Kernel methods

MATH-412 - Statistical Machine Learning

Making models non-linear with a feature map

Idea : make non-linear transformation of the data first

- Quadratic map :

$$\boldsymbol{\phi}(\mathbf{x}) = (x_1, \ldots, x_p, x_1^2, \ldots, x_p^2,\ x_1 x_2,\ x_1 x_3, \ldots,\ x_{p-1} x_p)$$

- Fourier basis, spline basis, wavelet basis

Regularized empirical risk minimization with a mapping $\boldsymbol{\phi}$:

$$\min_{\boldsymbol{w}} \frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{w}^\top \boldsymbol{\phi}(x_i), y_i) + \lambda \|\boldsymbol{w}\|^2.$$

Representer theorem (simple version with the feature map)

Theorem (Kimmeldorf and Wahba, 1971)

Consider the optimization problem

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} L(\boldsymbol{w}^\top\boldsymbol{\phi}(x_1),\ldots,\boldsymbol{w}^\top\boldsymbol{\phi}(x_n)) + \lambda\|\boldsymbol{w}\|^2$$

Then any local minimum is of the form $\quad \boldsymbol{w} = \sum_{i=1}^{n} \alpha_i \boldsymbol{\phi}(x_i),$

for some vector $\boldsymbol{\alpha} \in \mathbb{R}^n$*. Interpretation :* $\boldsymbol{w} \in$ *span*$\big(\boldsymbol{\phi}(x_1),\ldots,\boldsymbol{\phi}(x_n)\big)$.

So that $\quad f_{\boldsymbol{w}}(x) = \boldsymbol{w}^\top\boldsymbol{\phi}(x) = \sum_{i=1}^{n} \alpha_i \langle \boldsymbol{\phi}(x_i), \boldsymbol{\phi}(x)\rangle = \sum_{i=1}^{n} \alpha_i K(x_i, x).$

Applying the representer theorem to the ERM problem

$$\min_{\boldsymbol{w}} \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{w}^\top \boldsymbol{\phi}(x_i), y_i) + \lambda\|\boldsymbol{w}\|^2.$$

By the theorem of Kimmeldorf and Wahba, $\boldsymbol{w}^\star = \sum_{j=1}^{n} \alpha_j^\star \boldsymbol{\phi}(x_j)$.

So replacing in the previous expression, we get

$$\min_{\boldsymbol{\alpha}} \frac{1}{n}\sum_{i=1}^{n} \ell\Big(\sum_{j=1}^{n} \alpha_j \langle \boldsymbol{\phi}_j(x_j), \boldsymbol{\phi}_i(x_i)\rangle, y_i\Big) + \lambda\Big\|\sum_{j=1}^{n} \alpha_j \boldsymbol{\phi}(x_j)\Big\|^2.$$

$$\min_{\boldsymbol{\alpha}} \frac{1}{n}\sum_{i=1}^{n} \ell\Big(\sum_{j=1}^{n} \alpha_j K_{ij}, y_i\Big) + \lambda \sum_{1\le i,j\le n} \alpha_i\alpha_j K_{ij},$$

with $K_{ij} = K(x_i, x_j) = \langle \boldsymbol{\phi}(x_i), \boldsymbol{\phi}(x_j)\rangle$ the values of a kernel function on pairs of input datapoints.

The ERM expressed with the kernel matrix

We rewrote $\min_{\boldsymbol{w}} \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{w}^\top \boldsymbol{\phi}(x_i), y_i) + \lambda \|\boldsymbol{w}\|^2$ as :

$$\min_{\boldsymbol{\alpha}} \frac{1}{n}\sum_{i=1}^{n} \ell\Big(\sum_{j=1}^{n} \alpha_j K_{ij}, y_i\Big) + \lambda \sum_{1 \le i,j \le n} \alpha_i \alpha_j K_{ij},$$

with $K_{ij} = K(x_i, x_j) = \langle \boldsymbol{\phi}(x_i), \boldsymbol{\phi}(x_j) \rangle$.

