Problem Set 1

08/04/2017

1 Problem 1

1.1

We can form a cost function for logistic regression as

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\log\left(1+\exp\left(-y^{(i)}\theta^T x^{(i)}\right)\right)$$

where $y^{(i)} \in \{0,1\}$ is the response, $x^{(i)}$ is the vector of the ith sample's feature values, and θ is the vector of feature coefficients. m is the number of samples in the dataset. This question asked us to find the Hessian of J w.r.t θ, and show that it is positive semi-definite ($z^T Hz \geq 0 \;\forall\; z \in \mathbb{R}^n$).

By differentiating J w.r.t. each feature in x, I found that

$$\begin{aligned}\nabla_\theta J &= -\frac{1}{m}\sum_{i=1}^{m}\frac{y^{(i)}x^{(i)}}{1+\exp\left(y^{(i)}\theta^T x^{(i)}\right)}\\ H &= \frac{1}{m}\sum_{i=1}^{m}\frac{y^{(i)2}x^{(i)}x^{(i)T}}{(1+\exp\left(y^{(i)}\theta^T x^{(i)}\right))^2}\end{aligned} \tag{1}$$

From this we can see that the Hessian is indeed symmetric, and that all its elements are positive. By expanding the Hessian into the condition for positive semi-definiteness

$$z^T Hz \;=\; \frac{1}{m}\sum_{i=1}^{m}\frac{y^{(i)2}}{(1+\exp\left(y^{(i)}\theta^T x^{(i)}\right))^2}\; z^T x^{(i)}x^{(i)T}z$$

The fraction here will always be positive, as is the square on the right - $(x^T z)^2$. This means that we have have a sum of zero (for $z = \vec{0}$) or positive terms, thereby meaning that H is PSD.

1.2

After deriving the gradient and Hessian in part 1, we were then asked to apply what was found to fit logistic regression to some synthetic data. To make the fit, we minimized the cost function using Newton-Raphson. This optimization took the form

$$\theta_{j+1} := \theta_j + H^{-1}\nabla J$$

The coefficients I found from the fit were $\theta = (2.9, -0.18)$.

2 Problem 2

2.1

We were asked to show that that the Poisson distribution,

$$p(y;\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}$$

is a member of the Exponential Family. These are functions that have the form

$$p(y;\eta) = b(y)\exp\left(\eta^T T(y) - a(\eta)\right)$$

It's clear from inspection that

$$\begin{aligned} p(y;\lambda) &= \frac{1}{y!}\exp\left(y\log\lambda - \lambda\right) \\ : \quad b(y) &= \frac{1}{y!} \quad ; \quad \eta = \log\lambda \\ a(\eta) &= e^{\eta} \quad ; \quad T(y) = y \end{aligned} \tag{2}$$

2.2

The canonical response function for a generalized linear model is defined as the mean of $T(y)$ parameterized by η. For the Poisson distribution, this means the c.r.f. is the expectation of the response parameterized by η.

The expectation of the Poisson distribution is λ, which we can see from the part 1 is equal to e^{η}. It follows that if we assume that $\eta = \theta^T x$ - the linear assumption - then the canonical response function $E[y;\eta]$ is

$$g(\eta) = E[y;\eta] = E[y;\theta^T x] = e^{\theta^T x}$$

2.3

Deriving the stochastic gradient ascent rule is now straightforward, since the log-likelihood of a single example will be

$$\begin{aligned} \log l(\theta)^{(i)} &= \log p(y^{(i)}|x^{(i)};\theta) \\ &= \log\frac{1}{y^{(i)}!} + \theta^T x^{(i)} - \exp\left(\theta^T x^{(i)}\right) \\ : \nabla_\theta \log l(\theta)^{(i)} &= y^{(i)}x^{(i)} - x^{(i)}\exp(\theta^T x^{(i)}) \end{aligned} \tag{3}$$

Hence the stochastic gradient ascent rule, which will maximize the log-likelihood across all parameters incrementally, will be

$$\{ \quad \theta := \theta + \alpha \cdot \nabla_\theta \log l(\theta) \quad \} \quad \text{for each } i \in \{1, ..., m\}$$