2 Practical: Linear Regression

In Nestor you will find the file *linreg.mat* and, alternatively, the files *xtrain.csv, ytrain.csv, xtest.csv, ytest.csv*. These files contain two sets of 500 25-dim. feature vectors and two sets of 500 continuous target labels.
You are supposed to perform linear regression, assuming an underlying dependence of the form $y = \mathbf{w} \cdot \mathbf{x}$ with parameters $\mathbf{w} \in I\!R^{25}$.

2.1 Implementation (3 points)

Implement linear regression in order to determine an estimate $\mathbf{w}^*$ from the first P feature vectors and corresponding labels in the set *xtrain* and *ytrain*, respectively. To this end, construct the pseudoinverse solution discussed in class, which minimizes the Mean Squared Error (MSE):[1]

$$E_{train} = \frac{1}{P} \sum_{\mu=1}^{P} \frac{1}{2} \Big(\mathbf{w}^* \cdot \mathbf{x}^{\mu}_{train} - y^{\mu}_{train} \Big)^2 .$$

You can use available built-in functions or available code for the pseudoinverse (e.g. matlab: pinv) but make sure that the matrix and vector dimensions are correct and that you obtain $\mathbf{w}^* \in I\!R^{25}$. Unlike on the lecture slides we normalize the MSE here by P. This does not matter for the actual optimization, but makes the resulting MSE comparable for training sets with different P and the test error (see below).
With the resulting vector of parameters, also determine the corresponding test error

$$E_{test} = \frac{1}{1000} \sum_{\mu=1}^{500} \left(\mathbf{w}^* \cdot \mathbf{x}^{\mu}_{test} - y^{\mu}_{test} \right)^2$$

which is always computed w.r.t. all of the 500 samples in the second data set, i.e. *xtest.csv, ytest.csv*

2.2 Perform the regression and report results (7 points in total)

You should hand in a report comprising

- (1 point) A brief introduction.
- (2 points) A figure displaying both E_{train} and E_{test} for at least the following training set sizes: $P = 30, 40, 50, 75, 100, 200, 300, 400, 500$. Of course you may obtain and present results for more intermediate values of P.
- (2 points) Figures displaying the estimated parameter vectors $\mathbf{w}^*$, for instance as bar graphs, for $P = 30, 40, 50, 75, 100$ and $P = 500$. The x-axis should correspond to the 25 components (indices $j = 1, 2, \ldots 25$) and the y-axis should display the component w^*_j of the parameter vector.
Can you guess what the true vector $\mathbf{w}$ was that generated the data?
- (2 points) A brief discussion of your results.

[1] see remark on notation at the end of the assignment

Bonus (suggestions)

1 point max. in total:

- Determine the best estimates for small training sets with $P < 25$ by introducing a small regularization term as discussed in class.
- Implement the MSE minimization by (stochastic) gradient descent and compare the resulting weight vectors with the pseudionverse solution.
- Study the influence of the learning rate in stochastic gradient minimization of the MSE.

Notation and Matlab:

For the dot product of two vectors $\mathbf{a} \in \mathbb{R}^N$ and $\mathbf{b} \in \mathbb{R}^N$ we use the equivalent notations

$$\mathbf{a}^\top \mathbf{b} = \mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^{N} a_i \, b_i,$$

depending on context and available space. Note that the result is a real number (scalar). In Matlab, the dot product between two column vectors can be obtained as

`dot(a,b),     sum(a.*b),     a'*b=b'*a`

all giving the same result, where `a.*b` is the element-wise multiplication which yields a vector and `a'` denotes the transposed of `a`, i.e. a row vector.