2. Kernels and Regularisation

COMP0078: Supervised Learning

Mark Herbster

7 October 2019

University College London
Department of Computer Science

SL-kernreg19v7

Today's Plan

Overview

- Inner product space review
- Convexity review
- Ridge Regression
- Basis Functions (Explicit Feature Maps)
- Kernel Functions (Implicit Feature Maps)

Overview

- We show how a linear method such as least squares may be lifted to a (potentially) higher dimensional space to provide a nonlinear regression.
- We consider both explicit and implicit feature maps
- A feature map is simply a function that maps the “inputs” into a new space.
- Thus the original method is now nonlinear in original “inputs” but linear in the “mapped inputs”
- Explicit feature maps are often known as the Method of Basis Functions
- Implicit feature maps are often known as the (reproducing) “Kernel Trick”

Review: Inner Product Space

Vector Space

Vector space over the reals

The triple $(\mathrm{X}, +, *)$ defines a vector space where X is a set and $+ : \mathrm{X} \times \mathrm{X} \to \mathrm{X}$ is vector addition and $* : \mathbb{R} \times \mathrm{X} \to \mathrm{X}$ is scalar multiplication with the abbreviation $\mathrm{ax} := \mathrm{a} * \mathrm{x}$. For which the following properties hold.

1. $\mathrm{x} + \mathrm{y} = \mathrm{y} + \mathrm{x}$
2. $(\mathrm{x} + \mathrm{y}) + \mathrm{z} = \mathrm{x} + (\mathrm{y} + \mathrm{z})$
3. There exists $0 \in \mathrm{X}$ such that for all $\mathrm{x} \in \mathrm{X}$ then $\mathrm{x} + 0 = \mathrm{x}$
4. $\gamma(\mathrm{x} + \mathrm{y}) = \gamma\mathrm{x} + \gamma\mathrm{y}$
5. $(\gamma + \mu)\mathrm{x} = \gamma\mathrm{x} + \mu\mathrm{x}$
6. $\gamma(\mu\mathrm{x}) = (\gamma\mu)\mathrm{x}$
7. $0\mathrm{x} = 0$ and $1\mathrm{x} = \mathrm{x}$

Notes

1. $\gamma + \mu$ and $\gamma\mu$ are the usual scalar addition and multiplication over $\mathbb{R}$
2. A vector space can be defined over other fields than the reals for example arithmetic mod-2. I.e $\mathrm{X} = \{0, 1\}$ and the scalar set is just $\{0, 1\}$.

Normed Space

A function $\|\cdot\| : X \to \mathbb{R}$ is a norm on a vector space if

1. $\|x\| = 0 \Leftrightarrow x = 0$
2. $\|x + y\| \leq \|x\| + \|y\|$
3. $\|\gamma x\| = |\gamma|\|x\|$

A norm defines a metric (distance) between vectors in the space via $d(x, y) := \|x - y\|$. (check that the triangle inequality holds $d(x, z) \leq (x, y) + d(y, z)$).

Examples

1. $\|x\|_p := (\sum_{i=1}^{n} |x_i^p|)^{\frac{1}{p}}$ where $x \in \mathbb{R}^n$ and $p \in [1, \infty)$.
2. $\|x\|_M := \sqrt{x^\top M x}$ where $x \in \mathbb{R}^n$ and M is an $n \times n$ symmetric positive definite matrix.

Definition

A real-valued symmetric $n \times n$ square matrix M is positive definite iff $x^\top M x > 0 \ \ (\forall x \in \mathbb{R}^n \setminus \{0\})$.

Normed Space – Intuitions

A norm generalises the notion of euclidean magnitude $\|\cdot\|_2$. The p-norm plays a particularly important role in machine learning where kernel methods may be seen as a generalisation of the case $p = 2$. The case $p = 1$ is particularly important when learning sparse predictors.

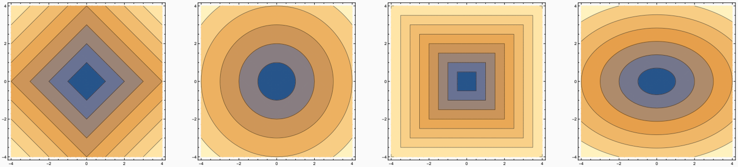


Figure 1: Level sets of $\|\cdot\|_1$, $\|\cdot\|_2$, $\|\cdot\|_\infty$, $\|\cdot\|_{\begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}}$

Where $\|x\|_\infty := \lim_{p\to\infty} \|x\|_p = \max_{i\in[n]} x_i$.

Level set of f at α is $\{x : f(x) \leq \alpha\}$.

Question : What metric does $\|\cdot\|_1$ correspond to in $\mathbb{R}^2$?

Real Inner product space

The quadruple $(\mathrm{X}, +, *, \langle\rangle)$ defines an inner product space where $(\mathrm{X}, +, *)$ is a vector space and $\langle\rangle : \mathrm{X} \times \mathrm{X} \to \mathbb{R}$ is an inner product s.t.

1. $\langle \mathrm{x}, \mathrm{x} \rangle = 0 \Leftrightarrow \mathrm{x} = 0$ and $\langle \mathrm{x}, \mathrm{x} \rangle \geq 0$
2. $\langle \mathrm{x}, \mathrm{y} \rangle = \langle \mathrm{y}, \mathrm{x} \rangle$
3. $\langle \gamma \mathrm{x}, \mathrm{y} \rangle = \gamma \langle \mathrm{x}, \mathrm{y} \rangle$ and $\langle \mathrm{x} + \mathrm{y}, \mathrm{z} \rangle = \langle \mathrm{x}, \mathrm{z} \rangle + \langle \mathrm{y}, \mathrm{z} \rangle$

The inner product induces a norm via $\|\mathrm{x}\| := \sqrt{\langle \mathrm{x}, \mathrm{x} \rangle}$. The geometric interpretation of the inner product so that if $\theta := \cos^{-1}(\frac{\langle \mathrm{x}, \mathrm{y} \rangle}{\|\mathrm{x}\| \|\mathrm{y}\|})$ then θ is the angle between the vectors.

Examples

1. The Euclidean inner product $\langle \mathrm{x}, \mathrm{y} \rangle := \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x}_\mathrm{i} \mathrm{y}_\mathrm{i}$
2. More generally $\langle \mathrm{x}, \mathrm{y} \rangle_\mathrm{M} := \mathrm{x}^\top \mathrm{M} \mathrm{y}$ where $\mathrm{x}, \mathrm{y} \in \mathbb{R}^\mathrm{n}$ and M is an $\mathrm{n} \times \mathrm{n}$ symmetric positive definite matrix. (Exercise: Check)

Note: Many different notations are used for inner product including $\langle \mathrm{x}, \mathrm{y} \rangle = (\mathrm{x}, \mathrm{y}) = \mathrm{x} \cdot \mathrm{y}$

A more complicated example – 1

Here the space is set of all functions from $[0, \infty) \to \mathbb{R}$ with the following three properties.

