Statistics Learning Theory:Logistic Regression I computational aspect

2019.02.23

Logistic Regression:Sigmoid Function

In this week slide, we will introduce the implementation of Logistic regression and related property.
To understand the method of logistic regression, we first introduce the sigmoid function. That is, the function $\sigma_{sig} : R \to [0, 1]$ over the class of linear functions L_d such that

$$\sigma_{sig}(z) = \frac{1}{1 + exp(-z)}$$

The hypothesis class becomes

$$\mathcal{H} = \sigma_{sig} \circ L_d = \{x \to \sigma_{sig}(\langle w, x \rangle)) : w \in R^d\}$$

Logistic Regression:Loss function

Given the classifier $h_w(x)$, we should define how bad it is to predict some $h_w(x) \in [0, 1]$ given that the true label is $y \in \{1, -1\}$
Therefore, we would like that $h_w(x)$ would be large if $y = 1$ and that $1 - h_w(x)$ would be large if $y = -1$. Since

$$1 - h_w(x) = \frac{1}{1 + exp(\langle w, x \rangle)}$$

Therefore, any resonable loss function would increase monotonically with $\frac{1}{1+exp(y\langle w,x \rangle)}$

Logistic Regression:Loss function

We can choose the log function, that is the loss function

$$l(h_w, (x, y)) = \log(1 + exp(-y\langle w, x\rangle)$$

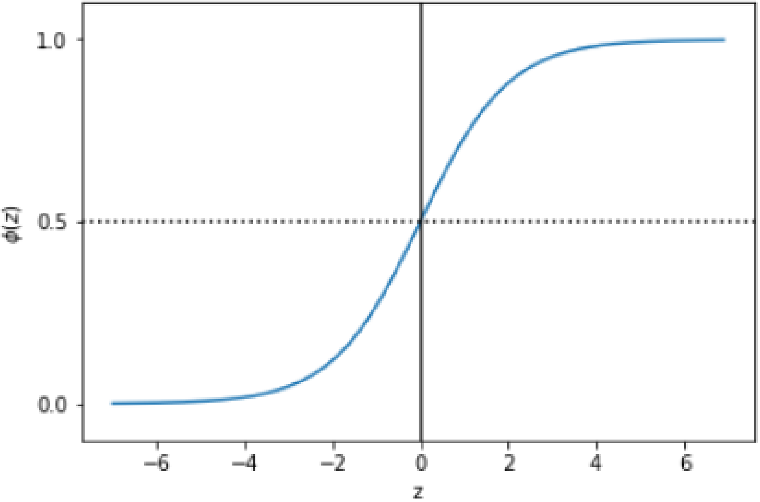
The ERM problem associated with logistic regression is

$$arg \min_{w \in R^d} \frac{1}{m} \sum_{i=1}^{m} \log(1 + exp(-y\langle w, x\rangle)$$

Logistic Regression: Remark

1. logistic loss function is convex function, the optimization can be solved efficienyly
2. The ERM problem associated with logistic regress is identical to the problem of finding a maximimum Likelihood Estimator.
3. One of the efficient algorithm is stochastic gradient descent

Sigmoid function



Gradient Descent:Motivation

The Taylor approximation tells us that

$$f(u) \approx f(w) + \langle u - w, \nabla f(w) \rangle$$

Therefore, given the point $w^{(t)}$, we can update the the next point by minimizing the approximation of $f(u)$, however, when u is far away from w^t, the approximation might become loose. Hence, we jointly minimize the distance between u and $w^{(t)}$ and approximation around $w^{(t)}$. That is

$$w^{(t+1)} = arg \min_u \frac{1}{2}|u - w^{(t)}|^2 + \eta(f(w^{(t)}) + \langle u - w^{(t)}, \nabla f(w^{(t)}) \rangle)$$

Gradient Gescent:Motivation

Taking the derivative with respect to u, we can obtain the following:

$$w^{(t+1)} = w^{(t)} - \eta \nabla f(w^{(t)})$$

The algorithm will update the value in the direction of the greatest rate of increase of f around $w^{(t)}$, η can be thought as learning rate, the rate we believe the approximation part.

