Kernels Cribsheet

kirstym

January 2017

Cauchy-Schwarz Inequality

$$|\langle u, v\rangle|^2 \leq \langle u, u\rangle . \langle v, v\rangle$$

$$|\langle u, v\rangle| \leq \|u\| \|v\|$$

where

$$\|u\| = \sqrt{\langle u, u\rangle}$$

Covariance stuff

Covariance in different spaces

The usual sample covariance is written:

$$C = \frac{1}{N} \sum_i^N (x_i - \bar{x})(x_i - \bar{x})^T$$

The (finite) feature space equivalent is:

$$C = \frac{1}{N} \sum_i^N (\phi(x_i) - \overline{\phi(x)})(\phi(x_i) - \overline{\phi(x)})^T$$

To extend this to an infinite dimensional feature space, we use the Kronecker product which is a generalization of the outer product from vectors to matrices.

$$\begin{aligned} C &= \frac{1}{N} \sum_i^N (\phi(x_i) - \overline{\phi(x)}) \otimes (\phi(x_i) - \overline{\phi(x)}) \\ &= \frac{1}{N} \sum_i^N \widetilde{\phi}(x_i) \otimes \widetilde{\phi}(x_i) \end{aligned}$$

From outer to inner products

With finite vectors, it is easy to see that

$$(ab^T)c = (b^T c)a$$

The infinite-dimensional analog is:

$$(a \otimes b)c = \langle b, c\rangle_{\mathcal{H}} a$$

This identity will come in handy in lots of derivations!

Centering using matrix multiplication

To compute covariance we have to centre our data. To make the algebra simpler, this is often done using a matrix operation, using the centering matrix $H = \boldsymbol{I}_{nxn} - \frac{1}{n}\mathbf{1}_{nxn}$. This allows us to compute:

- Row centred matrix $X_r = HX$
- Column centred matrix $X_c = XH$
- Row *and* column centred matrix $X_{rc} = HXH$

It's also important to note that $HH = H$ and $H^T = H$.
This allows us to write covariance (for example in the data space case) as:

$$\begin{aligned} C &= \frac{1}{N}\sum_{i}^{N}(x_i - \bar{x})(x_i - \bar{x})^T \\ &= \frac{1}{N}\sum_{i}^{N}(XH)(XH)^T \\ &= \frac{1}{N}\sum_{i}^{N}XHH^TX^T \\ &= \frac{1}{N}\sum_{i}^{N}XHX^T \end{aligned}$$

Miscellanea

Project a point onto a component f, assuming $\|f\| = 1$:

$$P_f\phi(x^*) = \langle \phi(x^*), f\rangle_{\mathcal{H}} f$$

Bounded operators and Riesz representation theorem

A linear operator $A : \mathcal{F} \to \mathcal{R}$ is bounded when (for some λ_A):

$$|Af| \leq \lambda_A \|f\|_{\mathcal{F}} \qquad \forall f \in \mathcal{F} \tag{1}$$

In a Hilbert space $\mathcal{F}$, all bounded linear operators can be written $\langle \cdot, g_A\rangle$ for some $g_A \in \mathcal{F}$:

$$Af = \langle f(\cdot), g_A(\cdot)\rangle \tag{2}$$

Mean embeddings

The mean embedding is defined as $\mu_P \in \mathcal{F}$ such that:

$$\mathbb{E}_P\left(f(X)\right) = \langle f, \mu_P\rangle_{\mathcal{F}} \tag{3}$$

These things are all true:

$$\mathbb{E}_{P,Q}\left(k(X,Y)\right) = \langle \mu_P, \mu_Q\rangle_{\mathcal{F}} \tag{4}$$

$$\mathbb{E}_{P,P'}\left(k(X,X')\right) = \langle \mu_P, \mu_P\rangle_{\mathcal{F}} \tag{5}$$

$$\mu_P = [..., \sqrt{\lambda_i}\mathbb{E}_P[e_i(X)], ...] \tag{6}$$

$$\begin{aligned}
\mu_P(x) &= \langle \mu_P, \phi(x) \rangle_{\mathcal{F}} && (7)\\
&= \langle \mu_P, k(\cdot, x) \rangle_{\mathcal{F}} && (8)\\
&= \mathbb{E}_P k(\cdot, x) && (9)
\end{aligned}$$

