annotated version

Machine Learning Course - CS-433

Kernel Ridge Regression and the Kernel Trick

Oct 31, 2019

changes by Martin Jaggi 2019, changes by Rüdiger Urbanke 2018, changes by Martin Jaggi 2016, 2017 ©Mohammad Emtiyaz Khan 2015

Last updated on: October 31, 2019

EPFL

Motivation

The ridge solution $\mathbf{w}^\star \in \mathbb{R}^D$ has a counterpart $\boldsymbol{\alpha}^\star \in \mathbb{R}^N$. Using duality, we will establish a relationship between $\mathbf{w}^\star$ and $\boldsymbol{\alpha}^\star$ which leads the way to kernels.

Ridge regression

cost of (A)
$\mathcal{O}(D^3 + ND^2)$
cost of (B)
$\mathcal{O}(N^3 + DN^2)$

Recall the ridge regression problem

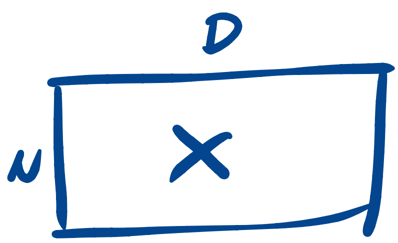
$$\min_{\mathbf{w}} \quad \tfrac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \tfrac{\lambda}{2}\|\mathbf{w}\|^2$$

For its solution, we have that

(A) $$\mathbf{w}^\star = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_D)^{-1}\mathbf{X}^\top\mathbf{y}$$
(B) $$= \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_N)^{-1}\mathbf{y} =: \mathbf{X}^\top\boldsymbol{\alpha}^\star,$$

case $N \gg D$
case $D \gg N$

where $\boldsymbol{\alpha}^\star := (\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_N)^{-1}\mathbf{y}$.



Lemma

This can be proved using the following identity: let $\mathbf{P}$ be an $N \times D$ matrix while $\mathbf{Q}$ be $D \times N$, and let both $\mathbf{PQ} + \mathbf{I}$ and $\mathbf{QP} + \mathbf{I}$ be invertible.

$$(\mathbf{PQ} + \mathbf{I}_N)^{-1}\mathbf{P} = \mathbf{P}(\mathbf{QP} + \mathbf{I}_D)^{-1}$$

$\Leftarrow$ $(PQ+I)^{-1} P (QP+I) = P$
$\Leftarrow$ $P(QP+I) = (PQ+I)P$
$\Leftarrow$ $PQP + P = PQP + P$

What are the computational complexities for the above two ways of computing $\mathbf{w}^\star$?

With this, we know that $\mathbf{w}^\star = \mathbf{X}^\top \boldsymbol{\alpha}^\star$ lies in the column space of $\mathbf{X}^\top$,

$$\text{where } \mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1D} \\ x_{21} & x_{22} & \dots & x_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{D2} & \dots & x_{ND} \end{bmatrix}$$

data point

The representer theorem

The representer theorem generalizes this result: for a $\mathbf{w}^\star$ minimizing the following function for any $\mathcal{L}_n$,

$\lambda > 0$

$$\min_{\mathbf{w}} \sum_{n=1}^{N} \mathcal{L}_n(\mathbf{x}_n^\top \mathbf{w}, y_n) + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

L2-Regularization

there exists $\boldsymbol{\alpha}^\star$ such that $\mathbf{w}^\star = \mathbf{X}^\top \boldsymbol{\alpha}^\star$.

Such a general statement was originally proved by *Schölkopf, Herbrich and Smola (2001)*.

Kernelized ridge regression

The representer theorem allows us to write an equivalent optimization problem in terms of $\boldsymbol{\alpha}$. For example, for ridge regression, the following two problems are equivalent:

primal problem

$$\mathbf{w}^\star = \arg\min_{\mathbf{w}} \quad \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

$$\boldsymbol{\alpha}^\star = \arg\max_{\boldsymbol{\alpha}} \quad -\frac{1}{2}\cancel{\boldsymbol{\alpha}}^\top(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_N)\boldsymbol{\alpha} + \cancel{\boldsymbol{\alpha}}^\top\mathbf{y} \quad = \mathcal{L}_D$$

dual problem

i.e. they both have the same optimal value. Also, we can always have the correspondence mapping $\mathbf{w} = \mathbf{X}^\top\boldsymbol{\alpha}$.