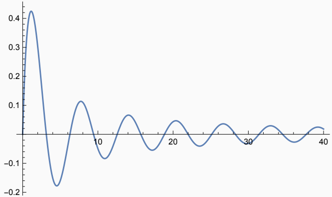
1. $f(0) = 0$
2. f is absolutely continuous (hence $f(b) - f(a) = \int_a^b f'(x)dx$)
3. $\int_0^\infty [f'(x)]^2 \, dx < \infty$

Addition '+' is the usual addition of functions similarly multiplication by a scalar. The dot product is defined as

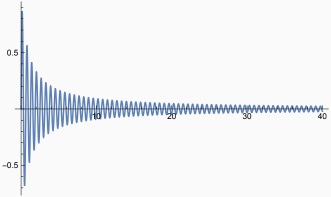
$$\langle f, g \rangle := \int_0^\infty f'(x)g'(x)dx$$

Exercise : Argue that this is an inner product space.

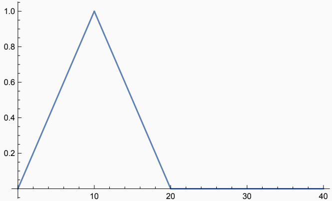
A more complicated example – 2



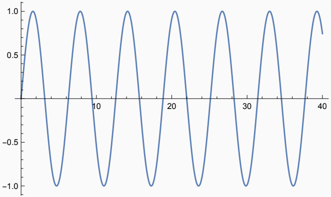
(a) $f(x) = \frac{\sin(x)}{x+1}$; $\|f\| \approx 0.68$



(b) $f(x) = \frac{\sin(10x)}{x+1}$; $\|f\| \approx 7.06$



(c) $f(x) = .1x[x \leq 10] + (2 - .1x)[10 < x \leq 20]$; $\|f\| = \frac{1}{\sqrt{5}}$



(d) $f(x) = \sin(x)$; Not in space (why?)

Figure 2: Example functions and their norms

Review: Convexity

Convexity

Definition

A function $f : \mathcal{X} \to \mathbb{R}$ is convex iff $\forall p, q \in \mathcal{X}$ and $\alpha \in (0, 1)$ we have

$$f(\alpha p + (1 - \alpha)q) \leq \alpha f(p) + (1 - \alpha)f(q)$$

A function f is concave if $-f$ is convex. A function is additionally strictly convex if we can replace"$\leq$" with "$<$".

Definition

A set $\mathcal{X}$ is convex if $p, q \in \mathcal{X} \implies (\alpha p + (1 - \alpha)q) \in \mathcal{X}$ for $\forall \alpha \in (0, 1)$.

Some results

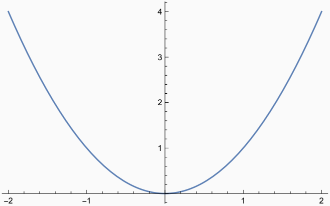
1. If f and g are convex then $f + g$ is convex.
2. If f is convex and g is affine (linear + a constant) then $f(g(\cdot))$ is convex.
3. Suppose M is a symmetric matrix then M is a PSD matrix iff $f(x) = x^\top Mx$ is convex.
4. A level set of a convex function is convex.

Checking for convexity

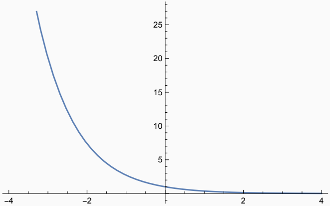
Some Results

1. For $f : (a, b) \to \mathbb{R}$ if f' is increasing then f is convex.
2. For $f : (a, b) \to \mathbb{R}$ if $f'' \geq 0$ then f is convex.
3. For $f : \mathcal{X} \to \mathbb{R}$ if $\mathcal{X} \subseteq \mathbb{R}^n$, $\mathcal{X}$ is a convex set and if the Hessian of f evaluated at x denoted $H(x)$ is positive semidefinite for all $x \in \mathcal{X}$ then f is convex.

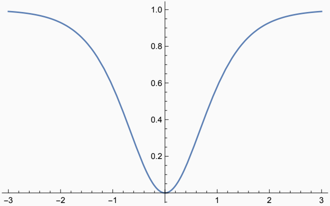
Examples



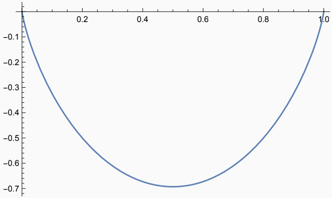
(a) $f(x) = x^2$ "Quadratic"



(b) $f(x) = e^{-x}$ "Exponential decay"



(c) $f(x) = \tanh(x)^2$ "Neural" (not convex)



(d) $f(x) = x \log(x) + (1 - x) \log(1 - x)$; "negative entropy"

Why Convexity?

- At the heart of many ML algorithms is an optimisation problem.
- For example, minimize error, minimise regularised error, minimise “energy”, maximise likelihood.
- If the (unconstrained) optimisation prob. is convex and there is a minima[1] then methods gradient-based methods can be applied to smooth problems.
- On the other hand. The existence of “many” minima each associated with a distinct function suggests that for many optimisation approaches there will be a high “variance”.
- Take-away : everything else equal convex objectives in ML are simpler to work with.

[1] Or more technically a minimal surface.

Ridge Regression

Linear interpolation

Problem

We wish to find a function $f(x) = w^\top x$ which best interpolates a data set $S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \subseteq \mathbb{R}^n \times \mathbb{R}$

- If the data have been generated in the form $(x, f(x))$, the vectors x_i are linearly independent and $m = n$ then there is a unique interpolant whose parameter w solves

$$Xw = y$$

  where, recall, $y = (y_1, \ldots, y_m)^\top$ and $X = [x_1, \ldots, x_m]^\top$
- Otherwise, this problem is ill-posed

Ill-posed problems

A problem is well-posed – in the sense of Hadamard (1902) – if

(1) a solution exists

(2) the solution is unique

(3) the solution depends continuously on the data

A problem is ill-posed if it is not well-posed

Learning problems are in general ill-posed (usually because of (2))

Regularization theory provides a general framework to solve ill-posed problems

Ridge Regression

Motivation:

1. Give a set of k hypothesis classes $\{\mathcal{H}_r\}_{r \in \mathbb{N}_k}$ we can choose an appropriate hypothesis class with cross-validation
2. An alternative compatible with linear regression is to choose a single "complex" hypothesis class and then modify the error function by adding a "complexity" term which penaltizes complex functions
3. This is known as regularization
4. Cross-validation may still be needed to set the regularization parameter (see below) and other parameters defining the complexity term

We minimize the regularized (penalized) empirical error

$$\mathcal{E}_{\mathrm{emp}_\lambda}(\mathrm{w}) := \sum_{i=1}^{m}(y_i - \mathrm{w}^\top x_i)^2 + \lambda \sum_{\ell=1}^{n} \mathrm{w}_\ell^2 \equiv (y - X\mathrm{w})^\top (y - X\mathrm{w}) + \lambda \|\mathrm{w}\|_2^2$$

