Linear Classification - Stochastic Gradient

Abdelouahed Benjelloun

February 2017

1 Introduction

The statistical machine learning approach begins with the collection of a sizable set of examples $(x_1, y_1), ..., (x_N, y_N)$, where for each $i \in 1..n$ the vector x_i represents the features and the scalar y_i a label indicating whether x_i belongs ($y_i = 1$) or not ($y_i = 1$) to a particular class. With such a set of examples, one can construct a classification program, defined by a prediction function h, and measure its performance by counting how often the program prediction h(xi) differs from the correct prediction y_i. To avoid rote memorization, one should aim to find a prediction function that generalizes the concepts that may be learned from the examples. One way to achieve good generalized performance is to choose between a carefully selected class of prediction functions.

Thanks to such a high-dimensional sparse representation of documents, it has been deemed empirically sufficient to consider prediction functions of the form $h(x; w,) =< w, x >$. The risk to be optimized can be written as :

$$R_N(W) = \frac{1}{N} \sum_{1}^{N} (y_i - < w, x_i >)^2$$

2 The algorithm

To get the appropriate $w^\star$ the stochastic gradient descent is a good algorithm that gets a good results faster than the classical optimization algorithms, especially when it's applied to a very large data : N very large.

Algorithm 1 Sctochastic Gradient

1: Initialisation : $w^\star = w_0$
2: *loop* $k = 1 : K$

$$i \sim U([0, N])$$

$$w^\star = w^\star - \alpha_k * (-2y_i * x_i + 2 < x_i, w^\star > x_i)$$

3 Results

In the following exemples we choosed $\alpha_k = \frac{1}{k}$, the index i with a uniform distribution. The ground truth data has been simulated with $W = (1, -1)$

2. Results for 1000 data points without noise, we find $w^\star = (1.1, -1.06)$

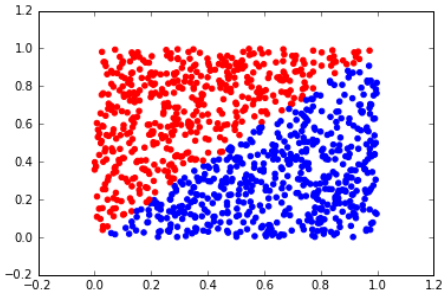

Figure 1: True data

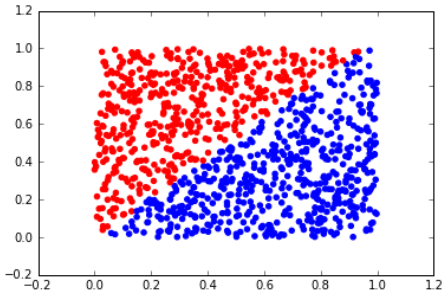

Figure 2: GS results, K=100

3. Results for 1000 data points with gaussien noise : $\epsilon \sim \frac{1}{5} * N(0, 1)$: we find $w^\star = (1.07, -1.02)$

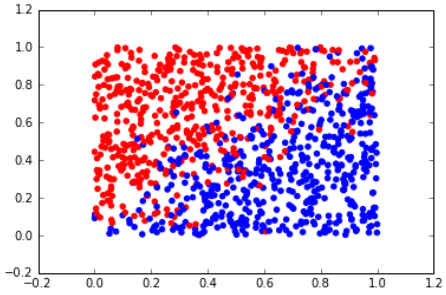

Figure 3: True noisy data

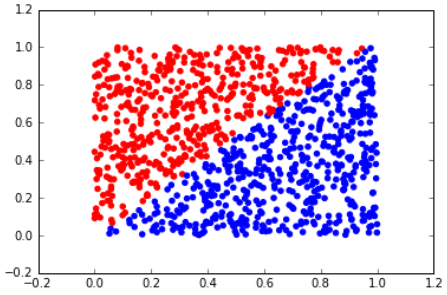

Figure 4: GS results, K=100