An Identity for Kernel Ridge Regression

Yuri Kalnishkan

Computer Learning Reseach Centre and
Department of Computer Science
Royal Holloway, University of London

March 2012

Outline

References

the results of this talk appeared in

- An Identity for Kernel Ridge Regression by F. Zhdanov and Y. Kalnishkan (to appear in Theoretical Computer Science) — see arXiv:1112.1390
- Competing with Gaussian linear experts by F. Zhdanov and V. Vovk (arXiv:0910.4683).

Batch Learning Problem

- suppose we are given a training set of pairs $(x_1, y_1), (x_2, y_2), \dots, (x_T, y_T)$, where
  — signals (examples, objects) x_t come from a set X
  — outcomes y_t are reals
- the task is to predict labels for new yet unseen signals $x \in X$

Ridge Regression

- kernel ridge regression (KRR) suggests the function $f_{\mathrm{RR}}(x) = Y'(K + aI)^{-1}k(x)$, where

$$K = \begin{pmatrix} \mathcal{K}(x_1, x_1) & \mathcal{K}(x_1, x_2) & \dots & \mathcal{K}(x_1, x_T) \\ \mathcal{K}(x_2, x_1) & \mathcal{K}(x_2, x_2) & \dots & \mathcal{K}(x_2, x_T) \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{K}(x_T, x_1) & \mathcal{K}(x_T, x_2) & \dots & \mathcal{K}(x_T, x_T) \end{pmatrix},$$

$$k(x) = \begin{pmatrix} \mathcal{K}(x_1, x) \\ \mathcal{K}(x_2, x) \\ \vdots \\ \mathcal{K}(x_T, x) \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{pmatrix}$$

- a is a parameter called ridge and $\mathcal{K}$ is a kernel

Kernels

- a kernel is a function of two arguments $\mathcal{K} : X \times X \to \mathbb{R}$ which is
  - symmetric, i.e., $\mathcal{K}(x_1, x_2) = \mathcal{K}(x_2, x_1)$
  - positive-semidefinite, i.e.,
    — the matrix $(\mathcal{K}(x_i, x_j))_{i,j=1}^n$ is always positive-semidefinite, i.e.,
    — for all n, all $x_1, x_2, \dots, x_n \in X$ and all $u_1, u_2, \dots, u_n \in \mathbb{R}$ we have

$$\sum_{i,j=1}^{n} u_i u_j \mathcal{K}(x_i, x_j) \geq 0$$

- if $\mathcal{K}$ is a kernel and $a > 0$, then the matrix $K + aI$ is positive-definite and therefore nonsingular

Examples of Kernels

- let $X = \mathbb{R}^n$ (or a subset of $\mathbb{R}^n$); the following popular kernels are used:
  — linear kernel $\mathcal{K}(x_1, x_2) = x_1' x_2$
  — Vapnik's polynomial kernel $\mathcal{K}_d(x_1, x_2) = (1 + x_1' x_2)^d$
  — radial-based (rbf) kernel $\mathcal{K}_\sigma(x_1, x_2) = e^{\|x_1 - x_2\|^2/\sigma^2}$ (and other functions depending on $\|x_1 - x_2\|$)
  — ANOVA kernels
  — spline kernels
  — etc

Justification

- ridge regression always specifies a function on the set of signals X
- why use f_{RR}?
  1. performs well in practice
  2. functional analysis: f_{RR} is optimal in a certain class
  3. probability theory: $f_{\mathrm{RR}}(x)$ is conditional expectation

RKHS

- each kernel $\mathcal{K}$ specifies a unique reproducing kernel Hilbert space (RKHS) $\mathcal{F}$, which
  - is a Hilbert space consisting of functions on X
  - contains functions $\mathcal{K}(x, \cdot)$ for all $x \in X$
  - has the scalar product $\langle \cdot, \cdot \rangle_{\mathcal{F}}$ satisfying the reproducing property $f(x) = \langle f, \mathcal{K}(x, \cdot) \rangle_{\mathcal{F}}$ for all $f \in \mathcal{F}$ and $x \in X$
- $\mathcal{F}$ contains all $\mathcal{K}(x, \cdot)$ and their linear combinations $\sum_i c_i \mathcal{K}(x_i, \cdot)$
  — the combinations are dense in $\mathcal{F}$