The positive parameter λ defines a trade-off between the error on the data and the norm of the vector w (degree of regularization)

Setting $\nabla \mathcal{E}_{\mathrm{emp}_\lambda}(\mathrm{w}) = 0$, we obtain the modified normal equations

$$-2X^\top(y - X\mathrm{w}) + 2\lambda \mathrm{w} = 0 \tag{1}$$

whose solution (called regularized solution) is

$$\mathrm{w} = (X^\top X + \lambda I_n)^{-1} X^\top y \tag{2}$$

Dual representation

It can be shown that the regularized solution can be written as

$$\mathrm{w} = \sum_{\mathrm{i}=1}^{\mathrm{m}} \alpha_{\mathrm{i}} \mathrm{x}_{\mathrm{i}} \quad \Rightarrow \quad \mathrm{f}(\mathrm{x}) = \sum_{\mathrm{i}=1}^{\mathrm{m}} \alpha_{\mathrm{i}} \mathrm{x}_{\mathrm{i}}^{\top} \mathrm{x} \qquad (*)$$

where the vector of parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{\mathrm{m}})^{\top}$ is given by

$$\boldsymbol{\alpha} = (\mathrm{X}\mathrm{X}^{\top} + \lambda \mathrm{I}_{\mathrm{m}})^{-1} \mathrm{y} \tag{3}$$

• Function representations: we call the functional form (or representation) $\mathrm{f}(\mathrm{x}) = \mathrm{w}^{\top}\mathrm{x}$ the primal form and (*) the dual form (or representation)

The dual form is computationally convenient when $\mathrm{n} > \mathrm{m}$

Dual representation (continued – 1)

We rewrite eq.(1) as

$$\mathrm{w} = \frac{\mathrm{X}^\top(\mathrm{y} - \mathrm{X}\mathrm{w})}{\lambda}$$

Thus we have

$$\mathrm{w} = \sum_{\mathrm{i}=1}^{\mathrm{m}} \alpha_{\mathrm{i}} \mathrm{x}_{\mathrm{i}} \tag{4}$$

with

$$\alpha_{\mathrm{i}} = \frac{\mathrm{y}_{\mathrm{i}} - \mathrm{w}^\top \mathrm{x}_{\mathrm{i}}}{\lambda} \tag{5}$$

Consequently, we have that

$$\mathrm{w}^\top \mathrm{x} = \sum_{\mathrm{i}=1}^{\mathrm{m}} \alpha_{\mathrm{i}} \mathrm{x}_{\mathrm{i}}^\top \mathrm{x}$$

proving eq.(*).

Dual representation (continued – 2)

Plugging eq.(4) in eq.(5) we obtain

$$\alpha_i = \frac{y_i - (\sum_{j=1}^{m} \alpha_j x_j)^\top x_i}{\lambda}$$

Thus (with defining $\delta_{ij} = 1$ if $i = j$ and as 0 otherwise)

$$y_i = (\sum_{j=1}^{m} \alpha_j x_j)^\top x_i + \lambda \alpha_i$$

$$y_i = \sum_{j=1}^{m} (\alpha_j {x_j}^\top x_i + \alpha_j \lambda \delta_{ij})$$

$$y_i = \sum_{j=1}^{m} ({x_j}^\top x_i + \lambda \delta_{ij}) \alpha_j$$

Hence $(XX^\top + \lambda I_m)\boldsymbol{\alpha} = y$ from which eq.(3) follows.

Computational Considerations

Training time:

- Solving for w in the primal form requires $O(mn^2 + n^3)$ operations while solving for $\boldsymbol{\alpha}$ in the dual form requires $O(nm^2 + m^3)$ (see $(*)$) operations

If $m \ll n$ it is more efficient to use the dual representation Running

(testing) time:

- Computing $f(x)$ on a test vector x in the primal form requires $O(n)$ operations while the dual form (see $(*)$) requires $O(mn)$ operations

Sparse representation

We can benefit even further in the dual representation if the inputs are sparse!

Example

Suppose each input $x \in \mathbb{R}^n$ has most of its components equal to zero (e.g., consider images where most pixels are 'black' or text documents represented as 'bag of words')

- If k denotes the number of nonzero components of the input then computing $x^\top t$ requires at most $O(k)$ operations.
  How do we do this?
- If $km \ll n$ (which implies $m, k \ll n$) the dual representation requires $O(km^2 + m^3)$ computations for training and $O(mk)$ for testing

Basis Functions

Basis Functions – Explicit Feature Map

The above ideas can naturally be generalized to nonlinear function regression

By a feature map we mean a function $\boldsymbol{\phi} : \mathbb{R}^n \to \mathbb{R}^N$

$$\boldsymbol{\phi}(x) = (\phi_1(x), \ldots, \phi_N(x))^\top, \quad x \in \mathbb{R}^n$$

- The $\phi_1, \ldots, \phi_N$ are called called basis functions
- Vector $\boldsymbol{\phi}(x)$ is called the feature vector and the space

  $$\{\boldsymbol{\phi}(x) : x \in \mathbb{R}^n\}$$

  the feature space

The non-linear regression function has the primal representation

$$f(x) = \sum_{j=1}^{N} w_j \phi_j(x)$$

Feature Maps (Example 1 : [BIAS])

We've already seen one example with $\boldsymbol{\phi} : \mathbb{R}^n \to \mathbb{R}^{n+1}$

$$\boldsymbol{\phi}(x) = (x, 1)^\top$$

In the context of linear regression before application of the feature map we had

$$\operatorname*{arg\,min}_{w \in \mathbb{R}^n} \sum_{i=1}^{m} \left(w^\top x_i - y\right)^2$$

After the feature map we have

$$\operatorname*{arg\,min}_{w \in \mathbb{R}^n, b \in \mathbb{R}} \sum_{i=1}^{m} \left(w^\top x_i + b - y_i\right)^2$$

thus allowing us to learn a linear fit with a constant offset.

Feature Maps (Example 2 : [XOR])

Consider the XOR function defined as

x_1	x_2	x_1 XOR x_2
1	1	-1
1	-1	1
-1	1	1
-1	-1	-1

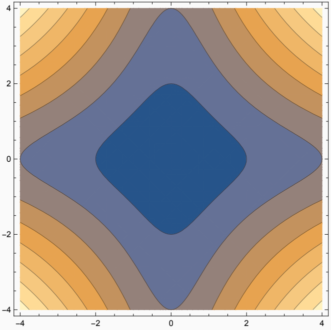
Does there exist a linear classifier that fits XOR perfectly? What if we add a bias term? Why?

What if instead we first apply the feature map $\phi(x) := (x, x_1 x_2)^\top$?