This can be rewritten in matrix vector form as

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \frac{1}{n}\sum_{i=1}^{n} \ell\Big(\mathbf{K}_{i \cdot} \boldsymbol{\alpha}, y_i\Big) + \lambda\, \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha}.$$

Furthermore to make a prediction, our predictor is computed as

$$\widehat{f}(x) = \boldsymbol{w}^{\star\top} \boldsymbol{\phi}(x) = \sum_{j=1}^{n} \alpha_j^\star\, K(x_j, x).$$

The kernel matrix when $\phi(\mathbf{x}) = \mathbf{x}$.

Based on the design matrix $\mathbf{X}$, two symmetric p.s.d. matrices are natural :

- the *empirical covariance matrix* (assuming $\boldsymbol{X}$ is centered)

$$\boxed{\widehat{\boldsymbol{\Sigma}} = \tfrac{1}{n}\boldsymbol{X}^\top \boldsymbol{X}}$$

$$\widehat{\boldsymbol{\Sigma}}_{k\ell} = \widehat{\mathrm{Cov}}(X^{(k)}, X^{(\ell)}) = \left\langle \frac{1}{\sqrt{n}}\mathbf{x}^k, \frac{1}{\sqrt{n}}\mathbf{x}^\ell \right\rangle$$

- the *kernel matrix* or *Gram matrix*

$$\boxed{\boldsymbol{K} = \boldsymbol{X}\boldsymbol{X}^\top}$$

$$\boldsymbol{K}_{ij} = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$$

$\boldsymbol{K}$ is simply the matrix of all dot products. $\boldsymbol{K}$ encodes information about the data vectors $\mathbf{x}_i = \boldsymbol{X}_{i\cdot}^\top$ while $\widehat{\boldsymbol{\Sigma}}$ encodes information about the variables $\mathbf{x}^k = \boldsymbol{X}_{\cdot k}$

Properties of the kernel matrix when $\phi(\mathbf{x}) = \mathbf{x}$.

The kernel matrix contains a lot of information about the data :

- It contains the information about all the distances between all pairs of data points (and between each data points and the origin). Indeed,

$$\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 = \boldsymbol{K}_{ii} - 2\boldsymbol{K}_{ij} + \boldsymbol{K}_{jj}.$$

- As a consequence, any factorization of the matrix $\boldsymbol{K}$ of the form

$$\boldsymbol{K} = \mathbf{R}\mathbf{R}^\top,$$

  retrieves a representation of the data up to an isometry. This can be obtained for example by the Cholesky decomposition.

Why is this useful ?

Dot products in feature space

Let $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$ and $\boldsymbol{\phi}(\mathbf{x}) = (x_1, x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2)^\top$.

$$\begin{aligned}
\langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\boldsymbol{y}) \rangle &= x_1y_1 + x_2y_2 + x_1^2y_1^2 + x_2^2y_2^2 + 2x_1x_2y_1y_2 \\
&= x_1y_1 + x_2y_2 + (x_1y_1)^2 + (x_2y_2)^2 + 2(x_1y_1)(x_2y_2) \\
&= \langle \mathbf{x}, \boldsymbol{y} \rangle + \langle \mathbf{x}, \boldsymbol{y} \rangle^2
\end{aligned}$$

For $\boldsymbol{w} = (0, 0, 1, 1, 0)^\top$, $\quad \boldsymbol{w}^\top \boldsymbol{\phi}(\mathbf{x}) - 1 \le 0 \quad \Leftrightarrow \quad \|\mathbf{x}\|^2 \le 1.$

Linear separators in $\mathbb{R}^5$ correspond to conic separators in $\mathbb{R}^2$.
`https://www.youtube.com/watch?v=Q7vT0--5VII`

Let $\mathbf{x} = (x_1, \ldots, x_p) \in \mathbb{R}^p$ and

$$\boldsymbol{\phi}(\mathbf{x}) = (x_1, \ldots, x_p, x_1^2, \ldots, x_p^2, \sqrt{2}x_1x_2, \ldots, \sqrt{2}x_ix_j, \ldots \sqrt{2}x_{p-1}x_p)^\top.$$

Still have

$$\langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\boldsymbol{y}) \rangle = \langle \mathbf{x}, \boldsymbol{y} \rangle + \langle \mathbf{x}, \boldsymbol{y} \rangle^2$$

But explicit mapping too expensive to compute : $\boldsymbol{\phi}(\mathbf{x}) \in \mathbb{R}^{p+p(p+1)/2}$.

