AIC-4101C – Machine learning

1. Linear regression and gradient descent

Kevin Zagalo & Marwan W. El Khazen

<kevin.zagalo@inria.fr>

<marwan.wehaiba-el-khazen@inria.fr>

Fall 2021

Exercise 1 *Gradient descent*

We want to minimize the function $F : \mathbf{R}^d \to \mathbf{R}$, differentiable on $\mathbf{R}^d$. Let ∇F be its gradient. The gradient descent is an iterative algorithm :

— **Initialization** $x_0 \in \mathbf{R}^d$

— **Iteration** $x_{k+1} = x_k - \alpha \nabla F(x_k)$

Until convergence.

Question 1

Suggest a stopping criterion for the algorithm.

Question 2

Implement the algorithm in Python (we will define a function taking the function F and its gradient as input).

Question 3

What happens if F is not convex ?

Question 4

Discuss the importance of the initial point x_0 in the convex case, then in the non convex case.

■

Exercise 2 *Least squared method*

A linear regression takes the inputs $X = (\mathbf{x}_1, \mathbf{x}_2, \ldots)$, the associated responses $\mathbf{y} = (y_{\mathbf{x}_1}, y_{\mathbf{x}_2}, \ldots)$, and as output $\beta = (\beta_1, \ldots, \beta_d)$, and the bias β_0. Let us define $h_\beta(\mathbf{x}) = \mathbf{x} \cdot \beta + \beta_0$. Let the local and global errors, respectively :

$$e(\mathbf{x}; \beta) = \frac{1}{2}(y_{\mathbf{x}} - h_\beta(\mathbf{x}))^2 \; ; \; E(X; \beta) = \frac{1}{|X|} \sum_{\mathbf{x} \in X} e(\mathbf{x}; \beta) \tag{1}$$

Question 1

So far, we have treated β and β_0 separately. Show that we can consider them simultaneously. How should we adapt the linear regression model?

In order to estimate the optimal parameter $\hat{\beta}$, we minimize the global error. For that, we will implement the following algorithm :

— **Input** X, the associated responses $\mathbf{y}$ and $\alpha > 0$.
— $\beta^{(0)} = \vec{0}$
— **Do**
  — **Compute** $L(\beta^{(t)}) = E(X; \beta^{(t)})$
  — **Update** $\beta^{(t+1)} = \beta^{(t)} - \alpha \nabla L(\beta^{(t)})$
— **Until** all data is explored and *convergence* of the sequence of parameters.

Question 2

Write a function that updates the parameters.

Question 3

Implement the algorithm.

■

Exercice 3 *Proof question*

Let $f(\beta) = (y - X\beta)^\top (y - X\beta)$ and $\hat{\beta} = \text{argmin}_{\beta \in \mathbf{R}^{d+1}} f(\beta)$. Show that $\hat{\beta} = (X^\top X)^{-1} X^\top \mathbf{y}$.

Recall : (see here)

- Let $v, a \in \mathbf{R}^k$. Then $\frac{\partial v^\top a}{\partial v} = \frac{\partial a^\top v}{\partial v} = a$,
- Let $v \in \mathbf{R}^k, M \in \mathbf{R}^{k \times k}$. Then $\frac{\partial v^\top M v}{\partial v} = (M + M^\top) v$.

■