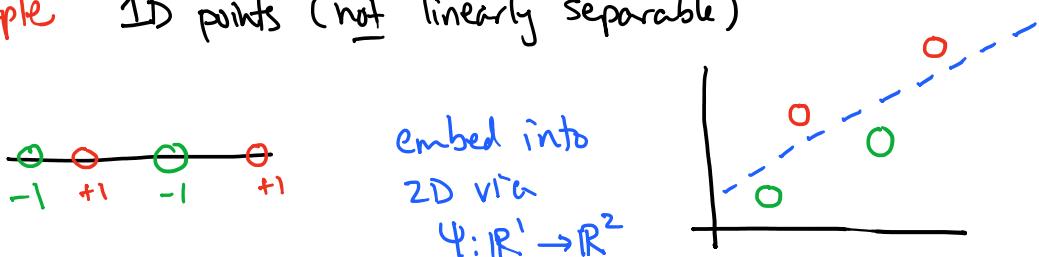


Ch 16 Kernel Methods

Friday, March 27, 2020 3:05 PM

A common trick is to embed data x into a higher dimension $\Psi(x)$, especially because if the embedding Ψ is nonlinear, then a linear classifier in high-dimensional space corresponds to a non-linear classifier in the original space (but the non-linearity is done in pre-processing, conceptually at least, so it doesn't hurt things like convexity)

Example 1D points (not linearly separable)

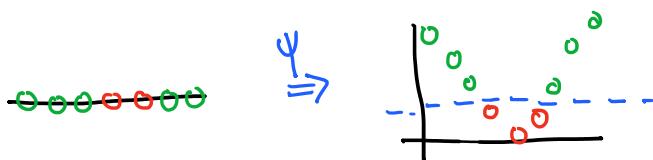


... now in 2D, the dataset is linearly separable
putting in #'s to make a real example:

Consider a dataset w/ "x" as $\{-10, -9, \dots, 9, 10\}$

let labels y be defined as $y = \begin{cases} +1 & |x| > 2 \\ -1 & |x| \leq 2 \end{cases}$
so not lin. separable

but define $\Psi: \mathbb{R}^1 \rightarrow \mathbb{R}^2$ as $\Psi(x) = (x, x^2)$, and
now it is lin. sep. in \mathbb{R}^2



So, in theory, it's simple:

$\Psi: X \rightarrow F \leftarrow$ new feature space,

preprocess data. Our distribution D over $X \times Y$ is now extended to D^Ψ over $F \times Y$

in the obvious way, so $L_{D^\Psi}(h) = L_D(h \circ \Psi)$

We've even seen this trick before: we never discussed **quadratic** or **polynomial** regression because we can treat them as a special case of **linear** regression (after adding extra features)

For most learning algorithms to work, we'll need F to at least be a **Hilbert Space**.

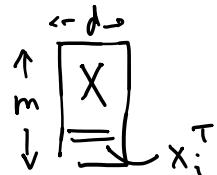
The Kernel trick

Often it's unclear what Ψ to use, and/or we don't want to work in F (why? it's larger-dim, even ∞ -dim!)

The Kernel trick is a way around this, applicable to many (not all) algorithms.

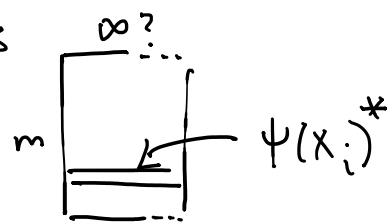
Ex Kernel Ridge Regression (Saunders et al., ICML '98)
(my simplified derivation w/o duality)

$$\min_w \frac{1}{2} \|Xw - y\|^2 + \lambda \|w\|^2$$



Solution is $w = (X^T X + \lambda I)^{-1} \cdot \underbrace{X^T y}_b = \sum_{i=1}^m y_i \psi(x_i)$

Now, suppose X is



our sol'n is still valid
in any Hilbert space,
but

$$X^T X = \sum_{i=1}^m \psi(x_i) \psi(x_i)^*$$

isn't easy to compute

Trick:

$$(X^T X + \lambda I)^{-1} = I - X^T (I + X X^T)^{-1} X \quad \begin{matrix} \text{choose } \lambda = 1 \text{ for} \\ \text{simplification} \end{matrix}$$

