Statistical learning

1. Linear regression and stochastic gradient descent

Kevin Zagalo

<kevin.zagalo@inria.fr>

2019 – 2020

Notations

- n – sample size
- k – number of features
- $X \in \mathbf{R}^{n \times d}$ – input data $(x_i)_{i=1,\dots,n}$
- $\hat{\beta} \in \mathbf{R}^d$ – estimator of the weight parameters to build
- $y \in \mathbf{R}^n$ – output data
- $\hat{y} \in \mathbf{R}^n$ – predicted output

Least squares approximation

- The fit of a model to a data point is measured by its **residuals**

$$y_i - \hat{y}_i, \;\; i = 1, \ldots, n$$

- The least squared loss (or cost) function focuses in the **mean squared error** of this measure, that is

$$L(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \langle x_i, w\rangle) \;\text{ where }\; \ell(y, y') = \frac{1}{2}|y - y'|^2$$

. We immediately note that

$$\nabla L(w) = \frac{1}{n}\sum_{i=1}^{n} \ell'(y_i, \langle x_i, w\rangle) x_i \;\text{ with }\; \ell'(y, y') = \partial_{y'}\ell(y, y')$$

Least squares estimator

- That leads us to the **optimization** problem :

$$\hat{\beta} = \underset{w}{\operatorname{argmin}}\, L(w) \Leftrightarrow \nabla L(\hat{\beta}) = \vec{0} \tag{1}$$

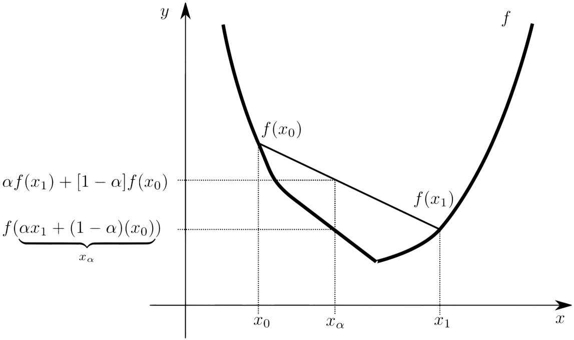
  - What are the conditions for uniqueness?
  - Why is (1) true?

- We are looking for the estimators $\hat{y}_i = \langle x_i, \hat{\beta} \rangle$.

Convex functions

A function $f : \mathbf{R}^d \to \mathbf{R}$ is **convex** if for all $x, y \in \mathbf{R}^d$ and all $\alpha \in [0, 1]$

$$f(\alpha x + (1-\alpha)y) \le \alpha f(x) + (1-\alpha) f(y)$$



Gradient Descent

Convergence of GD

One can show that this algorithm :

$$w_{t+1} = w_t - \gamma_t \cdot \nabla f(w_t) \tag{2}$$

converges to some $w_* \in \{\nabla f = 0\}$ for some **smooth functions** f with certain properties and some step size $\gamma_t := \gamma_t(f)$.

In our case, **convexity** is enough, but know that there is a more general class of functions that makes it still true.

Stochastic gradient descent (SGD)

If n is large, computing ∇L is expensive : it requires to to on the whole data set for just a step of the descent algorithm !

The idea of stochastic gradient is to build an **unbiased** estimator of ∇L : Choosing uniformly at random $N \in \{1, \dots, n\}$, then

$$\partial_w \mathbf{E}[\ell(y_N, \langle x_N, w \rangle)] = \frac{1}{n} \sum_{i=1}^{n} \ell`(y_i, \langle x_i, w \rangle) x_i = \nabla L(w)$$

- Question : Can we use this estimator in (2) ?

Convergence of SGD

Theorem : Robbins-Monroe

Let ℓ be defined as previously, and $N_t, t \in \mathbf{N}$ a sequence of *i.i.d.* random variables uniformly distributed on $\{1, \dots, n\}$. The algorithm

$$w_{t+1} = w_t - \gamma_t \cdot \partial_w \ell(y_{N_t}, \langle x_{N_t}, w_t \rangle)$$

converges to some $w_* \in \{\nabla L = 0\}$