Machine learning@Data science, 2021–Exercises.

Set 4. Due March 12, 2021

Problem 13 Consider the cost functional $A(f) = \mathbf{E}\phi(-f(X)Y)$ where $\phi : \mathbb{R} \to \mathbb{R}_+$ is a positive, increasing, strictly convex cost function, $f : \mathcal{X} \to \mathbb{R}$ is a real-valued function and $Y \in \{-1, 1\}$. Determine the function f^* that minimizes $A(f)$. Show that the classifier $g(x) = \mathrm{sgn}(f^*(x))$ is the Bayes classifier.

What is f^* in the case of the "hinge loss" $\phi(x) = (1 + x)_+$?

Problem 14 Let $\mathcal{H}$ be the Hilbert space of all sequences $s = \{s_n\}_{n=0}^{\infty}$ satisfying $\sum_{n=0}^{\infty} s_n^2 < \infty$ with inner product $\langle s, t\rangle = \sum_{n=0}^{\infty} s_n t_n$. Consider the feature map $\Phi : \mathbb{R} \to \mathcal{H}$ that assigns, to each real number x, the sequence $\Phi(x)$ whose n-th element equals

$$(\Phi(x))_n = \frac{1}{\sqrt{n!}} x^n e^{-x^2/2} , \quad n = 0, 1, 2, \dots .$$

Determine the kernel function $K(x, y) = \langle\Phi(x), \Phi(y)\rangle$ for $x, y \in \mathbb{R}$. (You may use the fact that $\sum_{n=0}^{\infty} x^n/n! = e^x$.)

Can you generalize the kernel so that it is defined on $\mathbb{R}^d \times \mathbb{R}^d$ instead of $\mathbb{R} \times \mathbb{R}$? What is the corresponding feature map?

Problem 15 Coinsider a binary classification problem in which X takes values in $\mathbb{R}^d$ and $Y \in \{-1, 1\}$. Suppose $\mathbf{P}\{Y = 1\} = 1/2$ and the class-conditional distributions are normal with identity covariance matrix but different mean vectors, say $(-a, 0, 0, \dots, 0)$ and $(a, 0, 0, \dots, 0)$.

Write a program that generates a training data of n i.i.d. pairs $D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ of random variables distributed as described above.

Train a linear classifier by performing stochastic gradient-descent minimization of the function $f(w) = \mathbf{E}\phi(-w^T XY)$ for $w \in \mathbb{R}^d$ where ϕ is an increasing convex function.

Estimate the probability of error of the obtained classifier and compare it to the optimal linear classifier (which is the Bayes classifier in this case). Try different choices of a, n, d and the function ϕ (including $\phi(x) = (1 + x)_+$, $\phi(x) = e^x$, $\phi(x) = \log_2(1 + e^x)$.

Play with the tuning parameter of the stochastic gradient descent algorithm.

Problem 16 Let $\mathcal{X}_n = \{0, 1\}^n$ be the set of binary strings of length n. Let $m < n$. We say that $s \in \{0, 1\}^m$ is a *substring* of $x = (x_1, \dots, x_n) \in \mathcal{X}_n$ if for some $i \in \{1, \dots, n - m + 1\}$, $s = (x_i, \dots, x_{i+m-1})$.

Define the function $K : \mathcal{X}_n \times \mathcal{X}_n \to \mathbb{R}$ as the number of common substrings of its arguments, that is,

$$K(x, y) = \sum_{s\in\{0,1\}^m} \mathbb{1}_{\{s \text{ is a substring of both } x \text{ and } y\}} \qquad \text{for } x, y \in \mathcal{X}_n.$$

Prove that K is a kernel function. Determine a feature map Φ defined of $\mathcal{X}_n$, mapping to some Hilbert space $\mathcal{H}$ for which $K(x, y)$ is the inner product of $\Phi(x)$ and $\Phi(y)$. What is the dimension of $\mathcal{H}$?