(matrix inversion lemma)

\rightarrow a $\tilde{d} \times \tilde{d}$ matrix if $\psi: \mathbb{R}^d \rightarrow \mathbb{R}^{\tilde{d}}$

How does that help? $X^T X = \sum_i \psi(x_i) \psi(x_i)^*$ is tricky

$$(X^T X)_{ij} = \psi(x_i)^* \psi(x_j) = \underline{\langle \psi(x_i), \psi(x_j) \rangle}$$

$K(x, \tilde{x}) = \langle \psi(x), \psi(\tilde{x}) \rangle$ is a kernel

So, ψ defines K , but...

you can define K without needing to know what ψ is!

In particular, let the kernel matrix be K such that aka Gram matrix

$$(K)_{ij} = K(x_i, x_j) = \langle \psi(x_i), \psi(x_j) \rangle, \quad K \in \mathbb{R}^{m \times m}$$

So, back to ridge regression,

$$\begin{aligned} w &= (I - X^T (I + K)^{-1} X) \cdot b, \quad b = X^T y = \sum_i y_i \psi(x_i) \\ &= b - X^T \tilde{K} X b, \quad \tilde{K} = (I + K)^{-1} \in \mathbb{R}^{m \times m} \end{aligned}$$

↑
w/o knowing ψ ,
this isn't computable

computable

but... we don't actually want w , we want to apply it, i.e., $h(x) = \langle w, x \rangle$

So, can we compute $h(x) = \langle w, x \rangle$?

or, since we're embedding x to $\psi(x)$, can we compute $h(x) = \langle w, \psi(x) \rangle$?

yes! $w = b - X^T \tilde{K} X b$

$$\begin{aligned} \langle b, \psi(x) \rangle &= \langle \sum_i y_i \psi(x_i), \psi(x) \rangle \\ &= \sum_i y_i \langle \psi(x_i), \psi(x) \rangle = \sum_i y_i \underline{\underline{k}(x_i, x)} \end{aligned}$$

computable

and

$$\begin{aligned}\langle X^T \tilde{K} X b, \psi(x) \rangle &= \langle \psi(x), X^T \tilde{K} X^T b \rangle \\ &= c^T \tilde{K} d \quad \text{where} \quad c_i = \langle \psi(x), \psi(x_i) \rangle \\ &\qquad\qquad\qquad d_i = \langle \psi(x_i), \sum_i y_i \psi(x_i) \rangle\end{aligned}$$

So all computable via only
Knowing $K(x, \tilde{x})$ (ψ is completely implicit)!

That's the kernel trick

General theory

The Kernel trick applied to ridge regression.

Does it apply to SVM? (yes)

What else?

Thm 1b.1 "Representer Thm"

Suppose $\psi: X \rightarrow \mathcal{H}$ Hilbert space now, not hypothesis class

and let $\phi \neq \arg \min_w f(\langle w, \psi(x_1) \rangle, \dots, \langle w, \psi(x_m) \rangle) + R(\|w\|)$

for arbitrary f and R is nondecreasing on \mathbb{R}_+

$$(\text{ex: } f(a_1, \dots, a_m) = \frac{1}{m} \sum_i l(a_i), \quad R(a) = a^2)$$

Note Hard + Soft + SVM, and Ridge Regression, satisfy the assumptions)

then $\exists \alpha \in \mathbb{R}^m$ s.t.

$$w = \sum \alpha_i \psi(x_i) \quad \begin{array}{l} (\text{ie, } w \in \text{span}\{\psi(x_1), \dots, \psi(x_m)\}) \\ \text{and we can use the kernel trick,} \end{array}$$

since now $\langle w, \psi(x) \rangle$ is computable

Proof:

let w^* be an optimal sol'n, decompose it as

$$w^* = w + u, \quad w \in \text{span}\{\psi(x_1), \dots, \psi(x_m)\} = V \subseteq \mathcal{H}$$

$$u \in V^\perp \quad \text{since } \mathcal{H} = V \oplus V^\perp \quad (\dim(V) < \infty)$$

so via orthogonality,

$$\|w^*\|^2 = \|w\|^2 + \|u\|^2 \Rightarrow \|w\| \leq \|w^*\|$$

$$\text{and } R \text{ nondecreasing} \Rightarrow R(\|w\|) \leq R(\|w^*\|)$$

Since w^* and w differ only on V^\perp ,
they both have the same value of
 $f(\langle w, \psi(x_1) \rangle, \dots, \langle w, \psi(x_n) \rangle)$

hence if w^* is optimal, w must be optimal
as well, and $w \in V$ as desired. \square