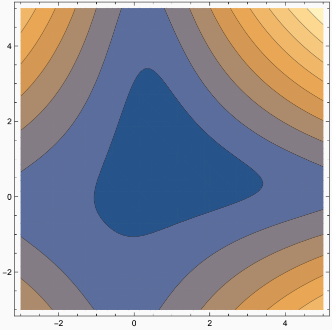
Let's visualize the unit balls associated with the "correlation" feature map

The distance between two points x, t after mapping in feature space is

$$\|\phi(x) - \phi(t)\|_2 = \sqrt{(x_1 - t_1)^2 + (x_2 - t_2)^2 + (x_1 x_2 - t_1 t_2)^2}$$



(a) Ball centred at $(0, 0)$.



(b) Ball centred at $(1, 1)$

Figure 3: Unit ball of 2d correlation feature map in "original space"

- Observe that the unit balls of in $\mathbb{R}^2$ w.r.t. to the feature-mapped

Feature (Example 3 : Correlations)

More generally the trick behind XOR was to add to the original vector the "correlate" x_1x_2.

More generally for second order correlations if $x \in \mathbb{R}^n$ we have

$$\phi(x) := (x, x_1x_1, x_1x_2, \ldots, x_1x_n, x_2x_2, x_2x_3, \ldots, x_2x_n, \ldots, x_nx_n)^\top$$

I.e., $\phi : \mathbb{R}^n \to \mathbb{R}^{\frac{n^2+3n}{2}}$.

What is the motivation for this feature map?

More generally we might also include higher order correlations.

What is a potential problem with this technique?

Kernels

Computational Considerations Revisited

Again, if $\mathrm{m} \ll \mathrm{N}$ it is more efficient to work with the dual representation

Key observation: in the dual representation we don't need to know $\boldsymbol{\phi}$ explicitly; we just need to know the inner product between any pair of feature vectors!

Example: Consider the following feature map with second order correlations $(\mathrm{N} = \mathrm{n}^2)$

$$\boldsymbol{\phi}(\mathrm{x}) = (\mathrm{x}_1\mathrm{x}_1, \mathrm{x}_1\mathrm{x}_2, \ldots, \mathrm{x}_\mathrm{n}\mathrm{x}_\mathrm{n})^\top$$

$$\begin{aligned}
\langle \boldsymbol{\phi}(\mathrm{x}), \boldsymbol{\phi}(\mathrm{t}) \rangle &= (\mathrm{x}_1\mathrm{x}_1, \mathrm{x}_1\mathrm{x}_2, \ldots, \mathrm{x}_\mathrm{n}\mathrm{x}_\mathrm{n})(\mathrm{t}_1\mathrm{t}_1, \mathrm{t}_1\mathrm{t}_2, \ldots, \mathrm{t}_\mathrm{n}\mathrm{t}_\mathrm{n})^\top \\
&= \mathrm{x}_1\mathrm{x}_1\mathrm{t}_1\mathrm{t}_1 + \mathrm{x}_1\mathrm{x}_2\mathrm{t}_1\mathrm{t}_2 + \ldots + \mathrm{x}_\mathrm{n}\mathrm{x}_\mathrm{n}\mathrm{t}_\mathrm{n}\mathrm{t}_\mathrm{n} \\
&= (\mathrm{x}_1\mathrm{t}_1 + \ldots + \mathrm{x}_\mathrm{n}\mathrm{t}_\mathrm{n})(\mathrm{x}_1\mathrm{t}_1 + \ldots + \mathrm{x}_\mathrm{n}\mathrm{t}_\mathrm{n}) \\
&= (\mathrm{x}^\top\mathrm{t})^2
\end{aligned}$$

Observe that $(\mathrm{x}^\top\mathrm{t})^2$ requires only $\mathrm{O}(\mathrm{n})$ computations whereas the more direct $(\mathrm{x}_1\mathrm{x}_1, \mathrm{x}_1\mathrm{x}_2, \ldots, \mathrm{x}_\mathrm{n}\mathrm{x}_\mathrm{n})(\mathrm{t}_1\mathrm{t}_1, \mathrm{t}_1\mathrm{t}_2, \ldots, \mathrm{t}_\mathrm{n}\mathrm{t}_\mathrm{n})^\top$ requires $\mathrm{O}(\mathrm{n}^2)$

Kernel Functions – Implicit Feature Map

Given a feature map ϕ we define its associated kernel function $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ as

$$K(x, t) := \langle \phi(x), \phi(t) \rangle, \quad x, t \in \mathbb{R}^n$$

- Key Point: for some feature map ϕ computing $K(x, t)$ is independent of N (only dependent on n). Where necessarily $\phi(x)$ depends on N.

Example (cont.) If $\phi(x) = (x_{i_1} x_{i_2} \cdots x_{i_r} : i_1, \ldots, i_r \in \{1, \ldots, n\})$ then we have that

$$K(x, t) = (x^\top t)^r$$

In this case $K(x, t)$ is computed with $O(n)$ operations, which is essentially independent of r or $N = n^r$. On the other hand, computing $\phi(x)$ requires $O(N)$ operations – Exponential in r!.

Question: So far the feature map has all r-order correlates how can we change it so that (r-1)-order, (r-2)-order, etc., correlates are included?

Redundancy of the feature map

Warning

The feature map is not unique! If $\boldsymbol{\phi}$ generates K so does $\hat{\boldsymbol{\phi}} = \mathrm{U}\boldsymbol{\phi}$ where U in an (any!) $\mathrm{N} \times \mathrm{N}$ orthogonal matrix. Even the dimension of $\boldsymbol{\phi}$ is not unique!

Proof.

$$(\mathrm{U}\boldsymbol{\phi})^\top(\mathrm{U}\boldsymbol{\phi}) = \boldsymbol{\phi}^\top \mathrm{U}^\top \mathrm{U}\boldsymbol{\phi} = \boldsymbol{\phi}^\top\boldsymbol{\phi}$$

Example

If $\mathrm{n} = 2$, $\mathrm{K}(\mathrm{x}, \mathrm{t}) = (\mathrm{x}^\top \mathrm{t})^2$ is generated by both $\boldsymbol{\phi}(\mathrm{x}) = (\mathrm{x}_1^2, \mathrm{x}_2^2, \mathrm{x}_1\mathrm{x}_2, \mathrm{x}_2\mathrm{x}_1)$ and $\hat{\boldsymbol{\phi}}(\mathrm{x}) = (\mathrm{x}_1^2, \mathrm{x}_2^2, \sqrt{2}\mathrm{x}_1\mathrm{x}_2)$.

