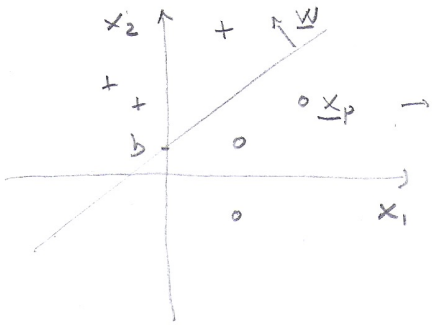
Fundamental limitations to statistical machine learning

Ex. of task: given a set of data points $\underline{x}_\mu$, $\mu = 1, \dots, P$ in $\mathbb{R}^N$ learn a given binary classification, i.e., a true or false assignment. $y_\mu \in \{-1, 1\}^P$.

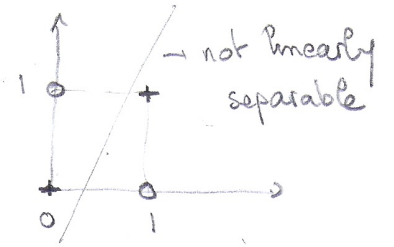
Ex. of algorithm: perceptron algorithm = linear classifier



$\underline{x}_\mu \rightarrow \underline{w}^T \cdot \underline{x}_\mu + b < 0 \rightarrow y_\mu = -1$ assignment.

Problem 1: Not all assignments are linearly separable.

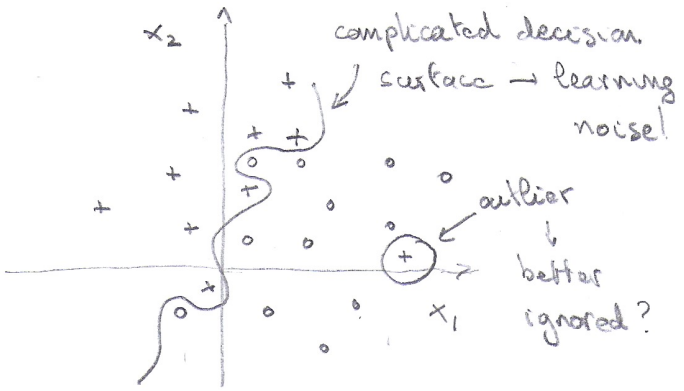
classical example: XOR logical gate



The space of decision surfaces generated by linear functions, i.e. the capacity of the perceptron is not rich enough.

This is the capacity limitation.

Problem 2: Suppose we have an algorithm with high capacity obtained by considering a very rich set of generating functions. Then arbitrary classifications can be realized



Learning algorithms with high capacity are proned to overfit: learning irrelevant details that do not generalized to new sampled data.

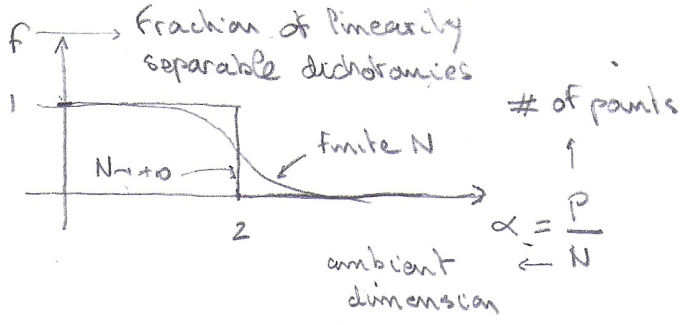
Support vector machines

* Support vector machines (SVM) were developped to jointly address the problem of limited capacity and the problem of overfitting. Vapnik (1992, 1993) with ideas dating back Vapnik (1963).
* The justification for support vector machines being able to realize a good trade-off between the two above limiting problems is rooted in a mathematical theory, that of "reproducing kernel Hilbert space" (RKHS). Aronszajn (1950).

Goal: Introducing informally the ideas leading to considering RKHS and proving the key theorem of RKHS theory by Aronszajn.

(I) Addressing capacity limitation

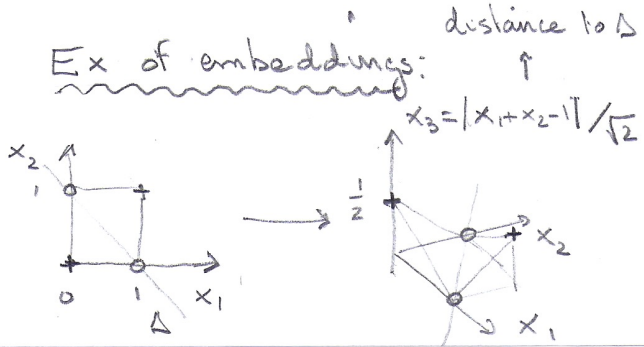
Recall Cover's counting function theorem:

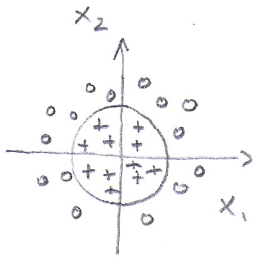


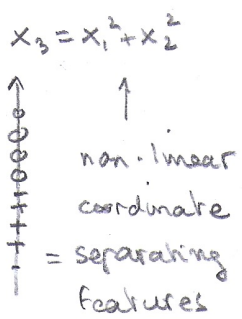
All patterns (in general position) are learnable if the dimension of the embedding space is large enough.

This suggests projecting the data in a high dimensional space via non linear embedding.

Ex of embeddings:







In principle, separating features are unknown, requiring to consider many features as coordinates of the embedding

Ex. $\phi_3 : \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} \longmapsto \left(x_1^3, \dots, x_N^3, x_1^2 x_2, \dots, x_N^2 x_{N-1} \right)$ ← all monomials of degree 3.

Such monomial embeddings ϕ_d corresponds to a feature space of $\left(\!\binom{N}{d}\!\right) = \binom{N+d-1}{d}$. For $\begin{cases} N = 28 \times 28 \leftarrow \#\text{pixels in MNIST} \\ d = 3 \rightarrow \sim 8.10^7 \text{ dimensions} \end{cases}$

⚠ Combinatorial explosion = "curse of dimensionality"
It is computationally infeasible to work with such high-dimensional embedding directly.

Idea: Many machine learning algorithms are based on rules involving computing similarity measure via dot product.

Ex: perceptron-like clustering algorithm.

Can one compute high-dimensional dot product without specifying the embedding of a configuration?

Yes! : Ex: $\phi_d(\underline{x}) \cdot \phi_d(\underline{y}) = \sum_{d_1=1}^{d} \cdots \sum_{d_N=1}^{d} \mathbb{1}_{\{d_1 + \dots + d_N = d\}} \left(x_1^{d_1} \cdots x_N^{d_N}\right)\left(y_1^{d_1} \cdots y_N\right)^{d_N}$

$$= \sum_{d_1=1}^{d} \cdots \sum_{d_N=1}^{d} \mathbb{1}_{\{d_1 + \dots + d_N = d\}} \left(x_1 y_1\right)^{d_1} \cdots \left(x_N y_N\right)^{d_N}$$

$$= \left(\sum_{i=1}^{N} x_i y_i \right)^d = \left(\underline{x} \cdot \underline{y}\right)^d \leftarrow \text{dot products in } N \text{ dimensions}$$

Goal of SVM: realizing implicit high-dimensional embedding via the choice of a dot product.

(II) Addressing the overfitting problem

* Recall the perceptron update rule: $\underline{w}_n = \underline{w}_{n-1} + \eta\, y_n \underline{x}_n$ if $y_n \cdot \underline{w}_{n-1}^T \underline{x}_n < 0$ (incorrect classification)

This rule can be interpreted as performing a gradient descent on the data-dependent cost function:

$$C(\underline{w}, \underline{x}_\mu, y_\mu) = -\sum_\mu y_\mu \cdot (\underline{w}^T \cdot \underline{x}_\mu)\, \mathbb{1}_{\{y_\mu \cdot \underline{w}^T \cdot x_\mu < 0\}}$$

* Many learning algorithms can be formulated as such cost minimization methods. These are called empirical risk minimization (ERM) methods and consist in finding a possibly nonlinear function f^* such that:

$$f^* = \arg\min_{f \in \mathcal{F}} C(f, \underline{x}_\mu, y_\mu)$$

$\mathcal{F}$ ← space of possible function depends on the complexity of the considered generating functions and the dimensionality of embeddings.

⚠ ERM methods are pronned to overfitting for high-capacity choices of generating functions.

Idea: The signature of overfitting is the occurrence of highly irregular, complicated decision surfaces. One may avoid overfitting by constraining the set of solutions to ERM methods to be regular.

Tikhonov regularization: $f^* = \arg\min \{ C(f, \underline{x}_\mu, y_\mu) + \lambda \|f\|_H \}$

(λ: Lagrange parameter; $\|f\|_H$: some norms that penalizes irregular function)

Goal of SVM: realizing Tikhonov regularization by choosing a dot-product for which the corresponding norm penalizes irregular functions.

Reproducing kernel Hilbert space (RKHS)

Theoretical framework of SVM methods.

Central ideas: * a choice of dot product defines a high-dimensional space of functions.

* this space of function must inherit some regularity property from the dot product.

