Linear Models

Exercise 1 : Derivative of $l_\sigma(c, y(\mathbf{x}))$

$$\text{Show that } \nabla l_\sigma(c, y(\mathbf{x})) = -\delta \cdot \mathbf{x},$$
$$\text{i.e., } \frac{\partial}{\partial w_i}[l_\sigma(c, y(\mathbf{x}))] = -\delta \cdot x_i$$

with $\delta = c - y(\mathbf{x})$ for $y(\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x})$.

Exercise 2 : Batch Gradient Descent

(a) Explain the difference of LMS and BGD that allows to overcome a significant flaw of the former.

(b) How is that related to the idea of a global loss?

(c) Verify that, indeed, $\nabla L(\mathbf{w}) = \sum_{(\mathbf{x},c)\in D} \nabla l(c, y(\mathbf{x}))$.

(d) Why is that needed to justify the BGD algorithm?

Exercise 3 : Convergence criterion

In algorithms like LMS and BGD, we analyze the global loss $L(\mathbf{w}_t)$ or the norm of the loss gradient $||\nabla L(\mathbf{w}_t)||$ at the time step t.

In this exercise, we will compare three choices of comparison values in the convergence criterion and see that they are approximately equivalent.

(a) By using the update rule $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \cdot \nabla L(\mathbf{w}_t)$, express $\Delta L = L(\mathbf{w}_t) - L(\mathbf{w}_{t+1})$ in terms of $||\nabla L(\mathbf{w}_t)||$.

Hint: You may estimate using the Taylor formula: $f(x) \approx f(a) + \nabla f(a)^T(x - a)$ for cleverly chosen f, x and a.

(b) Express $||\Delta\mathbf{w}|| = ||\mathbf{w}_{t+1} - \mathbf{w}_t||$ in terms of $||\nabla L(\mathbf{w}_t)||$.

(c) What does that mean for the choice of ε in the convergence criterion? Why would we prefer to use $||\nabla L(\mathbf{w}_t)||$?

Exercise 4 : Overfitting and train-test leakage

(a) What is the experimental setup of choice when trying to detect overfitting?

(b) What are methods to mitigate overfitting?

(c) What must be payed attention to when performing a train-validation split on the following datasets in the given problems?

(c1) Detecting pneumonia from chest x-rays. Data includes 112,120 unique images from 30,805 unique patients.

(c2) Given 1000 voice recordings (single sentences) of 100 people in total from 5 German cities. The model should be able to classify the dialects of arbitrary people into one of these cities.

(c3) Given 1000 voice recordings (single sentences) of 100 people in total from 5 German cities. The model should be able to rate the dialects of arbitrary people from all over Germany by intelligibility.

(c4) Given 1000 voice recordings (single sentences) of 100 people in total from 5 German cities. The model should be able to classify the person that said a given sentence.