ISML-II: Machine Learning **Spring 2014**

Lecture 14- Logistic Regression

Lecturer: Lorenzo Rosasco

We consider logistic regression, that is *Tikhonov* regularization

$$\min_{w \in \mathbb{R}^D} \hat{\mathcal{E}}(f_w) + \lambda \|w\|^2, \quad \hat{\mathcal{E}}(f_w) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_w(x_i)) \tag{14.1}$$

where the loss function is $\ell(y, f_w(x)) = \log(1 + e^{-y f_w(x)})$, namely the logistic loss function.

Since the logistic loss function is differentiable the natural candidate to compute a minimizer is a the gradient descent algorithm which we describe next.

14.1 Interlude: Gradient Descent and Stochastic Gradient

Before starting let's recall the following basic definition

- Gradient of $G : \mathbb{R}^D \to \mathbb{R}$,
$$\nabla G = (\frac{\partial G}{\partial w^1}, \dots, \frac{\partial G}{\partial w^D})$$
- Hessian of $G : \mathbb{R}^D \to \mathbb{R}$,
$$H(G)_{i,j} = \frac{\partial^2 G}{\partial w^i \partial w^j}$$
- Jacobian of $F : \mathbb{R}^D \to \mathbb{R}^D$
$$J(F)_{i,j} = \frac{\partial F^i}{\partial w^j}$$

Note that $H(G) = J(\nabla G)$.

Consider the minimization problem

$$\min_{w \in \mathbb{R}^D} G(w) \quad G : \mathbb{R}^D \to \mathbb{R}$$

when G is a differentiable (strictly convex) function. A general approach to find an approximate solution of the problem is the *gradient descent* (GD) algorithm, based on the following iteration

$$w_{t+1} = w_t - \gamma \nabla G(w_t) \tag{14.2}$$

for a suitable initialization w_0. Above $\nabla G(w)$ is the gradient of G at w and γ is a positive constant (or a sequence) called the step-size. Choosing the step-size appropriately ensures the iteration to converge to a minimizing solution. In particular, a suitable choice can be shown to be

$$\gamma = 1/L,$$

where L is the *Lipschitz constant* of the gradient, that is L such that

$$\|\nabla G(w) - \nabla G(w')\| \leq L\|w - w'\|.$$

It can be shown that L is less or equal than the biggest eigenvalue of the Hessian $H(G)(w)$ for all w. The term descent comes from the fact that it can be shown that

$$G(w_t) \geq G(w_{t+1}).$$

A related technique is called stochastic gradient or also incremental gradient. To describe this method, we consider an objective function is the form

$$G(w) = \sum_{i=1}^n g_i(w), \quad g_i : \mathbb{R}^D \to \mathbb{R}, \quad i = 1, \dots, n,$$

so that $\nabla G(w) = \sum_{i=1}^n \nabla g_i(w)$. The stochast gradient algorithm corresponds to replacing (14.2) with

$$w_{t+1} = w_t - \gamma \nabla g_{i_t}(w_t)$$

where i_t denotes a deterministic or stochastic sequence of indices. In this case, the step size needs to be chosen as sequence γ_t going to zero but not too fast. For example the choice $\gamma_t = 1/t$ can be shown to suffice.

14.2 Regularized Logistic Regression

The corresponding regularized empirical risk minimization problem is called regularized logistic regression. Its solution can be computed via gradient descent or stochastic gradient. Note that

$$\nabla \hat{\mathcal{E}}(f_w) = \frac{1}{n}\sum_{i=1}^n x_i \frac{-y_i e^{-y_i x_i^T w_{t-1}}}{1 + e^{-y_i x_i^T w_{t-1}}} = \frac{1}{n}\sum_{i=1}^n x_i \frac{-y_i}{1 + e^{y_i x_i^T w_{t-1}}}$$

so that, for $w_0 = 0$, the gradient descent algorithm applied to (14.1) is

$$w_t = w_{t-1} - \gamma \left(\frac{1}{n}\sum_{i=1}^n x_i \frac{-y_i}{1 + e^{y_i x_i^T w_{t-1}}} + 2\lambda w_{t-1} \right)$$

for $t = 1, \dots T$, where

$$\frac{1}{n}\sum_{i=1}^n \frac{-y_i x_i e^{-y_i x_i^T w}}{1 + e^{-y_i x_i^T w}} + 2\lambda w = \nabla(\hat{\mathcal{E}}(f_w) + \lambda\|w\|^2)$$

A direct computation shows that

$$J(\nabla \hat{\mathcal{E}}(f_w)) = \frac{1}{n}\sum_{i=1}^n x_i x_i^T \ell''(y_i w^T x_i) + 2\lambda I$$

where $\ell''(a) = \frac{e^{-a}}{(1+e^{-a})^2} \leq 1$ is the second derivative of the function $\ell(a) = \log(1 + e^{-a})$. In particular it can be shown that

$$L \leq \sigma_{\max}(\frac{1}{n}X_n^T X_n + 2\lambda I)$$

where $\sigma_{\max}(A)$ is the largest eigenvalue of a (symmetric positive semidefinite) matrix A.

14.3 Kernel Regularized Logistic Regression

The vector of coefficients can be computed by the following iteration

$$c_t = c_{t-1} - \gamma B(c_{t-1}), \quad t = 1, \dots, T$$

for $c_0 = 0$, and where $B(c_{t-1}) \in \mathbb{R}^n$ with

$$B(c_{t-1})^i = -\frac{1}{n}\frac{y_i}{1 + e^{y_i \sum_{k=1}^n x_k^T x_i c_{t-1}^k}} + 2\lambda c_{t-1}^i.$$

Here again we choose a constant step-size. Note that

$$\sigma_{\max}(\frac{1}{n}X_n^T X_n + \lambda I) = \sigma_{\max}(\frac{1}{n}X_n X_n^T + \lambda I) = \sigma_{\max}(\frac{1}{n}K_n + \lambda I).$$

14.4 Logistic Regression and confidence estimation

We end recalling that a main feature of logistic regression is that, as discussed, The solution of logistic regression can be shown to have probabilistic interpretation, in fact it can be derived from the following model

$$p(1|x) = \frac{e^{x^T w}}{1 + e^{x^T w}}$$

where the right hand side is called logistic function. This latter observation can be used to deduce a confidence from the on each prediction of the logistic regression estimator.