Some References

- Aronszajn, N. La theorie generale des noyaux reproduisants et ses applications (Proc. Cambridge Philos. Soc, vol 39, 1943)
  — in French
- Krein, M.G. Hermitian-Positive Kernels on Homogeneous Spaces, Parts I and II (AMS Translations, 1963, 34, 1)
  — a translation of a 1940s Russian paper
- `http://onlineprediction.net/?n=Main.KernelMethods`
  — a tutorial I have written

Optimality

- f_{RR} is a linear combination of $\mathcal{K}(x_i, \cdot)$ and therefore belongs to $\mathcal{F}$
- it is the minimum of

$$a\|f\|_{\mathcal{F}}^2 + \sum_{t=1}^{T}(f(x_t) - y_t)^2$$

over $f \in \mathcal{F}$
— the later term is quadratic loss
— the former term provides regularisation

- ridge regression can be thought of as curve fitting in the RKHS

Evaluation Functional

- in RKHS the evaluation functional $f \to f(x)$ is the scalar product by $\mathcal{K}(x, \cdot)$
— therefore it is continuous (on $\mathcal{F}$)
- this property characterises RKHSs: an RKHS is
— a Hilbert space of functions on X
— such that the evaluation functional is continuous for every $x \in X$
- the continuity of the evaluation functional means that $f(x) \to 0$ as $\|f\|_{\mathcal{F}} \to 0$
— the norm is consistent with the evaluation
- an RKHS is therefore ‘a reasonable Hilbert space’

Feature Spaces

- let Φ maps X into a Hilbert space H (feature space)
- for every $h \in H$ consider a ‘feature regressor’ $f_h(x) = \langle h, \Phi(x)\rangle_H$
— what are the functions f_h?
- the function

$$\mathcal{K}(x_1, x_2) = \langle \Phi(x_1), \Phi(x_2)\rangle_H$$

is a kernel
— and every kernel $\mathcal{K}$ can be represented in this way

- the set of functions f_h coincides with the RKHS $\mathcal{F}$ corresponding to $\mathcal{K}$ and the norm can be given by

$$\|f\|_{\mathcal{F}} = \min_{f_h = f} \|h\|_H$$

- thus RKHS consists of ‘regressors in a feature space’

On-line Learning Protocol

- in on-line learning the learner tries to predict each outcome y_t before it becomes available
- protocol:

  FOR $t = 1, 2, \dots$
  (1) $\mathfrak{A}$ observes x_t
  (2) $\mathfrak{A}$ outputs prediction γ_t
  (3) $\mathfrak{A}$ observes true outcome y_t
  END FOR
- in machine learning the performance is usually assessed by means of cumulative loss $\sum_{t=1}^{T}(\gamma_t - y_t)^2$
- prequential principle by Phil Dawid:
  — performance should be judged by what has happened and not by what could have happened

On-line Kernel Ridge Regression

- ridge regression can be applied in on-line mode
- on step t:
  — form a sample of known examples $(x_1, y_1), (x_2, y_2), \dots, (x_{t-1}, y_{t-1})$
  — populate matrices Y_{t-1}, K_{t-1}, and $k_{t-1}(x_t)$
  — output the prediction $\gamma_t^{\mathrm{RR}} = Y'_{t-1}(K_{t-1} + aI)^{-1}k_{t-1}(x)$
- question: how does the on-line cumulative loss of ridge regression $\sum_{t=1}^{T}(\gamma_t - y_t)^2$ compare against the optimal loss $\sum_{t=1}^{T}(f(x_t) - y_t)^2$?
  — how much do we loose by not knowing all (x_t, y_t) in advance?