Regularization-based learning algorithms

Let us open a short parenthesis and show that the dual form of ridge regression holds true for other loss functions as well

$$\mathcal{E}_{\mathrm{emp}_\lambda}(\mathrm{w}) = \sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{V}(\mathrm{y_i}, \langle \mathrm{w}, \boldsymbol{\phi}(\mathrm{x_i})\rangle) + \lambda \langle \mathrm{w}, \mathrm{w}\rangle, \quad \lambda > 0 \tag{6}$$

where $\mathrm{V} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a loss function

Theorem

If V is differentiable wrt. its second argument and w is a minimizer of E_λ then it has the form

$$\mathrm{w} = \sum_{\mathrm{i}=1}^{\mathrm{m}} \alpha_\mathrm{i} \boldsymbol{\phi}(\mathrm{x_i}) \quad \Rightarrow \quad \mathrm{f}(\mathrm{x}) = \langle \mathrm{w}, \boldsymbol{\phi}(\mathrm{x})\rangle = \sum_{\mathrm{i}=1}^{\mathrm{m}} \alpha_\mathrm{i} \mathrm{K}(\mathrm{x_i}, \mathrm{x})$$

This result is usually called the Representer Theorem

Representer theorem

Setting the derivative of E_λ wrt. w to zero we have

$$\sum_{i=1}^{m} V'(y_i, \langle w, \boldsymbol{\phi}(x_i)\rangle)\boldsymbol{\phi}(x_i) + 2\lambda w = 0 \Rightarrow w = \sum_{i=1}^{m} \alpha_i \boldsymbol{\phi}(x_i) \tag{7}$$

where V' is the partial derivative of V wrt. its second argument and we defined

$$\alpha_i = \frac{1}{2\lambda} V'(y_i, \langle w, \boldsymbol{\phi}(x_i)\rangle) \tag{8}$$

Thus we conclude that

$$f(x) = \langle w, \boldsymbol{\phi}(x)\rangle = \sum_{i=1}^{m} \alpha_i \langle \boldsymbol{\phi}(x_i), \boldsymbol{\phi}(x)\rangle = \sum_{i=1}^{m} \alpha_i K(x, x_i),$$

Some remarks

- Plugging eq.(7) in the rhs. of eq.(8) we obtain a set of equations for the coefficients α_i:

$$\alpha_i = \frac{1}{2\lambda} V' \left(y_i, \sum_{j=1}^{m} K(x_i, x_j)\alpha_j \right) \;, \quad i = 1, \ldots, m$$

  When V is the square loss and $\boldsymbol{\phi}(x) = x$ we retrieve the linear eq.(5)

- Substituting eq.(7) in eq.(6) we obtain an objective function for the α's:

$$\sum_{i=1}^{m} V(y_i, (K\boldsymbol{\alpha})_i) + \lambda \boldsymbol{\alpha}^\top K \boldsymbol{\alpha}, \quad \text{where} : K = (K(x_i, x_j))_{i,j=1}^{m}$$

Remark: the Representer Theorem holds true under more general conditions on V (for example V can be any continuous function)

What functions are "kernels"?

Question

Given a function $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, which properties of K guarantee that there exists a Hilbert space $\mathcal{H}$ and a feature map $\phi : \mathbb{R}^n \to \mathcal{H}$ such that $K(x, t) = \langle \phi(x), \phi(t) \rangle$?

Note 1

We've generalized the definition of finite-dimensional feature maps

$$\phi : \mathbb{R}^n \to \mathbb{R}^N$$

to now allow potentially infinite-dimensional feature maps

$$\phi : \mathbb{R}^n \to \mathcal{H}$$

Note (technical) 2

A Hilbert space is an inner product space which also contains the limit points of all its Cauchy sequences.

Positive Semidefinite Kernel

Definition

A function $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is positive semidefinite if it is symmetric and the matrix $(K(x_i, x_j) : i, j = 1, \ldots, k)$ is positive semidefinite for every $k \in \mathbb{N}$ and every $x_1, \ldots, x_k \in \mathbb{R}^n$

Theorem

K is positive semidefinite if and only if

$$K(x, t) = \langle \boldsymbol{\phi}(x), \boldsymbol{\phi}(t) \rangle, \quad x, t \in \mathbb{R}^n$$

for some feature map $\boldsymbol{\phi} : \mathbb{R}^n \to \mathcal{W}$ and Hilbert space $\mathcal{W}$

Note. We may replace domain $\mathbb{R}^n$ by any abstract set $\mathcal{X}$ in the above definitions.

Positive semidefinite kernel (cont.)

Proof of "$\Leftarrow$"

If $K(x,t) = \langle \boldsymbol{\phi}(x), \boldsymbol{\phi}(t) \rangle$ then we have that

$$\sum_{i,j=1}^{m} c_i c_j K(x_i, x_j) = \left\langle \sum_{i=1}^{m} c_i \boldsymbol{\phi}(x_i), \sum_{j=1}^{m} c_j \boldsymbol{\phi}(x_j) \right\rangle = \left\| \sum_{i=1}^{m} c_i \boldsymbol{\phi}(x_i) \right\|^2 \geq 0$$

for every choice of $m \in \mathbb{N}$, $x_i \in \mathbb{R}^d$ and $c_i \in R$, $i = 1, \ldots, m$

Note

the proof of '$\Rightarrow$' requires the notion of reproducing kernel Hilbert spaces. Informally, one can show that the linear span of the set of functions $\{K(x, \cdot) : x \in \mathbb{R}^n\}$ can be made into a Hilbert space H_K with inner product induced by the definition $\langle K(x, \cdot), K(t, \cdot) \rangle_K := K(x,t)$. In particular, the map $\boldsymbol{\phi} : \mathbb{R}^n \to H_K$ defined as $\boldsymbol{\phi}(x) = K(x, \cdot)$ is a feature map associated with K. Observe (check!) then with $f(\cdot) := \sum_{i=1}^{m} \alpha_i K(x_i, \cdot)$ that $\|f\|^2 = \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j K(x_i, x_j)$.

Two Example Kernels

Polynomial Kernel(s)

If $p : \mathbb{R} \to \mathbb{R}$ is a polynomial with nonnegative coefficients then $K(x, t) = p(x^\top t), x, t \in \mathbb{R}^n$ is a positive semidefinite kernel. For example if $a \geq 0$

- $K(x, t) = (x^\top t)^r$
- $K(x, t) = (a + x^\top t)^r$
- $K(x, t) = \sum_{i=0}^{d} \frac{a^i}{i!}(x^\top t)^i$

are each positive semidefinite kernels.

Gaussian Kernel

An important example of a "radial" kernel is the Gaussian kernel

$$K(x, t) = \exp(-\beta\|x - t\|^2), \quad \beta > 0, x, t \in \mathbb{R}^n$$

note: any corresponding feature map $\phi(\cdot)$ is ∞-dimensional.

Polynomial and Anova Kernel

Anova Kernel

$$K_a(x, t) = \prod_{i=1}^{n} (1 + x_i t_i)$$

Compare to the polynomial kernel $K_p(x, t) = (1 + x^\top t)^d$

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_n \end{pmatrix} \Rightarrow \phi_p(x) = \begin{pmatrix} 1 \\ \sqrt{d} x_1 \\ \sqrt{d} x_2 \\ \vdots \\ \sqrt{d} x_n \\ \sqrt{d(d-1)} x_1 x_2 \\ \vdots \\ \sqrt{\binom{d}{i_0, i_1, \dots, i_n} x_1^{i_1} x_2^{i_2} \cdots x_n^{i_n}} \\ \vdots \end{pmatrix} \qquad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_n \end{pmatrix} \Rightarrow \phi_a(x) = \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_n \\ x_1 x_2 \\ \vdots \\ \vdots \\ x_1 x_2 \dots x_n \end{pmatrix}$$

where $\sum_{j=0}^{n} i_j = d$

Problem: Argue that $\langle \phi_a(x), \phi_a(t) \rangle = K_a(x, t)$.