Which abstract space is a good predictor space ?

Require that

(1) the space should be a Hilbert space $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$

(2) $\forall x \in \mathcal{X}$, the *evaluation functional* $f \mapsto f(x)$ is *continuous* from $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ to $\mathbb{R}$.

- This is equivalent to requiring that for a given $x \in \mathcal{X}$:

  if $\|f - g\|_{\mathcal{H}}$ is small then $|f(x) - g(x)|$ should be small.

- The motivation is that we would like that

$$\left(\|\widehat{f}_n - f^*\|_{\mathcal{H}} \to 0 \right) \Rightarrow \left(\widehat{f}_n(x) \to f^*(x) \right)$$

Riesz Representation Theorem

Let $\mathcal{H}$ be a Hilbert space, and $\psi : \mathcal{H} \to \mathbb{R}$ be a *continuous* linear form, then there exists $h_\psi \in \mathcal{H}$ such that

$$\forall f \in \mathcal{H},\ \psi(f) = \langle h_\psi, f \rangle_{\mathcal{H}} .$$

Under (1) and (2) by this theorem, there must exist an element $h_x \in \mathcal{H}$ such that

$$\forall f \in \mathcal{H}, \quad f(x) = \langle h_x, f \rangle_{\mathcal{H}} .$$

Reproducing Kernel Hilbert Space

So if $\mathcal{H}$ is a Hilbert space of functions in which the *evaluation functionals* are continuous, then by the Riesz representation theorem, there must exist an element $h_x \in \mathcal{H}$ such that

$$\forall f \in \mathcal{H}, \quad f(x) = \langle h_x, f \rangle_{\mathcal{H}}.$$

But then by definition $h_y(x) = \langle h_x, h_y \rangle_{\mathcal{H}} = h_x(y)$.
Define the *reproducing kernel* as the function

$$\begin{aligned} K : \mathcal{X} \times \mathcal{X} &\to \mathbb{R} \\ (x, y) &\mapsto \langle h_x, h_y \rangle_{\mathcal{H}}. \end{aligned}$$

By definition $h_x(\cdot) = K(x, \cdot)$ so that

$$f(x) = \langle K(x,\cdot), f \rangle_{\mathcal{H}} \quad \text{and} \quad \langle K(x,\cdot), K(y,\cdot) \rangle_{\mathcal{H}} = K(x, y).$$

A space with these properties is called a *reproducing kernel Hilbert space* (RKHS).

Positive definite functions

$$(x, y) \mapsto K(x, y)$$

is a *positive definite function* if the matrix constructed as

$$\boldsymbol{K} = \begin{bmatrix} K(x_1, x_1) & \ldots, & \ldots & K(x_1, x_n) \\ K(x_2, x_1) & \ldots, & \ldots & K(x_2, x_n) \\ \vdots & & & \vdots \\ K(x_n, x_1) & \ldots, & \ldots & K(x_n, x_n) \end{bmatrix}$$

is a positive semi-definite matrix

$$\text{i.e.,} \quad \forall \boldsymbol{\alpha} \in \mathbb{R}^n, \quad \boldsymbol{\alpha}^\top \boldsymbol{K} \boldsymbol{\alpha} \geq 0,$$

for any choice of $x_1, \ldots, x_n$ and any value of n.

A reproducing kernel is a positive definite function

Proposition

A reproducing kernel is a positive definite function.

Proof of the claim The reproducing kernel is necessarily a *symmetric positive definite function* since for all $x_1, \ldots, x_n \in \mathcal{X}$, we have $\langle K(x_i, \cdot), K(x_j, \cdot)\rangle_{\mathcal{H}} = K(x_i, x_j)$, and thus for all $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$.

$$0 \leq \left\langle \textstyle\sum_i \alpha_i K(x_i, \cdot), \sum_j \alpha_j K(x_j, \cdot) \right\rangle_{\mathcal{H}} = \sum_{i,j} \alpha_i \alpha_j K(x_i, x_j),$$

with equality if and only if $\alpha_i = 0$ for all i.

Converse ?

Yes, any symmetric positive definite function is the reproducing kernel of a RKHS (Aronszajn, 1950).