Identity

- the main result of this talk:

$$\sum_{t=1}^{T}\frac{(\gamma_t^{\mathrm{RR}} - y_t)^2}{1 + d_t/a} = \min_{f\in\mathcal{F}}\left(\sum_{t=1}^{T}(f(x_t) - y_t)^2 + a\|f\|_{\mathcal{F}}^2\right)$$
$$= aY'_T(K_T + aI)^{-1}Y_T \quad ,$$

  where $d_t = \mathcal{K}(x_t, x_t) - k'_{t-1}(x_t)(K_{t-1} + aI)^{-1}k_{t-1}(x_t) > 0$
- the result holds for all sequences (x_t, y_t)
  — it is not a probabilistic statement

A Covariance is a Kernel

- a random field (random process) on X is a collection of random variables z_x, $x \in X$
  — we need to postulate that any finite number of them has a joint distribution
  — let $\mathbf{E}\, z_x = 0$
- the covariance $\mathcal{K}(x_1, x_2) = \mathbf{E}\, z_{x_1} z_{x_2}$ is a kernel on X
  — symmetry: obvious
  — positive-semidefiniteness:

$$0 \le \mathbf{E}\left(\sum_{i=1}^{n} u_i z_{x_i}\right)^2 = \sum_{i,j=1}^{n} u_i u_j\, \mathbf{E}\, z_{x_1} z_{x_2} = \sum_{i,j=1}^{n} u_i u_j \mathcal{K}(x_i, x_j)$$

A Kernel is a Covariance

- for every kernel $\mathcal{K}$ there is a Gaussian random field z_x such that $\mathcal{K}(x_1, x_2) = \mathbf{E}\, z_{x_1} z_{x_2}$
- proof:
  — for every finite set $x_1, x_2, \ldots, x_n$ there is a multivariate Gaussian distribution with means of 0 and covariances $\mathcal{K}$
  — the distributions can be 'joined together' by the Kolmogorov extension (or existence) theorem
- more on second order random functions and covariances:
  — M. Loève, Probability Theory II, Springer, 1963

Learning

- consider Gaussian noise ε_x such that
  — we have $\mathbf{E}\, \varepsilon_x = 0$ and $\mathbf{var}\, \varepsilon_x = \sigma^2 = a$
  — all ε_x are independent from each other and from all z_x (exists by Kolmogorov extension theorem)
- let us assume that outcomes y are a random process $y_x = z_x + \varepsilon_x$
  — we have $\mathbf{E}\, y_{x_1} y_{x_2} = \mathcal{K}(x_1, x_2) + a\delta_{x_1, x_2}$
- estimating y_x given a sample $(x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T)$ becomes a probabilistic task
  — note that xs are not stochastic: we just know the value of the process at some non-random points

Ridge Regression

- the conditional distribution of y_x given that $y_{x_1} = y_1$, $y_{x_2} = y_2, \ldots, y_{x_T} = y_T$ is
  — Gaussian
  — has the mean $f_{\mathrm{RR}}(x)$
  — has the variance
  $d_x + \sigma^2 = \mathcal{K}(x, x) - k'(x)(K + \sigma^2 I)^{-1} k(x) + a$
- references:
  — C. E. Rasmussen and C. K. I. Williams. Gaussian Processes for Machine Learning. MIT Press, 2006
  — C. M. Bishop. Pattern Recognition and Machine Learning. Springer, 2006

Repeating Signals

- in RKHSs there is no problem if some xs in the sample coincide and in the probabilistic model this is impossible — indeed, one cannot have two values for the same y_x
- solution: let us replace X by $X \times \{1, 2, 3, \dots\}$, define the kernel by
$$\mathcal{K}((x_1, t_1), (x_2, t_2)) = \mathcal{K}(x_1, x_2)$$
and pad each x_t to (x_t, t)

Probability

- the proof is by calculating the joint density $p_{y_{x_1}, y_{x_2}, \dots, y_{x_T}}(y_1, y_2, \dots, y_T)$ in three ways:
  1. as a chain of conditional probabilities;
  2. by marginalisation;
  3. directly.