Kernel construction

Which operations/combinations (eg, products, sums, composition, etc.) of a given set of kernels is still a kernel?

If we address this question we can build more interesting kernels starting from simple ones

Example

We have already seen that $K(x, t) = (x^\top t)^r$ is a kernel. For which class of functions $p : \mathbb{R} \to \mathbb{R}$ is $p(x^\top t)$ a kernel? More generally, if K is a kernel when is $p(K(x, t))$ a kernel?

General linear kernel

If A is an $n \times n$ psd matrix the function $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ defined by

$$K(x, t) = x^\top A t$$

is a kernel

Proof

Since A is psd we can write it in the form $A = RR^\top$ for some $n \times n$ matrix R. Thus K is represented by the feature map $\phi(x) = R^\top x$

Alternatively, note that:

$$\sum_{i,j} c_i c_j x_i^\top A x_j = \sum_{i,j} c_i c_j (R^\top x_i)^\top (R^\top x_j) =$$
$$\sum_i c_i [(R^\top x_i)]^\top [\sum_j c_j (R^\top x_j)] = \| \sum_i c_i R^\top x_i \|^2 \geq 0$$

Kernel composition

More generally, if $K : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ is a kernel and $\boldsymbol{\phi} : \mathbb{R}^n \to \mathbb{R}^N$, then

$$\tilde{K}(x, t) = K(\boldsymbol{\phi}(x), \boldsymbol{\phi}(t))$$

is a kernel from $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$.

Proof

By hypothesis, K is a kernel and so, for every $x_1, \ldots, x_m \in \mathbb{R}^n$ the matrix $(K(\boldsymbol{\phi}(x_i), \boldsymbol{\phi}(x_j)) : i, j = 1, \ldots, m)$ is psd

In particular, the previous example corresponds to $K(x, t) = x^\top t$ and $\boldsymbol{\phi}(x) = R^\top x$

Kernel construction (cont.)

Question

If $K_1, \ldots, K_q$ are kernels on $\mathbb{R}^n$ and $F : \mathbb{R}^q \to \mathbb{R}$, when is the function

$$F(K_1(x,t), \ldots, K_q(x,t)), \quad x, t \in \mathbb{R}^n$$

a kernel?

Equivalently: when for every choice of $m \in \mathbb{N}$ and $A_1, \ldots, A_q$ $m \times m$ psd matrices, is the following matrix psd?

$$(F(A_{1,ij}, \ldots, A_{q,ij}) : i, j = 1, \ldots m)$$

We discuss some examples of functions F for which the answer to these question is YES

Nonnegative combination of kernels

If $\lambda_j \geq 0$, $j = 1, \ldots, q$ then $\sum_{j=1}^{q} \lambda_j K_j$ is a kernel

This fact is immediate (a non-negative combination of psd matrices is still psd)

Example: Let $q = n$ and $K_j(x, t) = x_j t_j$.

In particular, this implies that

- aK_1 is a kernel if $a \geq 0$
- $K_1 + K_2$ is a kernel

Product of kernels

The pointwise product of two kernels K_1 and K_2

$$K(x,t) := K_1(x,t)K_2(x,t), \quad x,t \in \mathbb{R}^d$$

is a kernel

Proof

Idea: The fact that the element-wise product of PSD matrices is again PSD implies that product of kernels is again a kernel.

Thus we need to show that if A and B are psd matrices, so is $C = (A_{ij}B_{ij} : i,j = 1,\ldots,n)$ (C is also called the Schur product of A and B). Since A and B are psd we can write them in the form $A = UU^\top$ and $B = VV^\top$ for some $n \times n$ matrices U and V.

$$\begin{aligned}
\sum_{i,j=1}^{m} z_i z_j C_{ij} &= \sum_{ij} z_i z_j \sum_r U_{ir}U_{jr} \sum_s V_{is}V_{js} = \sum_{ij}\sum_{rs} z_i z_j U_{ir}U_{jr}V_{is}V_{js} \\
&= \sum_{rs}\sum_{ij} z_i z_j U_{ir}U_{jr}V_{is}V_{js} = \sum_{rs}\sum_{i} z_i U_{ir}V_{is} \sum_j z_j U_{jr}V_{js} \\
&= \sum_{rs}\Big(\sum_i z_i U_{ir}V_{is}\Big)^2 \geq 0
\end{aligned}$$

Summary of constructions

Theorem

If K_1, K_2 are kernels, $a \geq 0$, A is a symmetric positive semi-definite matrix, K a kernel on $\mathbb{R}^N$ and $\boldsymbol{\phi} : \mathbb{R}^n \to \mathbb{R}^N$ then the following functions are positive semidefinite kernels on $\mathbb{R}^n$

1. $x^\top A t$
2. $K_1(x, t) + K_2(x, t)$
3. $aK_1(x, t)$
4. $K_1(x, t)K_2(x, t)$
5. $K(\boldsymbol{\phi}(x), \boldsymbol{\phi}(t))$

Polynomial of kernels

Let $F = p$ where $p : \mathbb{R}^q \to \mathbb{R}$ is a polynomial in q variables with nonnegative coefficients. By properties 2,3 and 4 above we conclude that p is a valid function

In particular if $q = 1$,

$$\sum_{i=1}^{d} a_i (K(x,t))^i$$

is a kernel if $a_1, \ldots, a_d \geq 0$

Polynomial kernels

The above observation implies that if $p : \mathbb{R} \to \mathbb{R}$ is a polynomial with nonnegative coefficients then $p(x^\top t), x, t \in \mathbb{R}^n$ is a kernel on $\mathbb{R}^n$. In particular if $a \geq 0$ the following are valid polynomial kernels

- $(x^\top t)^r$
- $(a + x^\top t)^r$
- $\sum_{i=0}^{d} \frac{a^i}{i!}(x^\top t)^i$

'Infinite polynomial' kernel

If in the last equation we set $r = \infty$ the series

$$\sum_{i=0}^{r} \frac{a^i}{i!} (x^\top t)^i$$

converges everywhere uniformly to $\mathsf{exp}(ax^\top t)$ showing that this function is also a kernel.

Assume for simplicity that $n = 1$. A feature map corresponding to the kernel $\mathsf{exp}(axt)$ is

$$\boldsymbol{\phi}(x) = \left(1,\ \sqrt{a}x,\ \sqrt{\frac{a}{2}}x^2,\ \sqrt{\frac{a^3}{6}}x^3, \ldots\right) = \left(\sqrt{\frac{a^i}{i!}}x^i : i \in \mathbb{N}\right)$$

- The feature space has an infinite dimensionality!

