_i)=0$.

Thus, for all $f_{\perp}$, $C(f,x_i,y_i)=C(f_{//},x_i,y_i)$ while

$$\|f\|_H^2=\|f_{//}\|^2+\|f_{\perp}\|^2\geqslant\|f_{//}\|^2.$$

Regularization via Mercer's theorem

Consider a definite positive kernel $K: X\times X\to\mathbb{R}$ such that $\int_X dx\int_X dy\, K(y,x)^2<+\infty$ (For all f,g in $L^2(X)$, we have $\int_X dx\int_X dy\, K(y,x)f(y)f(x)\geqslant 0$.). This defines an operator

$$T_K: L^2(X)\longrightarrow L^2(X),\quad T_K[f](y)=\int_X K(y,x)f(x)\,dx.$$

Then: T_K has a spectral decomposition: there is a complete system of orthonormal functions e_i, $i\in\mathbb{N}$ such that $K(y,x)=\sum_i \lambda_i e_i(y)e_i(x)$, $\lambda_i\to 0$

Σ eigenvalues $\sum_i \lambda_i^2<+\infty$

The above theorem allows one to specify a high-dimensional embedding into the RKHS associated to K.

Consequences

Posit: $\phi(x) = \sum_i \sqrt{\lambda_i}\, e_i(x)\, e_i(\,) \in H \subset L_2(x)$

We have: $\langle \phi(x), \phi(y) \rangle_{L_2(x)} = \sum_i \sum_j \sqrt{\lambda_i \lambda_j}\, e_i(x) e_j(y) \underbrace{\langle e_i, e_j \rangle}_{\delta_{ij}}$

natural dot product

$= \sum_i \lambda_i e_i(x) e_i(y)$

$= K(x,y)$ ← reproducing kernel

How to relate $\langle , \rangle_H$ to $\langle , \rangle_{L_2}$?

By finding $\langle , \rangle_H$ such that $K(x,y) = \langle K(x,\cdot), K(y,\cdot) \rangle_H$ holds.

Posit $\langle f, g \rangle_H = \sum_i \frac{1}{\lambda_i} \int_x e_i(x) f(x) dx \int_x e_i(x) g(x) dx$

$$\langle K(x,\cdot), K(y,\cdot) \rangle_H = \sum_i \frac{1}{\lambda_i} \lambda_i e_i(x)\, \lambda_i e_i(x) = K(x,y)$$

$F \in H$, $\|f\|_H^2 = \sum_i \frac{\left(\int e_i(x) F(x) dx\right)^2}{\lambda_i}$ ← component of f on e_i

$\lambda_i \downarrow 0$ as $i \to +\infty$

coefficients must decays to 0 faster than λ_i as $i \to +\infty$

High-frequency components are penalized by $\| \ \|_H$

Translation-invariant kernel

$T_K[f](y) = \int_x K(y-x) f(x) dx$, eigenfunction $f_\omega(x) = e^{i\omega x}$

$T_K[f_\omega](y) = e^{i\omega y} \int_x K(y-x) e^{i\omega(x-y)} dx = \hat{K}(\omega) f_\omega(y)$

↑ Fourier transform

$\phi(x) = \left(\int_0^{+\infty} \sqrt{\hat{K}(\omega)}\, e^{i\omega x} d\omega \right) e^{i\omega(\cdot)}$ ← argument

↑ positive measure by Bochner theorem

$\langle f, g \rangle_H = \int_0^{+\infty} d\omega \frac{\overline{\hat{f}(\omega)}\, \hat{g}(\omega)}{\hat{K}(\omega)}$, $\|f\|_H^2 = \int_0^{+\infty} d\omega \frac{\|\hat{F}(\omega)\|^2}{\hat{K}(\omega)} \xrightarrow[\omega \to +\infty]{} 0$

↑ penalize irregularity