Stanford CS 229 - Machine Learning - Lecture 4

Giovani da Silva

June 2021

1 Newton's Method

Often runs much faster than gradient ascent.

Newton's Methods is going to take the tangent to the function at that point there to zero, and is going to sort of extend this tangent down until it intercepts the horizontal axis.To see what value this is, this is theta one,then we have one iteration of Newton's Method.

And what I'll do then is the same thing with the requested point. Take the tangent down here, and that's two iterations of the algorithm. And keep going, that's theta three and so on.

This is an algorithm for finding a value of theta for which F of theta equals zero.

$\Delta = \frac{f(\theta_0)}{f'(\theta_0)}$ Δ is the distance between θ_x and the next $\theta_x + 1$

So, one iteration of Newton's Method is:

$\theta_{t+1} = \theta_t - \frac{f(\theta_t)}{f'(\theta_t)}$

we want to maximize l(θ) , so l(θ)'= 0

$\theta_{t+1} = \theta_t - \frac{l'(\theta_t)}{l''(\theta_t)}$

A generalization when θ is a vector: $\theta_{t+1} = \theta_t + H_\theta^{-1} l$

Where H is the Hessian Matrix

$H_{ij} = \frac{\partial^2 l}{\partial \theta_i \partial \theta_j}$

Usually needs far fewer interations than gradient descent.The disadvantage of Newton's Method is that on every iteration you need to invert the Hessian.Hessian will be an N-by-N matrix, or an N plus one by N plus one-dimensional matrix if N is the number of features. And so if you have a large number of features in your learning problem, if you have tens of thousands of features, then inverting H could be a slightly computationally expensive step.

2 Generalized linear models

Assume:

1.$y|x : \theta \sim ExpFamily(\eta)$

2.Given x, E$[T(y)|x]$ our goal is to get our learning algorithms output or

h(x) = E$[T(y)|x]$

3.The relationship is linear or $\eta = \theta^T x$

3 Softmax regression

Softmax regression (or multinomial logistic regression) is a generalization of logistic regression to the case where we want to handle multiple classes. In logistic regression we assumed that the labels were binary: $y(i) \in \{0, 1\}$. Softmax regression allows us to handle $y(i) \in \{1, \ldots, K\}$ where K is the num-

ber of classes. The hypothesis takes form as: $h_\theta(x) = \begin{bmatrix} P(y=1|x;\theta) \\ P(y=2|x;\theta) \\ \vdots \\ P(y=K|x;\theta) \end{bmatrix} =$

$\frac{1}{\sum_{j=1}^{K} \exp(\theta^{(j)\top} x)} \begin{bmatrix} \exp(\theta^{(1)\top} x) \\ \exp(\theta^{(2)\top} x) \\ \vdots \\ \exp(\theta^{(K)\top} x) \end{bmatrix}$