Moore-Aronszajn theorem

Theorem

A symmetric function K on $\mathcal{X}$ is positive definite if and only if there exists a Hilbert space $\mathcal{H}$ and a mapping

$$\begin{aligned} \phi : \mathcal{X} &\to \mathcal{H} \\ x &\mapsto \phi(x) \end{aligned}$$

such that $K(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$.

- In fact, this mapping is $\phi(x) = h_x$
- Such symmetric p.d. functions are often called *Mercer kernels*.
- We will not show this theorem in this course.

Common RKHSes for $\mathcal{X} = \mathbb{R}^p$

Linear kernel

- $K(\mathbf{x}, \boldsymbol{y}) = \mathbf{x}^\top \boldsymbol{y}$
- $\mathcal{H} = \{f_{\boldsymbol{w}} : \mathbf{x} \mapsto \boldsymbol{w}^\top \mathbf{x} \mid \boldsymbol{w} \in \mathbb{R}^p\}$
- $\|f_{\boldsymbol{w}}\|_{\mathcal{H}} = \|\boldsymbol{w}\|_2$

Polynomial kernel

- $K_h(\mathbf{x}, \boldsymbol{y}) = (\gamma + \mathbf{x}^\top \boldsymbol{y})^d$
- $\mathcal{H}$

Radial Basis Function kernel (RBF)

- $K_h(\mathbf{x}, \boldsymbol{y}) = \exp\left(-\frac{\|\mathbf{x}-\boldsymbol{y}\|_2^2}{2h}\right)$
- $\mathcal{H} =$ Gaussian RKHS

Representer theorem

Theorem (Kimmeldorf and Wahba, 1971)

Consider the optimization problem

$$\min_{f \in \mathcal{H}} L(f(x_1), \ldots, f(x_n)) + \lambda \|f\|_{\mathcal{H}}^2$$

Then any local minimum is of the form $\quad f = \sum_{i=1}^{n} \alpha_i K(x_i, \cdot),$

where K is the reproducing kernel associated with the RKHS $\mathcal{H}$ and $\boldsymbol{\alpha}$ is a vector in $\mathbb{R}^n$.

Proof Indeed, let f be a local minimum and consider the subspace

$$\mathcal{S} = \{g \mid g = \sum_{i=1}^{n} \alpha_i K(x_i, \cdot), \quad \boldsymbol{\alpha} \in \mathbb{R}^n\}.$$

Representer theorem

We can decompose $f = f_{/\!/} + f_{\perp}$ with $f_{/\!/} = \mathsf{Proj}_{\mathcal{S}}(f)$. We then have

$$f_{\perp}(x_i) = \langle f_{\perp}, K(x_i, \cdot)\rangle_{\mathcal{H}} = 0 \quad \text{and} \quad \langle f_{\perp}, f_{/\!/}\rangle_{\mathcal{H}} = 0.$$

Thus

$$\begin{aligned} & L(f(x_1), \ldots, f(x_n)) + \lambda \|f\|_{\mathcal{H}}^2 \\ = \;& L(f_{/\!/}(x_1), \ldots, f_{/\!/}(x_n)) + \lambda \left(\|f_{/\!/}\|_{\mathcal{H}}^2 + 2\langle f_{\perp}, f_{/\!/}\rangle_{\mathcal{H}} + \|f_{\perp}\|_{\mathcal{H}}^2\right) \\ = \;& L(f_{/\!/}(x_1), \ldots, f_{/\!/}(x_n)) + \lambda \|f_{/\!/}\|_{\mathcal{H}}^2 + \lambda \|f_{\perp}\|_{\mathcal{H}}^2 \end{aligned}$$

So that we must have $f_{\perp} = 0$.

