Machine Learning
Exercise 7

Marc Toussaint
TAs: Janik Hager, Philipp Kratzer
Machine Learning & Robotics lab, U Stuttgart
Universitätsstraße 38, 70569 Stuttgart, Germany

May 28, 2019

(DS BSc students can skip the exercise 3.)

1 Kernels and Features (4 Points)

Reconsider the equations for Kernel Ridge regression given on slide 05:3, and – if features are given – the definition of the kernel function and $\kappa(x)$ in terms of the features as given on slide 05:2.

a) Prove that Kernel Ridge regression with the linear kernel function $k(x, x') = 1 + x^\top x'$ is equivalent to Ridge regression with linear features $\phi(x) = \begin{pmatrix} 1 \\ x \end{pmatrix}$. (1 P)

b) In Kernel Ridge regression, the optimal function is of the form $f(x) = \kappa(x)^\top \alpha$ and therefore linear in $\kappa(x)$. In plain ridge regression, the optimal function is of the form $f(x) = \phi(x)^\top \beta$ and linear in $\phi(x)$. Prove that choosing $k(x, x') = (1 + x^\top x')^2$ implies that $f(x) = \kappa(x)^\top \alpha$ is a second order polynomial over x. (2 P)

c) Equally, note that choosing the squared exponential kernel $k(x, x') = \exp(-\gamma \,|\, x - x' \,|^2)$ implies that the optimal $f(x)$ is linear in radial basis function (RBF) features. Does this necessarily impliy that Kernel Ridge regression with squared exponential kernel, and plain Ridge regression with RBF features are exactly equivalent? (Equivalent means, have the same optimal function.) Distinguish the cases $\lambda = 0$ (no regularization) and $\lambda > 0$. (1 P)

(Voluntary: Practically test yourself on the regression problem from Exercise 2, whether Kernel Ridge Regression and RBF features are exactly equivalent.)

2 Kernel Construction (4 Points)

For a non-empty set X, a kernel is a symmetric function $k : X \times X \to \mathbb{R}$. Note that the set X can be arbitrarily structured (real vector space, graphs, images, strings and so on). A very important class of useful kernels for machine learning are positive definite kernels. A kernel is called *positive definite*, if for all arbitrary finite subsets $\{x_i\}_{i=1}^n \subseteq X$ the corresponding *kernel matrix* K with elements $K_{ij} = k(x_i, x_j)$ is positive *semi*-definite,

$$\alpha \in \mathbb{R}^n \;\Rightarrow\; \alpha^\top K \alpha \geq 0 \;. \tag{1}$$

Let $k_1, k_2 : X \times X \to \mathbb{R}$ be two positive definite kernels. Often, one wants to construct more complicated kernels out of existing ones. Proof that

1. $k(x, x') = k_1(x, x') + k_2(x, x')$
2. $k(x, x') = c \cdot k_1(x, x') \quad \text{for} \quad c \geq 0$
3. $k(x, x') = k_1(x, x') \cdot k_2(x, x')$
4. $k(x, x') = k_1(f(x), f(x')) \quad \text{for} \quad f : X \to X$

are positive definite kernels.

3 Kernel logistic regression (2 Points)

The "kernel trick" is generally applicable whenever the "solution" (which may be the predictive function $f^{\text{ridge}}(x)$, or the discriminative function, or principal components...) can be written in a form that only uses the kernel function $k(x, x')$, but never features $\phi(x)$ or parameters β explicitly.

Derive a kernelization of Logistic Regression. That is, think about how you could perform the Newton iterations based only on the kernel function $k(x, x')$.

Tips: Reformulate the Newton iterations

$$\beta \leftarrow \beta - (\boldsymbol{X}^\top W \boldsymbol{X} + 2\lambda I)^{-1} \ [\boldsymbol{X}^\top(\boldsymbol{p} - \boldsymbol{y}) + 2\lambda I\beta] \tag{2}$$

using the two Woodbury identities

$$(X^\top W X + A)^{-1} X^\top W = A^{-1} X^\top (X A^{-1} X^\top + W^{-1})^{-1} \tag{3}$$

$$(X^\top W X + A)^{-1} = A^{-1} - A^{-1} X^\top (X A^{-1} X^\top + W^{-1})^{-1} X A^{-1} \tag{4}$$

Note that you'll need to handle the $\boldsymbol{X}^\top(\boldsymbol{p} - \boldsymbol{y})$ and $2\lambda I\beta$ differently.

Then think about what is actually been iterated in the kernalized case: surely we cannot iteratively update the optimal parameters, because we want to rewrite equations to never touch β or $\phi(x)$ explicitly.