(book has slight error: it didn't assume optimal
solution exists)

Conclusion: Kernel-SVM is practical

Examples of Kernels

Given $K(\cdot, \cdot)$, sometimes we have an explicit expression
for ψ , but not always

(and not just any $K(\cdot, \cdot)$ works ... more on that later)

① Polynomial Kernel

$$K(x, x') = (c + \langle x, x' \rangle)^k \quad (\psi \text{ not unique})$$

$$= \langle \psi(x), \psi(x') \rangle \text{ for } \begin{array}{l} \psi: \mathbb{R}^d \rightarrow \mathbb{R}^{\tilde{d}}, \\ \tilde{d} = \binom{d+k}{k} \text{ or } \tilde{d} = (d+1)^k \end{array}$$

ex, $d=2$,

$$K(x, x') = (c + x_1 x'_1 + x_2 x'_2)^2 = \begin{bmatrix} x_1^2 & & \\ x_2^2 & & \\ \sqrt{2} x_1 x_2 & & \\ \sqrt{2} c x_1 & & \\ \sqrt{2} c x_2 & & \\ c & & \end{bmatrix}^T \begin{bmatrix} x'_1^2 & & \\ x'_2^2 & & \\ \sqrt{2} x'_1 x'_2 & & \\ \sqrt{2} c x'_1 & & \\ \sqrt{2} c x'_2 & & \\ c & & \end{bmatrix}$$

for larger d , you can
form ψ , but you don't
want to

② Gaussian Kernel (aka RBF, Radial Basis Fun, or "SE")

most common in practice

$$K(x, x') = e^{-\frac{\|x-x'\|^2}{2\sigma^2}} \leftarrow \begin{array}{l} \text{for } \sigma > 0 \dots \\ \text{or often we write } e^{-\frac{\|x-x'\|^2}{2\sigma^2}} \\ \text{to make it clear } \sigma^2 > 0 \\ \text{and in analogy to normal distr.} \end{array}$$

what is ψ ?

consider $d=1$ case, $K(x, x') = e^{-\frac{(x-x')^2}{2}}$ (for now)

define $\psi: \mathbb{R} \rightarrow \ell^2$ by $\psi(x) = e^{-\frac{x^2}{2}} x$ $\ell^2 \text{ on } \{0, 1, 2, \dots\}$

$$(\psi(x))_n = \frac{1}{\sqrt{n!}} e^{-\frac{x^2}{2}} x^n$$

$$\text{So } \|\psi(x)\|_{\ell^2}^2 = e^{-\frac{x^2}{2}} \cdot \left(\sum \frac{1}{n!} (x^n)^2 \right)$$

$$= 1 \quad \text{since } e^a = \sum \frac{1}{n!} a^n \text{ has } \infty \text{ radius of convergence}$$

and note

$$\langle \psi(x), \psi(x') \rangle_{\ell^2} = \sum (\dots) \quad (\dots)$$

$$= \exp\left(-\frac{x^2}{2} - \frac{x'^2}{2}\right) \sum_{n=0}^{\infty} \underbrace{\frac{(xx')^n}{n!}}_{= e^{xx'}} = e^{-\frac{x^2}{2} + xx' - \frac{x'^2}{2}}$$

$$= \exp\left(-\frac{(x-x')^2}{2}\right)$$

③ Sigmoid Kernel

$$a, b \geq 0, \quad K(x, x') = \tanh(a \langle x, x' \rangle + b)$$

(w/ SVM, this is similar to a simple neural network)

④ discussion

The choice of Kernel is one way to encode prior knowledge
(just as choice of features was)

Similar to classic statistics: for regression,

if x_i is a covariate, sometimes you
choose to use $\log(x_i)$ instead

Often polynomial or Gaussian kernels are the only
choices considered (and their parameters are
hyperparameters that are fit)

... but kernels also used many places
 (ex: Gaussian processes, Kriging in spatial stat.)
 and there, often custom kernels are used.