Regularized ERM for f in a RKHS

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i) \; + \; \lambda \, \|f\|_{\mathcal{H}}^2 \tag{P}$$

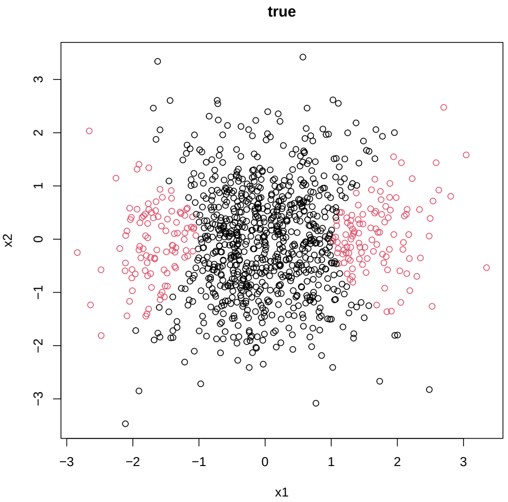
By the representer theorem, the solution of the regularized empirical risk minimization problem lies in the subspace of $\mathcal{H}$ generated by the point x_i, i.e.,

$$f^* = \sum_{j=1}^{n} \alpha_j K(x_j, \cdot) \quad \text{for some } \alpha_i \in \mathbb{R}. \tag{R}$$

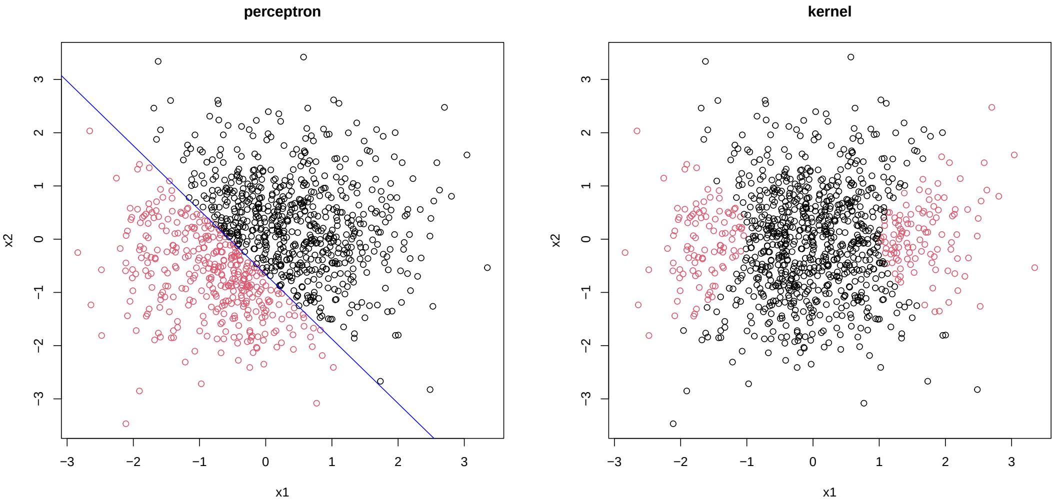
The solution of (P) is therefore of the form (R) with $\alpha \in \mathbb{R}^n$ the solution of

$$\min_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} \ell\Big(\sum_{j=1}^{n} \alpha_j K(x_j, x_i), y_i \Big) \; + \; \lambda \sum_{1 \leq i,j \leq n} \alpha_i \alpha_j K(x_i, x_j).$$