Maximum mean discrepancy

$$\begin{aligned}
MMD(P, Q; \mathcal{F}) &= \sup_{f \in \mathcal{F}} [\mathbb{E}_P f(X) - \mathbb{E}_Q f(Y)] && (10)\\
&= \sup_{f \in \mathcal{F}} [\langle f, \mu_P \rangle - \langle f, \mu_Q \rangle] && (11)\\
&= \sup_{f \in \mathcal{F}} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}} && (12)\\
&= \|\mu_P - \mu_Q\|_{\mathcal{F}} && (13)
\end{aligned}$$

$$\begin{aligned}
MMD^2(P, Q; \mathcal{F}) &= \langle \mu_P - \mu_Q, \mu_P - \mu_Q \rangle && (14)\\
&= \langle \mu_P, \mu_P \rangle + \langle \mu_Q, \mu_Q \rangle - 2\langle \mu_P, \mu_Q \rangle && (15)\\
&= \mathbb{E}_P \mathbb{E}_{P'} k(x, x') + \mathbb{E}_Q \mathbb{E}_{Q'} k(y, y') - 2\mathbb{E}_P \mathbb{E}_Q k(x, y) && (16)\\
&= \overline{K_{P,P}} + \overline{K_{Q,Q}} - 2\overline{K_{P,Q}} && (17)
\end{aligned}$$

HSIC

HSIC is nothing more than the MMD^2 between P_{XY} and $P_X P_Y$. Simples. The only tricky thing is keeping up with inconsistent notation over stacked expectations.

Let's start with a reminder of MMD:

$$MMD^2 = \mathbb{E}_P \mathbb{E}_{P'} k(x, x') + \mathbb{E}_Q \mathbb{E}_{Q'} k(y, y') - 2\mathbb{E}_P \mathbb{E}_Q k(x, y) \tag{18}$$

This means we can write out HSIC as:

$$\begin{aligned}
HSIC(P_{XY}, P_X P_Y) &= MMD^2(P_{XY}, P_X P_Y) && (19)\\
&= \|\mu_{P_{XY}} - \mu_{P_X P_Y}\|^2 && (20)\\
&= \mathbb{E}_{XY} \mathbb{E}_{X'Y'} k(x, x') l(y, y') + \mathbb{E}_X \mathbb{E}_Y \mathbb{E}_{X'} \mathbb{E}_{Y'} k(x, x') l(y, y') - && (21)\\
&\quad 2\mathbb{E}_{XY} \mathbb{E}_{X'} \mathbb{E}_{Y'} k(x, x') l(y, y') && (22)
\end{aligned}$$

Now we can separate expectations as much as possible:

$$\begin{aligned}
HSIC(P_{XY}, P_X P_Y) &= \mathbb{E}_{XY} \mathbb{E}_{X'Y'} k(x, x') l(y, y') + \mathbb{E}_X \mathbb{E}_{X'} k(x, x') \mathbb{E}_Y \mathbb{E}_{Y'} l(y, y') - && (23)\\
&\quad 2\mathbb{E}_{XY} \mathbb{E}_{X'} k(x, x') \mathbb{E}_{Y'} l(y, y') && (24)
\end{aligned}$$

Characteristic kernels

Characteristic kernels

A kernel is **characteristic** if there is a one-to-one (*'injective'*) mapping from probability distribution P to $\mu_P \in \mathcal{H}$. A characteristic kernel is a good choice for computing MMD, since it ensures that the result is a **metric** - that is, $MMD = 0$ iff $P = Q$.

Universal kernels

An kernel in an RKHS is **universal** if:

"k(x, x') is continuous, **X** is compact, and $\mathcal{F}$ dense in $C(X)$ with respect to L_∞"
Some translation:

- A *compact* set is closed (containing all its limit points), and bounded (having all its points lie within some fixed distance of each other).
- $C(\mathcal{X})$ is the set of continuous functions on $\mathcal{X}$.
- $\mathcal{F}$ dense in $C(\mathcal{X})$ implies that we can find different fs which are arbitrarily close to each other.

Universality implies that, for any given $\epsilon > 0$ and $f \in C(\mathcal{X})$, there exists $g \in \mathcal{F}$:

$$\|f - g\|_\infty \leq \epsilon \tag{25}$$

Showing characteristicness via FFT

Using a fourier decomposition, the mean embedding becomes a product of fourier series:

$$\mu_P(x) = \mathbb{E}_X k(X - x) = \int_{-\infty}^{\infty} k(x - t) dP(t) \tag{26}$$

$$\mu_{P,l} = \hat{k_l} \phi_{P,l} \tag{27}$$

where $\boldsymbol{\phi_P}$ is the fourier series for P.