Recall: * A Hilbert space is a complete and separable equipped with a norm $\|\|$ induced by a dot product $\langle , \rangle$

* A dot product is a bilinear, symmetric, positive definite form:

Ex: $L_2(0,1)$, $\langle f, g \rangle_{L_2(0,1)} = \int_0^1 f(t)\, g(t)\, dt$.

Def. RKHS: Let X be a set (i.e. empirical data). Let H be a class of function forming a Hilbert space with dot product $\langle , \rangle_H$. The function $K: X \times X \longrightarrow \mathbb{R}$ is a reproducing kernel of H if: (Γ canonical embedding)

1] H contains all functions $K_x : t \longmapsto K(x,t)$

2] For all x in X and f in H, the reproducing property holds:

$$f(x) = \langle f, K(x, \cdot) \rangle_H$$

($\Rightarrow$ kernel trick: $K(x,y) = \langle K(x,\cdot), K(y,\cdot) \rangle$)

Def. positive definite kernel: Let X be a set. The symmetric function $K: X \times X \longrightarrow \mathbb{R}$ is a positive definite kernel if and only if for all $x_1, \dots, x_n$ in X and all reals $c_1, \dots, c_n$, we have:

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

Remarks: * The definition of RKHS is equivalent to requiring that the evaluation functional $\delta_x : f \mapsto f(x)$ be a bounded linear functional. Then the Riesz representation theorem implies that $K(x, \cdot)$ is the representer of δ_x:

$$f(x) = \langle f, \delta_x \rangle = \langle f, K(x, \cdot) \rangle .$$

Thus $L^2(0,1)$ is not a RKHS as its evaluation functionals are not even well-defined

* Positive definite kernels can be easily generated via simple algebraic rules. For instance positive weighted sums and products of positive definite kernels are positive definite. Covariance functions are positive definite.

Key theorem by Aronszajn (1950)

For any set X, a function $K : X \times X \to \mathbb{R}$ is positive definite if and only if it is a reproducing kernel (RK) of a RKHS H.

Sketch of the proof:

($\Rightarrow$) Suppose K is a RK of RKHS H. Then K is symmetric

$$K(x,y) = \langle K(x,\cdot), K(y,\cdot) \rangle_H = \langle K(y,\cdot), K(x,\cdot) \rangle_H = K(y,x)$$

(kernel trick; symmetry of dot product)

Moreover, for all $x_1, \dots, x_n$ in X and for all reals $c_1, \dots, c_n$

$$\sum_i \sum_j c_i c_j K(x_i, x_j) = \sum_i \sum_j c_i c_j \langle K(x_i,\cdot), K(x_j,\cdot) \rangle_H$$

$$= \left\langle \sum_i c_i K(x_i,\cdot), \sum_j c_j K(x_j,\cdot) \right\rangle_H$$

$$= \left\| \sum_i c_i K(x_i,\cdot) \right\|_H^2 \geq 0 \qquad \checkmark$$

($\Leftarrow$) 1st step: build a pre-RKHS space associated to K, i.e., a space H_0 that has all the properties of a RKHS space except completeness.

2nd step: complete the pre-RKHS space H_0 into the full RKHS space H via natural limit arguments.

1st step: Define $H_0 = \left\{ \sum_{i=1}^{n} a_i K(x_i, \cdot) \;\middle|\; n \in \mathbb{N}, x_i \in X, a_i \in \mathbb{R} \right\}$

$= \{$ Finite linear combination of kernel evaluated on $X \} \subset \mathbb{R}^X$

For f, g in H_0, $f = \sum_{i=1}^{n} a_i K(x_i, \cdot)$, $g = \sum_{j=1}^{m} b_j K(y_j, \cdot)$,

define:

$$\langle f, g \rangle_{H_0} = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j K(x_i, y_j)$$

We have $\langle f, g \rangle_{H_0} = \sum_{i=1}^{n} a_i g(y_i) = \sum_{j=1}^{m} b_j f(x_i)$.

i) Thus $\langle \cdot, \cdot \rangle_{H_0}$ is a bilinear symmetric form that depends only on f and g.

ii) Moreover: $\langle f, K(x, \cdot) \rangle = \sum_{i=1}^{n} a_i K(x_i, x) = f(x)$.

iii) Finally: $\|f\|^2_{H_0} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j K(x_i, x_j) \geq 0$ by positive definiteness

$$|f(x)|^2 = |\langle f, K(x, \cdot) \rangle|^2 \underset{\text{Cauchy Schwarz}}{\leq} \|f\|_{H_0} K(x, x)^{1/2}$$

Thus: $\|f\|_{H_0} = 0 \Rightarrow f = 0$

i), ii) iii) prove that H_0 is a pre-RKHS for K.