Most importantly, the second problem is expressed in terms of the matrix $\mathbf{X}\mathbf{X}^\top$. This is our first example of a kernel matrix.

$$\nabla\mathcal{L}_D(\alpha) \overset{!}{=} 0$$

Note: To see the equivalence, you can show that we obtain equal optimal values for the two problems. Take the gradient of the second expression, to get $(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_N)\boldsymbol{\alpha} - \mathbf{y}$. $\overset{!}{=} 0$
Setting this to $\mathbf{0}$ and solving for $\boldsymbol{\alpha}$ results in $\boldsymbol{\alpha}^\star = \underbrace{(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_N)}^{-1}\mathbf{y}$.

If we combine this with the representer theorem $\mathbf{w}^\star = \mathbf{X}^\top\boldsymbol{\alpha}^\star$ we find back the dual solution.

Advantages of kernelized ridge regression

First, it might be computationally efficient in some cases when solving the system of equations.

Second, by defining $\mathbf{K} = \mathbf{X}\mathbf{X}^\top$, we can work directly with $\mathbf{K}$ and never have to worry about $\mathbf{X}$. This is the kernel trick.

Third, working with $\boldsymbol{\alpha}$ is sometimes advantageous, and provides additional information for each datapoint (e.g. as in SVMs).

original data matrix

D, N, X

Q, N, $\Phi(x)$

expanded data

$Q \gg D$

Kernel functions

The linear kernel is defined below:

$$\mathbf{K} = \mathbf{X}\mathbf{X}^\top = \begin{bmatrix} \mathbf{x}_1^\top\mathbf{x}_1 & \mathbf{x}_1^\top\mathbf{x}_2 & \dots & \mathbf{x}_1^\top\mathbf{x}_N \\ \mathbf{x}_2^\top\mathbf{x}_1 & \mathbf{x}_2^\top\mathbf{x}_2 & \dots & \mathbf{x}_2^\top\mathbf{x}_N \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_N^\top\mathbf{x}_1 & \mathbf{x}_N^\top\mathbf{x}_2 & \dots & \mathbf{x}_N^\top\mathbf{x}_N \end{bmatrix}.$$

$\mathbb{R}^D \ni x \longrightarrow \Phi(x) \in \mathbb{R}^Q$

Kernel with basis functions $\boldsymbol{\phi}(\mathbf{x})$ with $\mathbf{K} := \boldsymbol{\Phi}\boldsymbol{\Phi}^\top$ is shown below:

$$\begin{bmatrix} \boldsymbol{\phi}(\mathbf{x}_1)^\top\boldsymbol{\phi}(\mathbf{x}_1) & \boldsymbol{\phi}(\mathbf{x}_1)^\top\boldsymbol{\phi}(\mathbf{x}_2) & \dots & \boldsymbol{\phi}(\mathbf{x}_1)^\top\boldsymbol{\phi}(\mathbf{x}_N) \\ \boldsymbol{\phi}(\mathbf{x}_2)^\top\boldsymbol{\phi}(\mathbf{x}_1) & \boldsymbol{\phi}(\mathbf{x}_2)^\top\boldsymbol{\phi}(\mathbf{x}_2) & \dots & \boldsymbol{\phi}(\mathbf{x}_2)^\top\boldsymbol{\phi}(\mathbf{x}_N) \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\phi}(\mathbf{x}_N)^\top\boldsymbol{\phi}(\mathbf{x}_1) & \boldsymbol{\phi}(\mathbf{x}_N)^\top\boldsymbol{\phi}(\mathbf{x}_2) & \dots & \boldsymbol{\phi}(\mathbf{x}_N)^\top\boldsymbol{\phi}(\mathbf{x}_N) \end{bmatrix}.$$

The kernel trick

A big advantage of using kernels is that we do not need to specify $\boldsymbol{\phi}(\mathbf{x})$ explicitly, since we can work directly with $\mathbf{K}$.

We will use a kernel function $\kappa(\mathbf{x}, \mathbf{x}')$ and compute the (i, j)-th entry of $\mathbf{K}$ as $K_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$. A kernel function κ is usually associated with a feature map $\boldsymbol{\phi}$, such that

$$\kappa(\mathbf{x}, \mathbf{x}') := \boldsymbol{\phi}(\mathbf{x})^\top \boldsymbol{\phi}(\mathbf{x}') \ .$$

For example, for the linear kernel $\kappa(\mathbf{x}, \mathbf{x}') := \mathbf{x}^\top \mathbf{x}'$, the feature map is just the original features, $\boldsymbol{\phi}(\mathbf{x}') = \mathbf{x}'$.

