Theoretical homework 1

2-AIN-150, Winter 2023

Deadline: 30.10.2023, 23:59, Google classroom

Before you start solving the homework, please read the general instruction at the end of the document. Submitted solutions should be your own. Do not copy and do not tried to find solution in literature or over the internet.

Theory of ML

Consider problem of regression over set of hyphotheses $H = \{h_b(x) = 2x + b\}$.

a) Describe algorithm, which takes given training data $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$ and selects hyphothesis, which minimizes error function: $E(b) = \sum_{i=1}^{n} (h_b(x^{(i)}) - y^{(i)})^2$.

b) If data are generated using distribution $P_{x,y}$ defined as:

- distribution of x is uniform over interval $[0, 100]$.
- for given x is $\Pr(y = 2x + 7|x) = 0.4$ and $\Pr(y = 2x - 5|x) = 0.6$ (other values of y with given x are not possible).

What is the best possible test error for hyphothesis set H (i.e. test for best hyphothesis out of H) if we assume, that data are independent samples from $P_{x,y}$.

c) For distribution $P_{x,y}$ from part b) and $n = 1$ calculate expected training and test error (i.e. we generate one sample, we train and from this process we calculate expected errors).

d) For distribution $P_{x,y}$ from part b) and general n calculate expected training and test error. What is the relationship of there errors with optimal test error if $n \to \infty$? How would you set good size of training sample based on your results?

General instructions

Submit a solution in PDF format via Google classroom.