Analysis of GD for Convex-Lipschitz Functions

In the GD algorithm, we assume the output is $\bar{w} = \frac{1}{T}\sum_{t=1}^{T} w^{(t)}$ and denote w^* as the minimizer of $f(w)$. By the convexity and the definition of $\bar{w}$, we have

$$f(\bar{w}) - f(w^*) = f(\frac{1}{T}\sum_{t=1}^{T} w^{(t)}) - f(w^*)$$

$$\leq \frac{1}{T}\sum_{t=1}^{T}(f(w^{(t)}) - f(w^*)$$

$$= \frac{1}{T}\sum_{t=1}^{T}(f(w^{(t)}) - f(w^*))$$

because of the convexity of f, we have that

$$f(w^{(t)}) - f(w^*) \leq \langle w^{(t)} - w^*, \nabla f(w^{(t)})\rangle$$

Therefore

$$f(\bar{w}) - f(w^*) \leq \frac{1}{T}\sum_{t=1}^{T}\langle w^{(t)} - w^*, \nabla f(w^{(t)})\rangle$$

Analysis of GD for Convex-Lipschitz Functions

To estimate the converge rate of GD algorithm, we claim the following lemma

Theorem

Let $v_1, \cdots, v_T$ be an arbitrary sequence of vectors. Any algorithm with an initialization $w^{(1)} = 0$ and an update rule of the form

$$w^{(t+1)} = w^{(t)} - \eta v_t$$

satisfies

$$\sum_{t=1}^{T} \langle w^{(t)} - w^*, v_t \rangle \leq \frac{|w^*|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} |v_t|^2$$

In particular, if $|v_t| \leq \rho$ and $|w^*| \leq B$ then we set $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$ then

$$\frac{1}{T} \sum_{t=1}^{T} \langle w^{(t)} - w^*, v_t \rangle \leq \frac{B\rho}{\sqrt{T}}$$

Analysis of GD for Convex-Lipschitz Functions

Using albegraic manipulations, we have

$$\langle w^{(t)} - w^*, v_t \rangle = \frac{1}{\eta} \langle w^{(t)} - w^*, \eta v_t \rangle$$

$$= \frac{1}{2\eta} (-|w^{(t)} - w^* - \eta v_t|^2 + |w^{(t)} - w^*|^2 + \eta^2 |v_t|^2)$$

$$= \frac{1}{2\eta} (-|w^{(t+1)} - w^*|^2 + |w^{(t)} - w^*|^2) + \frac{\eta}{2} |v_t|^2$$

Summing the equality over t, we have

$$\sum_{t=1}^{T} \langle w^{(t)} - w^*, v_t \rangle = \frac{1}{2\eta} \sum_{t=1}^{T} (-|w^{(t+1)} - w^*|^2 + |w^{(t)} - w^*|^2) + \frac{\eta}{2} |v_t|^2$$

Analysis of GD for Convex-Lipschitz Functions

The first part is a telescopic sum equal to

$$|w^{(1)} - w^*|^2 - |w^{(T+1)} - w^*|^2$$

Therefore

$$sum_{t=1}^{T} \langle w^{(t)} - w^*, v_t \rangle = \frac{1}{2\eta}(|w^{(1)} - w^*|^2 - |w^{(T+1)} - w^*|^2) + \frac{\eta}{2} \sum_{t=1}^{T} |v_t|^2$$

$$\leq \frac{1}{2\eta}|w^{(1)} - w^*|^2 + \frac{\eta}{2} \sum_{t=1}^{T} |v_t|^2$$

$$= \frac{1}{2\eta}|w^*|^2 + \frac{\eta}{2} \sum_{t=1}^{T} |v_t|^2$$

Gereralized:Subgradients

We want to generalize to non-differentable functions. To do this, we use subgradients instead. Here, we recall the definition of subgradient.

Definition

A vector v that satisfies

$$\forall u \in S, \quad f(u) \geq f(w) + \langle u - w, v \rangle$$

is called a subgradient of f at w. The set of subgradients of f at w is called the differential set and denoted $\partial f(w)$. Note that if f is convex, such v must exist.

Stochastic Gradient Descent

Algorithm 1: Stochastic Gradient Descent(SGD)

Input :$\leftarrow$ Scalar $\eta > 0$, integer $T > 0$;

Initialize :$\leftarrow$ $w^{(1)} = (0, 0, \cdots, 0)$;

for $t = 1, 2, \cdots T$ **do**

  choose v_t at random from a distribution such that $E[v_t | w^{(t)}] \in \partial f(w^{(t)})$

  update $w^{(t+1)} = w^{(t)} - \eta v_t$;

end for

Output : $\bar{w} = \frac{1}{T} \sum_{t=1}^{T} w^{(t)}$

Stochastic Gradient Descent:theorem

Theorem

Let $B, \rho > 0$. Let f be a convex function and let $w^* \in arg\min_{w:|w|\leq B} f(w)$. Assume that SGD is run for T iterations with $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$. Assume also that for all t, $|v_t| \leq \rho$ with probability 1. Then,

$$E[f(\bar{w})] - f(w^*) \leq \frac{b\rho}{\sqrt{T}}$$