Statistical Machine Learning

UoC Stats 37700, Winter quarter

Lecture 7: Kernel methods.

Euclidean structure methods

- We have already seen a couple of methods depending only either on a notion of *distance* between training (and test) inputs, or of a scalar product between said points.
- One common method to make for example linear separators more flexible is to add more coordinates to the input observations, or more generally to map them explicitly into some higher-dimensional Euclidean “feature space”:

- In many common cases, for all purposes it is actually sufficient to be able to compute dot products $\langle \Phi(X), \Phi(X') \rangle$ of input points mapped in feature space.
- This begs the natural question: when is a real function

$$k : (x, x') \in \mathcal{X} \times \mathcal{X} \mapsto k(x, x') \in \mathbb{R}$$

  the dot product for some mapping $x \mapsto \Phi(x)$ into some Euclidean feature space E?
- For example

$$k(x, x') = (x \cdot x' + c)^2$$

  is the “kernel” for the mapping

$$x \mapsto \Phi(x) = \left[(x_i x_j)_{i,j}, (\sqrt{2c} x_i)_i, c \right] .$$

Fundamental characterization theorem

Theorem

Given a set $\mathcal{X}$ and a function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$, there exists a Hilbert space $\mathcal{H}$ and a mapping $\Phi \mathcal{X} \to \mathcal{H}$ such that $k(x, x') = \Phi(X) \cdot \Phi(X')$
if and only if
the function k is of positive type, i.e. for any integer $n > 0$, for any n-uple $(x_1, \ldots, x_n) \in \mathcal{X}^n$, and $(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$,

$$\sum_{i,j} \alpha_i \alpha_j k(x_i, x_j) \geq 0 \, .$$

Further properties of reproducing kernel Hilbert spaces (RKHS)

- A RKHS is a Hilbert space of real functions on a space $\mathcal{X}$.
- The kernel function $k : \mathcal{X}^2 \mapsto \mathbb{R}$ is such that for all $x \in \mathcal{X}$, the function $k(x,.)$ belongs to $\mathcal{H}$. Furthermore,

$$\forall x, y \in \mathcal{H}^2 \quad \langle k(x,.), k(y,.)\rangle = k(x,y)\,.$$

- The above implies the general "reproducing" property:

$$\forall f \in \mathcal{H}, x \in \mathcal{X}, \; \langle f, k(x,.)\rangle = f(x)\,.$$

- The above implies that the evaluation functional in a point $x \in \mathcal{X}$:

$$f \in \mathcal{H} \mapsto f(x) \in \mathbb{R}$$

  is a continuous function $\mathcal{H} \rightarrow \mathbb{R}$ for all x.
- Conversely, any Hilbert space of functions on $\mathcal{X}$ satisfying this last property is a RKHS and the two first properties characterize its kernel.

The representer theorem revisited

The representer theorem can be rewritten for RKHS spaces under an interesting form:

Theorem

Let $\mathcal{H}$ be a reproducing kernel Hilbert space. Consider an optimization problem of the form

$$\operatorname*{Arg\,Min}_{f \in \mathcal{H}, b} \Psi((f(X_i))_{1 \leq i \leq n}, \|f\|_{\mathcal{H}}) ,$$

where Ψ is a function nondecreasing in its last variable.
Then the solution $f^* \in \mathcal{H}$ is a linear combination of the $k(X_i, .)$'s,

$$f(x) = \sum_i a_i k(X_i, x) .$$

Kernels and regularity

- Suppose k is a reproducing kernel for RKHS $\mathcal{H}$; then we have the formula, for any $f \in \mathcal{H}$:

$$|f(x) - f(y)| \leq \|f\|_{\mathcal{H}} \, dist_k(x, y) \,,$$

  where $dist_k$ is the distance on $\mathcal{X}$ implicitly defined by the kernel.
- Hence the RKHS norm represents a bound on the Lipschitz constant of tbe function relative to the distance defined by the kernel (this is kind of auto-referential, but still interesting!).
- Note also that if the kernel is bounded, the norm of functions in $\mathcal{H}$ is an upper bound on their supremum norm.

When is a function a kernel?

- Of course the standard dot product $k(x, z) = x \cdot z$ is a kernel.
- Some basic kernel transformations: if k_1, k_2 are kernels and f is a real function on $\mathcal{X}$, and a a positive numer, then the following are kernels:
  - $k(x, z) = k_1(x, z) + k_2(x, z)$
  - $k(x, z) = ak_1(x, z)$
  - $k(x, z) = k_1(x, z)k_2(x, z)$
  - $k(x, z) = f(x)f(z)$

Applications

- The "normalization" of kernel is a kernel:

$$k'(x,y) = \frac{k(x,y)}{\sqrt{k(x,x)k(y,y)}} .$$

  Note that it corresponds to the transformed feature mapping $x \mapsto \Phi'(x) = \frac{\Phi(x)}{\|\Phi(x)\|_{\mathcal{H}}}$.
- A polynomial function with nonnegative coefficients of a kernel is a kernel.
- A convergent series with nonnegative coefficients of a kernel is a kernel.
- The Gaussian kernel is a kernel:

$$k_\sigma(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) .$$

Kernel distances and conditionally positive definite (CPD) kernels

- Many Euclidean learning methods are actually invariant by translation of the datapoints. Hence, they only depend on the distances between the points.
- Let k be a positive type kernel and d the associated distance, i.e.

$$d^2(x,y) = k(x,x) + k(y,y) - 2k(x,y) .$$

  Then d^2 is a conditionally negative function, i.e.

