Machine Learning Worksheet 7

Kernels and Constrainted Optimization

1 The Gaussian kernel

Problem 1: One of the nice things about kernels is that new kernels can be constructed out of already given ones. Use the five kernel construction rules from the lecture to prove that the function

$$K(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{|\boldsymbol{x}-\boldsymbol{y}|^2}{2\sigma^2}\right)$$

is a kernel.

(Hint: Use the Taylor expansion of the exponential function to prove that $\exp(K_1(\boldsymbol{x}, \boldsymbol{y}))$ is a kernel if $K_1(\boldsymbol{x}, \boldsymbol{y})$ is a kernel.)

2nd hint: It might help to apply the rule $K_3(\boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{y}))$ with the linear kernel $K_3(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^T \boldsymbol{y}$ and consider a feature map $\boldsymbol{\phi}(z)$ with only one feature.

Problem 2: Find an infinite-dimensional feature space $\boldsymbol{\phi}(\boldsymbol{x})$ corresponding to the Gaussian kernel, i.e. determine $\boldsymbol{\phi}(\boldsymbol{x})$ so that

$$\boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{\phi}(\boldsymbol{y}) = \exp\left(-\frac{|\boldsymbol{x}-\boldsymbol{y}|^2}{2\sigma^2}\right) .$$

(Hint: The multinomial formula turns a power of a sum into a weighted sum of products,

$$\left(\sum_{t=1}^{m} x_t\right)^n = \sum_{k_1+k_2+\cdots+k_m=n} \binom{n}{k_1, k_2, \ldots, k_m} \prod_{t=1}^{m} x_t^{k_t} ,$$

with $\binom{n}{k_1, k_2, \ldots, k_m} = \frac{n!}{k_1! k_2! \cdots k_m!}$.)

Problem 3: Informally explain the difference between doing linear regression with a Gaussian RBF basis and doing linear regression with a Gaussian kernel.

2 Kernel Perceptron

Problem 4: In this exercise, we develop a dual formulation of the perceptron algorithm. Using the perceptron learning rule you learned in the lecture, show that the learned weight vector can be written as a linear combination of the vectors $t^{(n)} \boldsymbol{\phi}(\boldsymbol{x}^{(n)})$ where $t^{(n)} \in \{-1, +1\}$. Denote the coefficients of this linear combination by α_n and derive a formulation of the perceptron learning algorithm, and the predictive function for the perceptron in terms of the α_n. Show that the feature vector $\boldsymbol{\phi}(\boldsymbol{x})$ enters only in the form of the kernel function $K(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{\phi}(\boldsymbol{y})$.

3 Kernelized k-nearest neighbours

To classify the point $\boldsymbol{x}$ the k-nearest neighbours finds the k training samples $\mathcal{N} = \{\boldsymbol{x}^{(s_1)}, \boldsymbol{x}^{(s_2)}, \ldots, \boldsymbol{x}^{(s_k)}\}$ that have the shortest distance $|\boldsymbol{x}-\boldsymbol{x}^{(s_i)}|_2$ to $\boldsymbol{x}$. Then the label that is mostly represented in the neighbour set $\mathcal{N}$ is assigned to $\boldsymbol{x}$.

Problem 5: Formulate the k-nearest neighbours algorithm in feature space by introducing the feature map $\boldsymbol{\phi}(\boldsymbol{x})$. Then rewrite the k-nearest neighbours algorithm so that it only depends on the scalar product in feature space $K(\boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{y})) = \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{\phi}(\boldsymbol{y})$.

4 Convex functions

Problem 6: Given two convex functions $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ show that the sum $h(\boldsymbol{x}) = f(\boldsymbol{x}) + g(\boldsymbol{x})$ and the scaled function $u(\boldsymbol{x}) = cf(\boldsymbol{x})$ with $c \geq 0$ are convex.

Problem 7: Consider the family of convex functions $f_\lambda(\boldsymbol{x})$, that is for every $\lambda \in \mathbb{R}$ the function $f_\lambda(\boldsymbol{x})$ is convex. Prove that the pointwise maximum $g(\boldsymbol{x}) = \max_\lambda f_\lambda(\boldsymbol{x})$ is convex.

Problem 8: Show that the Lagrange dual function $g(\boldsymbol{\alpha}) = \min_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{\alpha})$ is concave. (A function $f(x)$ is concave if and only if $-f(x)$ is convex.)

5 A simple constrained optimization problem

Consider the simple optimization problem

$$\begin{aligned} \text{minimize } & f_0(x) = x^2 + 1 \\ \text{subject to } & f_1(x) = (x-2)(x-4) \leq 0\,. \end{aligned}$$

Problem 9: Plot the objective $f_0(x)$ and the constraint $f_1(x)$ versus x in one plot. Show the feasible points. Use this plot to directly give the solution of the optimization problem.

Problem 10: Derive the Lagrangian $L(x, \alpha)$ and use a computer program to plot it for $\alpha \in \{0, 0.5, 1, 1.5, 2, 3, 4, 5, 8\}$. For which regions is the value of the Lagrangian larger than the objective function? For which regions is the value of the Lagrangian smaller than the objective function? Which points are unaffected? What is the upper bound of $\min_x L(x, \alpha)$ for all $\alpha \geq 0$?

Problem 11: Derive and plot the Lagrange dual function $g(\alpha)$. State the dual problem.

Problem 12: Find the dual optimal value and the dual optimal solution α^*.

Problem 13: Is the dual optimal value also the minimum of the original optimization problem?

Problem 14: Is the constraint f_1 active or inactive? Can you also see this from the plot of the primal problem? What does it mean when a constraint is active, i.e. what is the effect of an active constraint on the solution?