CS
AML
2018 - 2019

Stochastic approximation methods

1 Problem formulation

A binary linear classifier can be modeled as a function that predicts the output $y \in \{-1, +1\}$ associated to a given input $\mathbf{x} \in \mathbb{R}^d$. This prediction is defined through a linear combination of the input components, yielding the decision variable $\text{sign}(\mathbf{x}^\top \mathbf{w})$ where $\mathbf{w} \in \mathbb{R}^d$ is the weight vector to be estimated. In supervised learning, this weight vector is determined from a set of input-output pairs

$$\{(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \{-1, +1\} \mid \ell \in \{1, \ldots, n\}\},$$

which is called training set. The learning task is defined through the minimization of a cost function ensuring a trade-off between fitting the training data and reducing the model complexity. Here, we will consider the penalized cost function :

$$(\forall \mathbf{w} \in \mathbb{R}^d) \quad F(\mathbf{w}) = \sum_{i=1}^{n} \log(1 + \exp(-y_i \mathbf{w}^\top \mathbf{x}_i)) + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

with $\lambda > 0$.

2 Batch and online solutions

1. Express the gradient of F, and implement the gradient descent algorithm for minimizing it.
   (**Hint :** The maximum value for the stepsize θ is $\left(\frac{1}{4}\sum_{i=1}^{n} \|\mathbf{x}_i \mathbf{x}_i^\top\| + \lambda\right)^{-1}$).
2. What is the influence of λ parameter on the quality of the classification results ?
3. Show that F can be written in the form :

   $$(\forall \mathbf{w} \in \mathbb{R}^d) \quad F(\mathbf{w}) = \sum_{i=1}^{n} f_i(\mathbf{w})$$

   with each f_i depending only on $(\mathbf{x}_i, y_i)$. Deduce a stochastic gradient algorithm to minimize F, and implement it. Display the evolution of the cost function along iterations, for several strategies for the setting of the learning rate, with and without averaging, and comment the results.

4. Let us set a mini-batch size m, such that $n = km$, with $k \in \mathbb{N}^*$. Show that F can be written in the form :

$$(\forall \mathbf{w} \in \mathbb{R}^d) \quad F(\mathbf{w}) = \sum_{j=1}^{k} F_j(\mathbf{w})$$

with each F_j depending on $\{(\mathbf{x}_i, y_i)\}_{m(j-1)+1 \leq i \leq jm}$. Deduce a mini-batch version of the previous stochastic gradient approach, and implement it. Study the influence of m on the convergence speed.

5. Try the acceleration strategies seen in the course.