2nd step: A] How to extend H_0 to H?

Consider a Cauchy sequence f_n in H_0: $\forall \varepsilon, \exists N > 0$

$n, m > N \quad \|f_n - f_m\|_{H_0} \leq \varepsilon$.

By Cauchy Schwarz inequality, we have:

$$|f_n(x) - f_m(x)| \leq \|f_n - f_m\|_{H_0} K(x,x)^{1/2} \leq \varepsilon K(x,x)^{1/2}$$

Thus $f_n(x)$ converges as a Cauchy sequence of real numbers

Denote $f(x) = \lim_{n \to +\infty} f_n(x)$ the pointwise limit of f_n.

Define $H = \{ f \in \mathcal{X}^{\mathbb{R}} \mid f(x) = \lim_{n \to \infty} f_n(x)$ for some Cauchy sequence in $H_0 \}$

B] How to extend $\langle , \rangle_{H_0}$ on H?

Consider f, g in H. There are f_n, g_n in H_0 such that $f_n \to f$ and $g_n \to g$ point wise,

Cauchy Scharwz inequality implies that

$$|\langle f_n, g_n \rangle_{H_0} - \langle f_m, g_m \rangle_{H_0}| \leq |\langle f_n - f_m, g_n \rangle_{H_0} + \langle f_m, g_n - g_m \rangle_{H_0}|$$
$$\leq \|f_n - f_m\|_{H_0} \|g_n\|_{H_0} + \|f_m\|_{H_0} + \|g_n - g_m\|_{H_0}$$

Thus $\langle f_n, g_n \rangle_{H_0}$ is a real Cauchy sequence.

Moreover its limits only depends on the point wise limits f and g. The latter point follows from the lemma:

Lemma: f_n is H_0 Cauchy and point-wise convergent to 0.
Then $\|f_n\|_{H_0} \longrightarrow 0$. $n \to +\infty$.

Proof: $\|f_n\|$ is necessarily bounded by, say, $B > 0$

By Cauchy property, $\exists N$, $\forall n \geq N$, $\|f_n - f_N\|_{H_0} < \varepsilon / B$

Moreover f_N can be written as: $F_N = \sum_{i=1}^{p} c_i K(x_i, \cdot)$

For some p, x_i, c_i. Then:

$$\|f\|^2_{H_0} = \langle f_n - f_N, f_n \rangle_{H_0} + \langle f_N, f_n \rangle_{H_0} \leq \underbrace{\frac{\varepsilon}{B} \|f_n\|_{H_0}}_{\leq \varepsilon} + \sum_{i=1}^{p} c_i f_n(x_i)$$
$$\underbrace{\qquad\qquad}_{n \to +\infty} 0$$

The previous Lemma implies that if we have ($f_n \to F$, $g_n \to g$; $f_n' \to F$, $g_n' \to g$) pointwise for f_n, f_n', g_n, g_n' H_0 Cauchy

then $\|f_n - f_n'\|_{H_0} \to 0$, $\|g_n - g_n'\|_{H_0} \to 0$ and by Cauchy Schwarz

$$\lim_{n\to+\infty} \langle f_n', g_n' \rangle_{H_0} = \lim_{n\to+\infty} \langle f_n, g_n \rangle_{H_0} = \underset{\text{def}}{\uparrow} \langle F, g \rangle_H$$

4] Proving that H is a RKHS for K

i] Definite positiveness of $\| \ \|_H$:

Suppose $f \in H$ such that $\|f\|_H = 0$.

There is f_n in H_0 such that $f_n \longrightarrow f$ pointwise

$$|f(x)| = \lim_{n\to+\infty} |f_n(x)| = \lim_{n\to+\infty} |\langle f_n, K(x,\cdot)\rangle_{H_0}|$$

$$\leq K(x,x)^{1/2} \lim_{n\to+\infty} \|f_n\|_{H_0} = 0$$

Thus $\|f\|_H = 0 \Rightarrow f = 0$.

ii] Completeness of H

By construction, H_0 is dense in H.

Consider a H Cauchy sequence f_n. There is a H_0 sequence f_n' such that $\|f_n - f_n'\|_H \to 0$ by density. For all $\varepsilon > 0$ there is $N > 0$ such that $\forall n, m > N$

$$\|f_n' - f_m'\|_{H_0} \leq \|f_n' - f_m\|_H + \|f_m - f_m\|_H + \|f_m' - f_m\|$$

$$\leq \varepsilon/3 + \varepsilon/3 + \varepsilon/3 = \varepsilon$$

Thus f_n' is a H_0 Cauchy sequence and f_n converge point-wise to some function F in H.