Homework Assignment 1

2-INF-150: Machine Learning, Fall 2021

Deadline: 25.10.2021 23:55
(electronic submission through classroom)

1. Matrix Algebra.

a) For the statement:

$$\nabla_A \mathrm{tr}(BA) = B^T,$$

specify necessary conditions for the statement to hold and prove the statement.

b) Consider linear regression with L_2 norm error function. Let X be the design matrix, y be the target vector and θ be the vector of linear regression parameters. In the lecture, we have shown that to mimize the error function, it must hold:

$$X^T X\theta = X^T y.$$

If $X^T X$ is a regular matrix, we can multiply from the left both sides of the equation with the inverse matrix $(X^T X)^{-1}$ obtaining:

$$\theta = (X^T X)^{-1} X^T y = X^{-1} X^{T^{-1}} X^T y = X^{-1} y,$$

Thus the solution for the vector θ can be easily computed directly as $\theta = X^{-1}y$.

Where is the problem with this derivation?

2. Machine Learning Theory. Consider a regression problem with the set of hypotheses:

$$H = \{h_b : x \to 2x + b\}.$$

a) Describe an algorithm that for the training set $(x_1, y_1), \ldots, (x_t, y_t)$ finds the hypothesis minimizing the error function $J(b) = \frac{1}{t}\sum_{i=1}^{t}(h_b(x_i) - y_i)^2$.

(Your algorithm should be MUCH simpler than the algorithms used for the traditional linear regression.)

b) Consider a probabilistic distribution $P_{x,y}$ defined as follows:

- distribution of x is uniform over the interval $[0, 100]$
- for given x, $\Pr(y = 2x + 3 \,|\, x) = 0.3$ and $\Pr(y = 2x - 2 \,|\, x) = 0.7$ (there are no other values of y that can occur in combination with x).

Compute bias of the set of hypotheses H if we assume that the data will be independent samples from $P_{x,y}$.

c) For distribution $P_{x,y}$ from part b) and number of training examples $t = 1$, compute expected training and expected testing error.

d) For distribution $P_{x,y}$ from part b) and general value of t, compute expected training and testing error. How are these errors related to the bias computed in part b) as $t \to \infty$? (You can illustrate this by drawing a graph.)

Based on your results, would you be able to give a guidance as to an appropriate size of the training set in this scenario?