Translation invariant and radial kernels

We say that a kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is

- Translation invariant if it has the form

$$K(x, t) = H(x - t), \quad x, t \in \mathbb{R}^d$$

  where $H : \mathbb{R}^d \to \mathbb{R}$ is a differentiable function

- Radial if it has the form

$$K(x, t) = h(\|x - t\|), \quad x, t \in \mathbb{R}^d$$

  where $h : [0, \infty) \to [0, \infty)$ is a differentiable function

The Gaussian kernel

An important example of a radial kernel is the Gaussian kernel

$$K(x,t) = \mathsf{exp}(-\beta\|x-t\|^2), \quad \beta > 0, x, t \in \mathbb{R}^d$$

It is a kernel because it is the product of two kernels

$$K(x,t) = \left(\mathsf{exp}(-\beta(x^\top x + t^\top t))\right)\mathsf{exp}(2\beta x^\top t)$$

(We saw before that $\mathsf{exp}(2\beta x^\top t)$ is a kernel. Clearly $\mathsf{exp}(-\beta(x^\top x + t^\top t))$ is a kernel with one-dimensional feature map $\phi(x) = \mathsf{exp}(-\beta x^\top x)$)

Exercise:

Can you find a feature map representation for the Gaussian kernel?

Other kernels

In are examples we have mainly focused on kernels defined on $\mathbb{R}^n$, more generally and usefully they can be defined on other input spaces X for example.

1. Kernels between sets
2. Kernels on text and strings
3. Kernels between graphs
4. Kernel between vertices on a graph

• Defining useful kernels between on new domains X allows for a host of ML algorithms to be transferred to that domain. For example ridge regression, k-NN, SVMs k-means, PCA, etc.

Further examples (Kernels)

From Kernel Methods for Pattern Analysis, Shawe-Taylor.J, and Cristianini N., Cambridge University Press (2004)

Kernels

(to give you an idea)

- Definition 9.1 Polynomial kernel 286
  Computation 9.6 All-subsets kernel 289
  Computation 9.8 Gaussian kernel 290
  Computation 9.12 ANOVA kernel 293
  Computation 9.18 Alternative recursion for ANOVA kernel 296
  Computation 9.24 General graph kernels 301
  Definition 9.33 Exponential difiusion kernel 307
  Definition 9.34 von Neumann difiusion kernel 307
  Computation 9.35 Evaluating difiusion kernels 308
  Computation 9.46 Evaluating randomised kernels 315
  Definition 9.37 Intersection kernel 309
  Definition 9.38 Union-complement kernel 310
  Remark 9.40 Agreement kernel 310
  Section 9.6 Kernels on real numbers 311
  Remark 9.42 Spline kernels 313
  Definition 9.43 Derived subsets kernel 313
  Definition 10.5 Vector space kernel 325
  Computation 10.8 Latent semantic kernels 332
  Definition 11.7 The p-spectrum kernel 342
  Computation 11.10 The p-spectrum recursion 343
  Remark 11.13 Blended spectrum kernel 344
  Computation 11.17 All-subsequences kernel 347
  Computation 11.24 Fixed length subsequences kernel 352
- Computation 11.33 Naive recursion for gap-weighted subsequences kernel 358
  Computation 11.36 Gap-weighted subsequences kernel 360
  Computation 11.45 Trie-based string kernels 367
  Algorithm 9.14 ANOVA kernel 294
  Algorithm 9.25 Simple graph kernels 302
  Algorithm 11.20 All–non-contiguous subsequences kernel 350
  Algorithm 11.25 Fixed length subsequences kernel 352
  Algorithm 11.38 Gap-weighted subsequences kernel 361
  Algorithm 11.40 Character weighting string kernel 364
  Algorithm 11.41 Soft matching string kernel 365
  Algorithm 11.42 Gap number weighting string kernel 366
  Algorithm 11.46 Trie-based p-spectrum kernel 368
  Algorithm 11.51 Trie-based mismatch kernel 371
  Algorithm 11.54 Trie-based restricted gap-weighted kernel 374
  Algorithm 11.62 Co-rooted subtree kernel 380
  Algorithm 11.65 All-subtree kernel 383
  Algorithm 12.8 Fixed length HMM kernel 401
  Algorithm 12.14 Pair HMM kernel 407
  Algorithm 12.17 Hidden tree model kernel 411
  Algorithm 12.34 Fixed length Markov model Fisher kernel 427

www.kernel-methods.net

Further examples (Algorithms)

From Kernel Methods for Pattern Analysis, Shawe-Taylor.J, and Cristianini N., Cambridge University Press (2004)

Algorithms

(to give you an idea)

Computation 2.5 Ridge regression 30
Computation 5.14 Regularised Fisher discriminant 131
Computation 5.15 Regularised kernel Fisher discriminant 133
Computation 6.3 Maximising variance 141
Computation 6.18 Maximising covariance 154
Computation 6.30 Canonical correlation analysis 163
Computation 6.32 Kernel CCA 165
Computation 6.34 Regularised CCA 169
Computation 6.35 Kernel regularised CCA 169
Computation 7.1 Smallest enclosing hypersphere 193
Computation 7.7 Soft minimal hypersphere 199
Computation 7.10 nu-soft minimal hypersphere 202
Computation 7.19 Hard margin SVM 209
Computation 7.28 1-norm soft margin SVM 216
Computation 7.36 2-norm soft margin SVM 223
Computation 7.40 Ridge regression optimisation 229
Computation 7.43 Quadratic e-insensitive SVR 231
Computation 7.46 Linear e-insensitive SVR 233
Computation 7.50 nu-SVR 235
Computation 8.8 Soft ranking 254
Computation 8.17 Cluster quality 261
Computation 8.19 Cluster optimisation strategy 265
Computation 8.25 Multiclass clustering 272
Computation 8.27 Relaxed multiclass clustering 273
Computation 8.30 Visualisation quality 277

- Algorithm 5.1 Normalisation 110
Algorithm 5.3 Centering data 113
Algorithm 5.4 Simple novelty detection 116
Algorithm 5.6 Parzen based classifier 118
Algorithm 5.12 Cholesky decomposition or dual Gram–Schmidt 126
Algorithm 5.13 Standardising data 128
Algorithm 5.16 Kernel Fisher discriminant 134
Algorithm 6.6 Primal PCA 143
Algorithm 6.13 Kernel PCA 148
Algorithm 6.16 Whitening 152
Algorithm 6.31 Primal CCA 164
Algorithm 6.36 Kernel CCA 171
Algorithm 6.39 Principal components regression 175
Algorithm 6.42 PLS feature extraction 179
Algorithm 6.45 Primal PLS 182
Algorithm 6.48 Kernel PLS 187
Algorithm 7.2 Samllest hypersphere enclosing data 194
Algorithm 7.8 Soft hypersphere minimisation 201
Algorithm 7.11 nu-soft minimal hypersphere 204
Algorithm 7.21 Hard margin SVM 211
Algorithm 7.26 Alternative hard margin SVM 214
Algorithm 7.29 1-norm soft margin SVM 218
Algorithm 7.32 nu-SVM 221
Algorithm 7.37 2-norm soft margin SVM 225
Algorithm 7.41 Kernel ridge regression 229
Algorithm 7.45 2-norm SVR 232
Algorithm 7.47 1-norm SVR 234
Algorithm 7.51 nu-support vector regression 236
Algorithm 7.52 Kernel perceptron 237
Algorithm 7.59 Kernel adatron 242
Algorithm 7.61 On-line SVR 244
Algorithm 8.9 nu-ranking 254
Algorithm 8.14 On-line ranking 257
Algorithm 8.22 Kernel k-means 269
Algorithm 8.29 MDS for kernel-embedded data 276
Algorithm 8.33 Data visualisation 280