The MMD becomes:

$$MMD(P, Q; \mathcal{F}) = \left\| \sum_{l=-\infty}^{\infty} [(\phi_{P,l} - \phi_{Q,l})\hat{k_l}]e^{ilx} \right\|_{\mathcal{F}} \tag{28}$$

It isn't too hard to see that if any $\hat{f_l}$ are zero, we might be able to get an MMD of zero without $P = Q$. In fact, we can state that a kernel is characteristic if the FFT is either

- non-zero everywhere; or
- zero only at countably many points

This means that if the FFT of the kernel has limited support, it will not be characteristic, and $MMD(P, Q; \mathcal{F})$ may be zero for non-equal P, Q (for example if P and Q differ only in the regions of frequency space where the kernel has no support).

The allowance for countably many zero-points only holds because P and Q are pdfs, therefore they must integrate to one. This means you can't get delta functions in their FFTs, since cosines do not decay.

Primal, dual, KKT

The primal function

Given a general optimisation problem:

$$\min_{x \in \mathbb{R}^n} f_0(x); \qquad \text{subject to} \quad f_i(x) \leq 0 \; i = 1, ..., m \tag{29}$$

$$h_i(x) = 0 \; i = 1, ..., m \tag{30}$$

If we want to put the constraints directly into the optimisation, it would look like this:

$$\min_{x \in \mathbb{R}^n} f_0(x) + \sum_i^m l_-(f_i(x)) + \sum_i^p l_0(h_i(x)) \tag{31}$$

where $l_-(u)$ is zero for $u \leq 0$ and ∞ otherwise, and l_0 is zero for $u = 0$ and ∞ otherwise. This isn't an easy thing to solve, since it's non-differentiable and not even continous. Instead let's solve something simple - the Lagrangian, which gives us an **lower bound** for the original problem.

$$L(x, \lambda, \nu) = f_0(x) + \sum_i^m \lambda_i f_i(x) + \sum_i^p \nu_i h_i(x) \tag{32}$$

where all $\lambda_i > 0$ (but ν_i are allowed to be negative).

The dual function

The Lagrange dual function is given by:

$$g(\lambda, \nu) := \inf_{x \in \mathcal{D}} L(x, \lambda, \nu)$$

A **dual feasible** pair (λ, ν) is a pair for which all $\lambda_i \geq 0$ and $(\lambda, \nu) \in \text{dom}(g)$.

For any $\lambda \geq 0$ and ν, the dual is a lower bound wherever the original constraints are met:

$$g(\lambda, \nu) \leq f_0(x); \quad \text{wherever} \tag{33}$$

$$f_i(x) \leq 0 \tag{34}$$

$$h_i(x) = 0 \tag{35}$$

Since the optimum $p*$ obeys the above constraints, we have:

$$g(\lambda, \nu) \leq f_0(x*) \; [= p*] \tag{36}$$

So, the Lagrangian function gives us a **lower bound** on the thing we're trying to minimise. So, we maximise the dual, **and** try to set things up so that the bound is strict.

$$\text{maximise:} \quad g(\lambda, \nu) \tag{37}$$

$$\text{subject to:} \quad \text{all } \lambda_i \geq 0 \tag{38}$$

Strong duality

It's only worth doing this if strong duality holds - that is, $g(\lambda*, \nu*) = f(x*)$. The best known sufficient condition for duality is:

- Primal problem is **convex**, i.e. all f_i are convex and h_i are affine.
- **Slater's condition** holds: there exists some *strictly* feasible point $\tilde{x} \in relint(\mathcal{D})$ for which all inequality constraints are strictly satisfied:

  $$f_i(\tilde{x}) < 0 \quad i = 1, ..., m \quad A\tilde{x} = b$$

Complementary slackness

A consequence of strong duality is complementary slackness:

$$\sum_{i}^{m} \lambda_i^* f_i(x^*) = 0 \tag{39}$$

If we remember that all $\lambda_i \geq 0$ and all $f_i(x*) \leq 0$, then:

$$\lambda_i^* > 0 \implies f_i(x^*) = 0 \tag{40}$$

$$f_i(x^*) < 0 \implies \lambda_i^* = 0 \tag{41}$$

Every inequality constraint becomes either strict (equality) or it doesn't contribute at all (corresponding $\lambda = 0$).

KKT conditions

For an unconstrained convex optimization problem, we know we are at the global minimum if the gradient is zero. The KKT conditions are the equivalent conditions for the global minimum of a constrained convex optimization problem.

Lagrangian stationarity

$$\nabla f_0(x*) + \sum_{i}^{m} \nabla f_i(x*) + \sum_{i}^{p} \nabla h_i(x*) = 0 \tag{42}$$

Primal feasibility

$$f_i(x*) \leq 0, \quad h_i(x*) = 0 \qquad \text{for all i} \tag{43}$$

Dual feasibility

$$\lambda_i \geq 0 \qquad \text{for all i} \tag{44}$$

Complementary slackness

$$\lambda_i f_i(x*) = 0 \qquad \text{for all i} \tag{45}$$