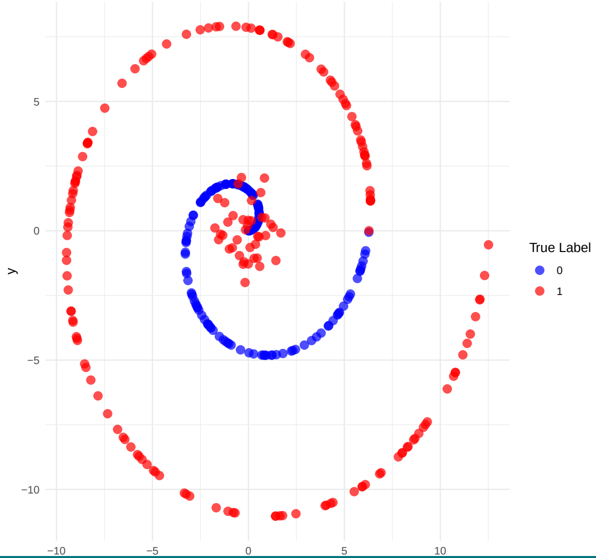
Hyperbola example



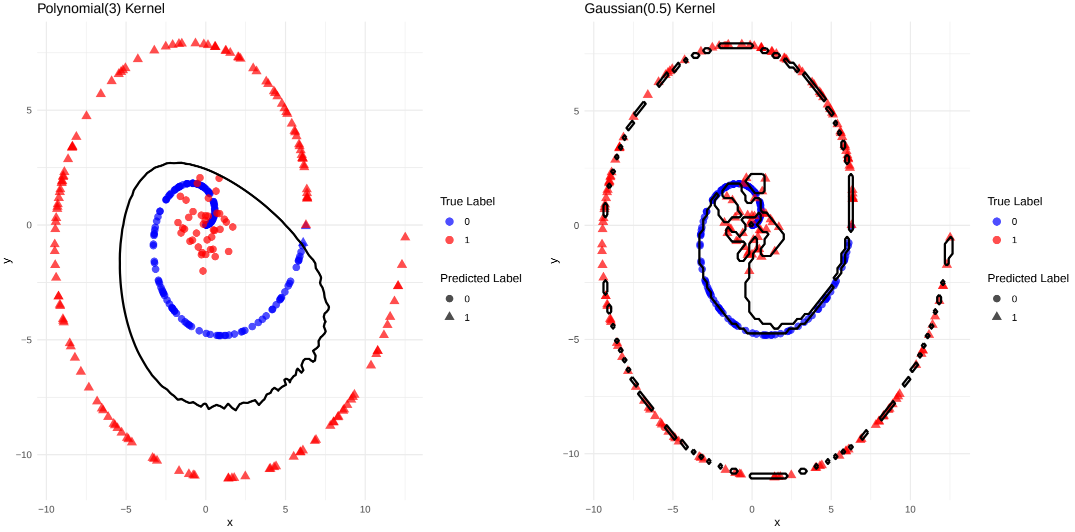
Hyperbola example



"Roll" example



"Roll" example



$\|f\|_{\mathcal{H}}$ measures the smoothness of the function f

Indeed :

$$|f(x) - f(x')| = |\left\langle f, K(x,\cdot) - K(x',\cdot)\right\rangle_{\mathcal{H}}| \leq \|f\|_{\mathcal{H}}\|K(x,\cdot) - K(x',\cdot)\|_{\mathcal{H}}$$

$\rightarrow$ f is Lipschitz with respect to the ℓ^2 distance induced by the RKHS

$$d(x,x') = \|K(x,\cdot) - K(x',\cdot)\|_{\mathcal{H}} = \sqrt{K(x,x) + K(x',x') - 2K(x,x')}$$

$\rightarrow$ $\|f\|_{\mathcal{H}}$ is the Lipschitz constant

Kernel combinations

Assume K, K_1 and K_2 are positive definite functions,
then the following are still p.d. kernel functions :

Sum of kernels : For $\alpha_1, \alpha_2 > 0,\ \tilde{K}(x,y) = \alpha_1 K_1(x,y) + \alpha_2 K(x,y)$

Limits of kernels : $K(x,y) = \lim_{n\to\infty} K_n(x,y)$

Pointwise product : $\tilde{K}(x,y) = K_1(x,y)\, K_2(x,y)$

Pairwise kernel : $\tilde{K}(x,y) = \sum_{z\in\mathcal{Z}} K(x,z)K(z,y)$

Normalized kernel : $\tilde{K}(x,y) = \dfrac{K(x,y)}{\sqrt{K(x,x)K(y,y)}} = \cos\angle\big(\boldsymbol{\phi}(x), \boldsymbol{\phi}(y)\big)$

In terms of kernel matrices

Pointwise product : $\tilde{\boldsymbol{K}} = \boldsymbol{K}_1 \odot \boldsymbol{K}_2$ (Hadamard product)

Pairwise kernel : $\tilde{\boldsymbol{K}} = \boldsymbol{K}^2$ (Matrix product)

Scaling...

The kernelized form $$\min_{\boldsymbol{\alpha}} \frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{\alpha}^\top \boldsymbol{k}_i, y_i) + \frac{\lambda}{2} \boldsymbol{\alpha}^\top \boldsymbol{K} \boldsymbol{\alpha}$$

requires to compute $\boldsymbol{K} \in \mathbb{R}^{n \times n}$.