Conditional Probabilities

- by decomposing the density we get

$$\begin{gathered} p_{y_{x_1}, y_{x_2}, \dots, y_{x_T}}(y_1, y_2, \dots, y_T) = \\ p_{y_{x_T}}(y_T \mid y_{x_1} = y_1, y_{x_2} = y_2, \dots, y_{x_{T-1}} = y_{T-1}) \cdot \\ p_{y_{x_{T-1}}}(y_T \mid y_{x_1} = y_1, y_{x_2} = y_2, \dots, y_{x_{T-2}} = y_{T-2}) \cdot \\ \dots \\ p_{y_{x_1}}(y_1) \end{gathered}$$

- each y_{x_t} has a Gaussian distribution with the mean of γ_t^{RR} and the variance $d_t + a$

Marginalisation (1)

- let us compress $y_{x_1}, y_{x_2}, \ldots, y_{x_T}$ to Y_{X_T}
  — the same for Z
- the density is the integral of a joint density:

$$p_{Y_{X_T}}(Y_T) = \int_{\mathbb{R}^T} p_{Y_{X_T}, Z_{X_T}}(Y_T, Z_T) dZ_T$$

  where

$$p_{Y_{X_T}, Z_{X_T}}(Y_T, Z_T) = p_{Y_{X_T}}(Y_T \mid Z_T) p_{Z_{X_T}}(Z_T)$$

Marginalisation (2)

- evaluation of the integral reduces to the following:
  — let $Q(\theta) = \theta' A \theta + \theta' b + c$, where the matrix A is symmetric positive-definite
  — then

$$\int_{\mathbb{R}^n} e^{-Q(\theta)} d\theta = e^{-Q(\theta_0)} \frac{\pi^{n/2}}{\sqrt{\det A}}$$

  where $\theta_0 = \arg\min_{\mathbb{R}^n} Q$.
- hence the infimum in the middle term of the identity

Direct Evaluation

- all ys are Gaussian with the covariance
  $\mathbf{E}\, y_{x_i} y_{x_j} = \mathcal{K}(x_i, x_j) + a\delta_{x_i, x_j}$
- the density can be written down easily

A Determinant Identity

- it remains to take the logarithm and to apply the following identity to kill off extra terms:

$$(d_1 + \sigma^2)(d_2 + \sigma^2) \ldots (d_T + \sigma^2) = \det(K_T + \sigma^2 I)$$

- the identity follows from Frobenius's identity

$$\det \begin{pmatrix} A & u \\ v' & d \end{pmatrix} = (d - v' A^{-1} u) \det A \ ,$$

1. Ridge Regression

2. Ridge Regression and Reproducing Kernel Hilbert Spaces

3. Regression in On-line Learning

4. Ridge Regression and Random Fields

5. Proof of the Identity

6. Corollaries

7. An Alternative Proof

Multiplicative Bound

- let $\mathcal{K}(x, x) \le c_{\mathcal{F}}$ on X
  — this constant uniformly bounds the norm of the evaluation functional
  — for the existence of a finite $c_{\mathcal{F}}$ it is sufficient for X to be compact and $\mathcal{K}$ to be continuous on X^2
- then

$$\sum_{t=1}^{T}(\gamma_t^{\mathrm{RR}} - y_t)^2 \le \left(1 + \frac{c_{\mathcal{F}}^2}{a}\right) \min_{f \in \mathcal{F}} \left(\sum_{t=1}^{T}(f(x_t) - y_t)^2 + a\|f\|_{\mathcal{F}}^2\right) = (a + c_{\mathcal{F}}^2) Y_T'(K_T + aI)^{-1} Y_T$$

Clipped Regression

- suppose that true outcomes are bounded: $|y| \le Y$
  — it makes no sense to output prediction outside $[-Y, Y]$.
  — clipped ridge regression outputs the ridge regression prediction if it is inside the interval or the closest point of the interval otherwise
  — we have $(\gamma^{\mathrm{RR},Y} - y)^2 \le 4Y^2$
- we get:

$$\sum_{t=1}^{T}(\gamma_t^{\mathrm{RR},Y} - y_t)^2 \le \min_{f \in \mathcal{F}} \left(\sum_{t=1}^{T}(f(x_t) - y_t)^2 + a\|f\|_{\mathcal{F}}^2\right) + 4Y^2 \ln\det\left(I + \frac{1}{a}K_T\right)$$