For more ex., see exercises in ch 6 in Mohri et al.

What are the restrictions on K ?

Lemma 16.2 (Mercer's Thm, Simplified)

A symmetric function $K: X \times X \rightarrow \mathbb{R}$ implements an inner product in some Hilbert space iff K is psd (pos semidef.)

A kernel is psd if $\forall m, \forall \{x_1, \dots, x_m\}$, the Gram matrix

$G_{ij} = K(x_i, x_j)$ is a psd matrix

$$G = G^T, G \geq 0 \text{ iff } \lambda_i(G) \geq 0 \quad \forall i$$

Turns out not all Hilbert spaces can be described w/ a kernel

Those that can are **Reproducing Kernel Hilbert Spaces (RKHS)**

In particular, a RKHS has the property

that it is over an underlying domain \mathcal{S} (think of $L^2(\mathcal{S})$)

for $\mathcal{S} = [0, 1]$

and if $x \in \mathcal{S}$,

or $\mathcal{S} = \mathbb{R}$,

$f \in \mathcal{H}$, then $f_x \in \mathcal{H}' (\cong \mathcal{H})$

or $L^2(\mathcal{S})$ for $\mathcal{S} = \mathbb{N}$ or \mathbb{Z})

ex: $\mathcal{H} = l^2(\mathbb{N})$, $\mathcal{S} = \mathbb{N}$,

$f \in l^2$ is $f = (f_1, f_2, f_3, \dots)$

then $f_n(f) = f_n$ is a bounded linear functional

$l^2(\mathbb{N})$ is a RKHS (so is $l^2(\mathbb{Z})$)

ex: $H = L^2(\Omega = [0,1])$

$f \in L^2, x \in [0,1],$

$\delta_x(f) = f(x) \quad (= \int_{\Omega} f(x') \delta(x-x') dx')$
in physics notation

this is not bounded

(not even well-defined, since L^2 is
not defined pointwise)

So $L^2([0,1])$ is not a RKHS
(nor is $L^2(\mathbb{R})$)

History

Kernels in general, Mercer (1909)

Schoenberg (1938)

* Aronszajn (1950)

In ML, Boser, Guyon, Vapnik '92 (SVM)

Schölkopf et al '98 (general)

Supplement

Filling the Gram matrix is slow!

$O(m^2d)$ computation

many attempts to improve

One broad strategy is the Nyström method

(also used in integral eq'n) which
picks out representative pts.

A well-known 2007 paper (Rahimi, Recht)

makes approx. maps Ψ st.

$$K(x, x') \approx \langle \Psi(x), \Psi(x') \rangle$$

for Shift-invariant kernels

$$K(x, x') = G(x - x')$$
 Ex Gaussian $\exp\left(-\frac{1}{2}\|x - x'\|^2\right)$
Laplacian $\exp\left(-\|x - x'\|\right)$
Cauchy $\prod_{i=1}^d \frac{1}{1 + (x_i - x'_i)^2}$

relying on

Bochner's Thm A cts kernel of the form $K(x, x') = G(x - x')$ defined over a locally compact set X is pos. def. iff G is the Fourier transform of a non-neg. measure, i.e., $G(x) = \int_X e^{i\omega x} p(\omega) d\omega$

For standard Shift-inv. kernels, $p(\omega)$ is known

Idea: approximate $G(x)$ via Monte Carlo integration: draw a set

$$\{\omega_i\}_{i=1}^N, \text{ iid}, \omega_i \sim p(\omega), \text{ equiv. to an}$$

approx. feature map into \mathbb{R}^N . Of course MC converges very slowly but not too bad if X has small covering number

(and in practice, low-accuracy is ok, since we don't need to approximate the kernel very well, just enough to give us "features" to learn from)