- The cost of working with kernels quadratic in n.
- ... unless $\boldsymbol{K}$ is low rank, e.g. for the linear kernel
- This is a price to pay to work in very high/infinite dimensional spaces
- It is however possible to
  - compute low rank approximations to $\boldsymbol{K}$ using *Nyström's method* (Williams and Seeger, 2001; Gittens and Mahoney, 2016) or
  - use greedy approximation schemes (Smola et al., 2000)
  - compute directly lower/finite dimensional approximation to the feature map using *random features expansions* (Rahimi and Recht, 2007; Bach, 2017; Yang et al., 2017)

Kernel ridge regression

$$\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\big(y_i - f(x_i)\big)^2 + \lambda\|f\|_{\mathcal{H}}^2$$

$$\min_{f\in\mathcal{H}} \frac{1}{n}\|\boldsymbol{y}-\boldsymbol{f}\|_2^2 + \lambda\|f\|_{\mathcal{H}}^2 \qquad \text{with} \quad \boldsymbol{f} = (f(x_1),\dots,f(x_n)).$$

By the representer property $\widehat{f}(x) = \sum_{i=1}^{n}\alpha_i K(x_i,x)$, so that $\frac{1}{2}\|\boldsymbol{f}-\boldsymbol{y}\|_2^2 = \frac{1}{2}\|\boldsymbol{K\alpha}-\boldsymbol{y}\|_2^2$.

The regularized empirical risk is $\frac{1}{2n}\|\boldsymbol{K\alpha}-\boldsymbol{y}\|_2^2 + \frac{\lambda}{2}\boldsymbol{\alpha}^\top\boldsymbol{K\alpha}$

and the minimizers are of the form $\boldsymbol{\alpha}^\star + \boldsymbol{h}$ with $\boldsymbol{\alpha}^\star = (\boldsymbol{K}+\lambda n\boldsymbol{I})^{-1}\boldsymbol{y}$, and $\boldsymbol{h}\in\mathrm{Ker}(\boldsymbol{K})$.

Finally $\widehat{f}(x) = \sum_{i=1}^{n}\alpha_i^\star K(x_i,x)$ because $\sum_{i=1}^{n} h_i K(x_i,\cdot) = 0,\ \forall \boldsymbol{h}\in\mathrm{Ker}(\boldsymbol{K})$.

Convolution vs Mercer kernels

In this course we encountered two type of kernels

Convolution kernels

Used for density estimation and by the Nadaraya-Watson estimator

$$K_\delta(x-y) = h\left(\frac{\|x-y\|}{\delta}\right)$$

- e.g. Epanechnikov, tricube or Gaussian kernel

Mercer kernels

... or simply positive definite kernel functions, which by Aronszajn's theorem provide the inner product of a RKHS

$$K(x,y)$$

- e.g. linear, polynomial, Laplace, Gaussian kernel and more

- Some are actually both

Summary

- Every positive definite function (Mercer kernel) is associated with a RKHS
- Regularized ERM can be kernelized (e.g. ridge regression)
- Many other algorithms in ML can be kernelized (e.g. kernel PCA)
- The representer theorem of Kimmeldorf and Wahba (1971) guarantees that a large class of optimization problems in RKHS can be reformulated as a finite-dimensonal optimization problem.
- Using kernels directly has complexity n^2 but there are efficient approximation schemes (Nyström, random feature expansions)
- Mercel kernels should not be confused with convolution kernels used for density estimation and by Nadaraya-Watson estimators.

References I

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3) :337–404.

Bach, F. (2017). On the equivalence between kernel quadrature rules and random feature expansions. *The Journal of Machine Learning Research*, 18(1) :714–751.

Gittens, A. and Mahoney, M. W. (2016). Revisiting the Nyström method for improved large-scale machine learning. *The Journal of Machine Learning Research*, 17(1) :3977–4041.

Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pages 1177–1184. Curran Associates Inc.

Smola, A., Schölkopf, B., and Langley, P. (2000). Sparse greedy matrix approximation for machine learning. In *17th International Conference on Machine Learning, Stanford, 2000*, pages 911–911.

Williams, C. K. and Seeger, M. (2001). Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, pages 682–688.

Yang, Y., Pilanci, M., and Wainwright, M. (2017). Randomized sketches for kernels : Fast and optimal nonparametric regression. *The Annals of Statistics*, 45(3) :991–1023.