www.kernel-methods.net

Computational Summary for ridge regression

Summary : Computation with Basis Functions

Data: X, $(m \times n)$; y, $(m \times 1)$
Basis Functions: $\phi_1, \ldots, \phi_N$ where $\phi_i : \mathbb{R}^n \to \mathbb{R}$
Feature Map: $\boldsymbol{\phi} : \mathbb{R}^n \to \mathbb{R}^N$

$$\boldsymbol{\phi}(x) = (\phi_1(x), \ldots, \phi_N(x)), \quad x \in \mathbb{R}^n$$

Mapped Data Matrix:

$$\Phi := \begin{pmatrix} \boldsymbol{\phi}(x_1) \\ \vdots \\ \boldsymbol{\phi}(x_m) \end{pmatrix} = \begin{pmatrix} \phi_1(x_1) & \ldots & \phi_N(x_1) \\ \vdots & \ddots & \vdots \\ \phi_1(x_m) & \ldots & \phi_N(x_m) \end{pmatrix}, \quad (m \times N)$$

Regression Coefficients: $w = (\Phi^\top \Phi + \lambda I_N)^{-1} \Phi^\top y$
Regression Function: $\hat{y}(x) = \sum_{i=1}^{N} w_i \phi_i(x)$

Summary : Computation with Kernels

Data: $\mathrm{X}, \quad (\mathrm{m} \times \mathrm{n});\ \ \mathrm{y}, \quad (\mathrm{m} \times 1)$

Kernel Function: $\mathrm{K} : \mathbb{R}^{\mathrm{n}} \times \mathbb{R}^{\mathrm{n}} \to \mathbb{R}$

Kernel Matrix:

$$\mathrm{K} := \begin{pmatrix} \mathrm{K}(\mathrm{x}_1, \mathrm{x}_1) & \dots & \mathrm{K}(\mathrm{x}_1, \mathrm{x}_\mathrm{m}) \\ \vdots & \ddots & \vdots \\ \mathrm{K}(\mathrm{x}_\mathrm{m}, \mathrm{x}_1) & \dots & \mathrm{K}(\mathrm{x}_\mathrm{m}, \mathrm{x}_\mathrm{m}) \end{pmatrix}, \quad (\mathrm{m} \times \mathrm{m})$$

Regression Coefficients: $\boldsymbol{\alpha} = (\mathrm{K} + \lambda \mathrm{I}_\mathrm{m})^{-1}\mathrm{y}$

Regression Function: $\hat{\mathrm{y}}(\mathrm{x}) = \sum_{\mathrm{i}=1}^{\mathrm{m}} \alpha_\mathrm{i} \mathrm{K}(\mathrm{x}_\mathrm{i}, \mathrm{x})$

Suggested Readings

Chapters 2,3 (Additionally read chapter 9 for more depth). Kernel Methods for Pattern Analysis, Shawe-Taylor.J, and Cristianini N., Cambridge University Press (2004)

Going Deeper ...

- Regularization Matters: Generalization and Optimization of Neural Nets v.s. their Induced Kernel. One of series of paper on the Neural Tangent Kernel (connecting overparameterised neural networks to a particular RKHS). One of aim of this research is to get a better understanding for why NNs generalise.
- Convolution Kernels on Discrete Structures. Classic paper with a variety of nice ideas on Kernels for discrete structures.

Problems – 1

1. Prove results on page 9.
2. Consider the solution to linear regression optimisation problem. When is it advantageous to compute it via the primal solution? When is it advantageous to compute it via the dual solution? Explain why
3. Given a kernel $K : \mathbb{R}^2 \times \mathbb{R}^2 \to R$ where $K(x, t) := (1 + \langle x, t \rangle)^2$. Find a feature map $\phi : \mathbb{R}^2 \to \mathbb{R}^6$ which corresponds to the kernel.
4. For each of the following functions $K : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ argue whether they are a valid kernel (i.e. the kernel can be written as an inner product is some feature space) and when the answer is positive derive an associated feature map representation.
   4.1 $K(x, t) = x^\top Dt$, where D is the matrix
   $$\begin{bmatrix} 1 & 0 \\ 0 & 5 \end{bmatrix}$$
   4.2 $K(x, t) = x^\top Dt$, where D is the matrix
   $$\begin{bmatrix} -1 & 2 \\ 2 & 4 \end{bmatrix}$$
   4.3 $K(x, t) = \exp(x_1 t_1)$, where x_1 is the first component of the vector x and, likewise, t_1 is the first component of the vector t.
   4.4 $K(x, t) = x^\top t - (x^\top t)^2$.
   4.5 Now prove that if $K : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ is a given valid kernel then the following transformed kernels are also valid:
      4.5.1 $K(Ax, At)$, where A is a given 2×2 matrix.
      4.5.2 $f(x)K(x, t)f(t)$, where f is a given real-valued function.
5. Consider a Gaussian kernel function $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ defined by $K(x, z) := e^{-\|x-z\|^2}$, does there exist a finite-dimensional feature map representation? I.e., does theres exist a $\phi : \mathbb{R}^n \to \mathbb{R}^d$ such that $K(x, z) = \langle \phi(x), \phi(z) \rangle$? Indicate an answer “yes” or “no” and provide an argument supporting your answer. [hard]

Problems – 2

1. Argue that the vector space definition implies that every element $x \in X$ has an additive inverse $-x \in X$ such that $x + (-x) = 0$
2. Kernels between sets. Let X be a finite set define $K : 2^X \times 2^X \to R$ as

$$K(A, B) := 2^{|A \cap B|}$$

   where $A, B \subseteq X$. Prove that K is a kernel.
3. Min Kernel.
   3.1 Argue that $\mathsf{min}(x, t)$ (where $x, t \in [0, \infty)$) is a kernel for "a more complicated example" on page 9. See discussion on 41, on going from a kernel to a Hilbert space. [technical]
   3.2 Determining an explicit feature map. Find a set of basis functions $\phi_i : [0, \infty) \to R$ $(i = 1, 2, \ldots, \infty)$ such that

$$\mathsf{min}(x, t) = \sum_{i=1}^{\infty} \phi_i(x)\phi_i(t)$$

   [technical, very difficult]
   Note: Mercer's Theorem, which we did not cover, implies such a feature map exists but does not give a way of constructing the