$K(x, x') = (x x')^2$

Another example: The kernel $\kappa(x, x') := x^2 (x')^2$ corresponds to $\phi(x) = x^2$, and $\kappa(\mathbf{x}, \mathbf{x}') := (x_1 x_1' + x_2 x_2' + x_3 x_3')^2$ corresponds to $= (x^\top x')^2$ Polynomial Kernel

$$\boldsymbol{\phi}(\mathbf{x})^\top = [\, x_1^2,\ x_2^2,\ x_3^2,\ \sqrt{2} x_1 x_2,\ \sqrt{2} x_1 x_3,\ \sqrt{2} x_2 x_3 \,]$$

$Q = 6$

The good news is that the evaluation of a kernel is often faster when using κ instead of $\boldsymbol{\phi}$.

Visualization

Why would we want such general feature maps?

See video explaining linear separation in the kernel space (where $\phi(\mathbf{x})$ maps to) corresponding to non-linear separation in the original $\mathbf{x}$-space: https://www.youtube.com/watch?v=3liCbRZPrZA

Examples of kernels

The above kernel is an example of the polynomial kernel. Another example is the Radial Basis Function (RBF) kernel.

Polynomial kernel

$= (x^T x')^7$

$K(x,x') = h(x^T x')$

$$\kappa(\mathbf{x}, \mathbf{x}') = \exp\left[-\frac{1}{2}\underbrace{(\mathbf{x} - \mathbf{x}')^\top(\mathbf{x} - \mathbf{x}')}_{\|x-x'\|^2}\right]$$

$K(x,x') = h(\|x-x'\|)$

$h = \exp(-\ldots)$

See more examples in Section 14.2 of Murphy's book.

A natural question is the following: how can we ensure that there exists a $\boldsymbol{\phi}$ corresponding to a given kernel $\mathbf{K}$? The answer is: as long as the kernel satisfies certain properties.

Is there Φ s.t. $K(x,x) = \Phi(x)^T \Phi(x')$?

Properties of a kernel

A kernel function must be an inner-product in some feature space. Here are a few properties that ensure it is the case.

$\exists \Phi$ s.t.

$K = \Phi \Phi^T$?

1. $\mathbf{K}$ should be symmetric, i.e. $\kappa(\mathbf{x}, \mathbf{x}') = \kappa(\mathbf{x}', \mathbf{x})$.
2. For any arbitrary input set $\{\mathbf{x}_n\}$ and all N, $\mathbf{K}$ should be positive semi-definite.

$\alpha^T A \alpha \geq 0$

An important subclass is the positive-definite kernel functions, giving rise to infinite-dimensional feature spaces.

given α^*,

How to do prediction on test data (new)

	no kernel	kernel
new datapoint	x_{test}	$\Phi(x_{test})$
prediction	$x_{test}^T w$	$\Phi(x_{test})^T w = \sum_i \Phi(x_t) \Phi(x_i) \alpha_i$ $k(x_t, x_i)$
correspondence $w^* = X^T \alpha^*$	$w = \sum_i x_i \alpha_i$	$w = \sum_i \Phi(x_i) \alpha_i$
cost	$O(D)$	$O(N \cdot \text{kernel})$

Exercises

1. Understand the relationship $\mathbf{w}^\star = \mathbf{X}^\top \boldsymbol{\alpha}^\star$. Understand the statement of the representer theorem.

2. Show that the primal and dual formulations of ridge regression are equivalent. Hint: show that the optimization problems corresponding to $\mathbf{w}$ and $\boldsymbol{\alpha}$ have the same optimal value.

3. Get familiar with various examples of kernels. See Section 6.2 of Bishop on examples of kernel construction. Read Section 14.2 of Murphy's book for examples of kernels.

4. Revise and understand the difference between positive-definite and positive-semi-definite matrices.

5. If you are interested in more details about kernels, read about Mercer and Matern kernels from Kevin Murphy's Section 14.2. There is also a small note by Matthias Seeger on the git repository under lectures/07, in particular for the case of infinite dimensional $\boldsymbol{\phi}$.