Finite-Dimensional Case

- let $X = \mathbb{R}^n$ and $\mathcal{K}(x_1, x_2) = x_1' x_2$
  — the RKHS is $\mathbb{R}^n$ with the quadratic norm
- for the clipped regression we get

$$\sum_{t=1}^{T}(\gamma_t^{\mathrm{RR},Y} - y_t)^2 \le \min_{\theta \in \mathbb{R}^n} \left(\sum_{t=1}^{T}(\theta' x_t - y_t)^2 + a\|\theta\|^2\right) + 4Y^2 n \ln\left(1 + \frac{TB^2}{an}\right)$$

where $\|x_t\| \le B$

Asymptotic Comparison (1)

- let X be a compact metric space and a kernel $\mathcal{K}$ be continuous on X^2
- consider a sequence $(x_1, y_1), (x_2, y_2), \ldots$
- if there is $f \in \mathcal{F}$ such that

$$\sum_{t=1}^{\infty} (y_t - f(x_t))^2 < +\infty$$

then

$$\sum_{t=1}^{\infty} (y_t - \gamma_t^{\mathrm{RR}})^2 < +\infty$$

— this follows from the multiplicative bound

Asymptotic Comparison (2)

- if for all $f \in \mathcal{F}$ we have

$$\sum_{t=1}^{\infty} (y_t - f(x_t))^2 = +\infty$$

then

$$\lim_{T\to\infty} \frac{\sum_{t=1}^{T} (y_t - \gamma_t^{\mathrm{RR}})^2}{\min_{f\in\mathcal{F}} \left(\sum_{t=1}^{T} (y_t - f(x_t))^2 + a\|f\|^2\right)} = 1$$

— this holds because $d_t \to 0$ for continuous kernels on compact domains

Prediction with Expert Advice

- in prediction with expert advice the learner reads to experts' predictions before making its own:

for $t = 1, 2, \ldots$
  experts and the learner observe x_t
  experts $\theta \in \Theta$ announce predictions $\gamma_t^{\theta} \in \Gamma$
  learner outputs prediction $\gamma_t \in \Gamma$
  reality announces outcome $y_t \in \Omega$
  each expert $\theta \in \Theta$ suffers loss $\lambda(\gamma_t^{\theta}, y_t)$
  learner suffers loss $\lambda(\gamma_t, y_t)$
endfor

- a loss function $\lambda : \Gamma \times \Omega \to [0, +\infty)$ measures the deviation between predictions and outcomes

Game

- it is important that the sets of outcomes and predictions may differ
- we take $\Omega = \mathbb{R}$ and Γ to be the set of all continuous density functions, i.e., continuous $\xi : \mathbb{R} \to [0, +\infty)$ such that $\int_{-\infty}^{+\infty} \xi(t) dt = 1$
- the loss is logarithmic likelihood:

$$\lambda(\xi, y) = -\ln \xi(y)$$

Experts

- let the signals be real vectors from $\mathbb{R}^n$
- let the experts be Gaussian densities: an expert θ predicts

$$\xi_t^\theta(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\theta' x_t - y)^2}{2\sigma^2}}$$

Merging

- the aggregating algorithm assigns weights to experts, updates the weights on every trial and mixes the experts with the current weights
- for this game the aggregating algorithm amounts to the Bayesian mixture
- assume the prior

$$p_0(\theta) = \frac{1}{(2\pi)^{n/2}} e^{-\|\theta\|^2/2}$$

- the learner will then output Gaussian distribution with the mean of γ_t^{RR} and the variance $d_t + \sigma^2$
— the kernel is linear

Identity

- a key lemma about the aggregating algorithm states

$$\mathrm{Loss}_t = -\ln \int_\Theta e^{-\mathrm{Loss}_t(\theta)} P_0(d\theta).$$

- this leads to the linear case of the identity
- the general kernel formula can be obtained using a standard procedure