$$\forall (x_i), (\lambda)_i \text{ with } \sum_i \lambda_i = 0 \; : \; \sum_{i,j} \lambda_i \lambda_j d(x_i, x_j) \leq 0 .$$

- Conversely, any d^2 satisfying the above condition is a squared distance corresponding to a positive type kernel. To see this, pick any "origin" $x_0 \in \mathcal{X}$ and use the formula to compute dot product from square distances:

$$k_0(x,y) = -\frac{1}{2} \left(d^2(x,y) - d^2(x,x_0) - d^2(y,x_0) \right) .$$

Properties of cpd functions

Theorem

If $f : \mathcal{X}^2 \to \mathcal{R}^+$ *is a conditionally negative function taking nonnegative (!) values, then so are* f^α, $\alpha \in [0, 1]$, *and* $\log(1 + f)$.

An interesting consequence is that for any Euclidean norm $\|x\|$, the function $d^2(x, y) = \|x - y\|^\beta$, for $\beta \in [0, 2]$ is conditionally negative definite.

One striking aspect of ν-SVM based on this family of distances is that it is invariant by translation and scale change in the input space.

When is a kernel useful?

- Kernels seem to be a wonderful general tool... Are all kernels potentially useful?
- Example of a useless kernel: $k(x, x) = 1$ and $k(x, y) = 0$ if $x \neq y$.
- In general, one should try to embed some kind of prior knowledge in the kernel used.
- Remember a kernel is implicitly a Euclidean strtucture: the underlying "distance" should somehow reflect what we think is important to compare examples.

Separable RKHSs

- A separable Hilbert space has by definition a countable dense subset, or, equivalently, a countable Hilbert basis $(\phi_1, \phi_2, \ldots)$.
- If a RKHS is separable, then the kernel can be but under the form

$$k(x,y) = \sum_{i,j} \phi_i(x)\phi_j(y) \,,$$

  for any Hilbert basis of $\mathcal{H}$.
- If the kernel is bounded, the sum converges uniformly in x for any fixed y.
- Conversely, any function $f(x,y)$ of the above form is a reproducing kernel.

Example: translation invariant kernels on a compact interval

- Consider a kernel of the following form, for $x, y \in [0, 1]$:

$$k(x, y) = k_0(x - y)$$

- Assume $k_0 : [-1, 1] \to \mathbb{R}$ is the sum of its Fourier series

$$k_0(t) = \sum_{n=0}^{\infty} a_n \cos(nt),$$

  with $\sum_k |a_k| < \infty$ ensuring absolute convergence.
- Then the kernel k can be expanded as

$$k(x - y) = a_0 + \sum_{n \geq 1} a_n \sin(nx) \sin(ny) + \sum_{n \geq 1} a_n \cos(nx) \cos(ny).$$

- Hence k is a reproducing kernel iff $a_i \geq 0$ for all i.

A criterion for separable RKHS

Theorem

Let k be a positive type symmetric kernel.
If k is continuous, then the "feature mapping"

$$x \in \mathcal{X} \mapsto k(x, .) \in \mathcal{H}$$

is continuous.
If additionally $\mathcal{X}$ is a separable space $\mathcal{X}$, then the associated RKHS is a separable Hilbert of continuous functions.
Furthermore, for any Hilbert basis (ϕ_i), the representation

$$k(x,y) = \sum_{i,j} \phi_i(x)\phi_j(y)$$

converges uniformly in (x, y) on any compact.

Kernels, smoothness and differential operators

- Consider the Hilbert space of real functions f on $[0, 1]$, with $f(0) = 0$, a.e. derivable, with the scalar product

$$\langle f, g\rangle = \int_0^1 f'(x)g'(x)dx.$$

- This space is a RKHS with kernel $k(x, y) = \min(x, y)$.
- This can be extended to more general differential operators D, then the kernel is the Green function for operator D^*D.

Kernels and integral operators

- For ν a distribution on $\mathcal{X}$, assume that $\int k(x,x)d\nu(x) < \infty$. An important operator is the kernel integral operator

$$L_k : f \in L^2(\nu) \mapsto L_k f(x) \int k(x,y) f(y) d\nu(y) \,.$$

- This operator is Hilbert-Schmidt and can be written as TT^*, where T is the inclusion operator from $\mathcal{H}$ to $L^2(\nu)$.
- If $\mathcal{X}$ is compact, $L^2(\nu)$ is a separable Hilbert and there exists a diagonalizing basis (ψ_i, λ_i) for the operator L_k.
- If additionally k is continuous, and ν has full support, $\sqrt{\lambda_i}\psi_i$, forms an orthogonal basis of $\mathcal{H}$.
- The unit ball of $\mathcal{H}$ can be seen as a compact (in fact Hilbert-Schmidt) ellipsoid in $L^2(\nu)$.

An interesting extension of a previous result is the case of translation-invariant kernels on $\mathbb{R}^d$:

$$k(x,y) = k_0(x-y)\,.$$

Theorem (Bochner's theorem (more or less))

If k_0 is (Lebesgue) integrable and its Fourier transform $\widehat{k_0}$ is real nonnegative and integrable, k is a reproducing kernel and the associated RKHS consists in continuous, integrable functions f satisfying

$$\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\frac{\left|\widehat{f}(u)\right|^2}{\widehat{k_0}(u)}du < \infty\,,$$

with the inner product

$$\langle f,g\rangle_{\mathcal{H}} = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\frac{\widehat{f}(u)\widehat{g}^*(u)}{\widehat{k_0}(u)}du\,.$$