CS229 Fall 2017, Problem Set #1: Supervised Learning

Armand Sumo – armandsumo@gmail.com

April 28, 2021

Collaborators:

By turning in this assignment, I agree by the Stanford honor code and declare that all of this is my own work.

---

1. Logistic regression

Average empirical loss for logistic regression:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} log(h_\theta(y^i x^i))$$

where $y^{(i)} \in \{-1, 1\}$, $h_\theta(x)) = g(\theta^T x)$ and $g(z) = 1/(1 + e^{-z})$

(a)

$$\begin{aligned}
\nabla_\theta J(\theta) &= -\frac{1}{m}\sum_{i=1}^{m} \frac{1}{g(\theta^T y^{(i)} x^{(i)})} \nabla_\theta g(\theta^T y^{(i)} x^{(i)}) \\
&= -\frac{1}{m}\sum_{i=1}^{m} \frac{1}{g(\theta^T y^{(i)} x^{(i)})} y^{(i)} x^{(i)} g(\theta^T y^{(i)} x^{(i)})(1 - g(\theta^T y^{(i)} x^{(i)})) \\
&= -\frac{1}{m}\sum_{i=1}^{m} \frac{1}{y}^{(i)} x^{(i)}(1 - g(\theta^T y^{(i)} x^{(i)}))
\end{aligned}$$

$$\begin{aligned}
H_{i,j} = \frac{\partial}{\partial \theta_j}\left[\nabla_\theta J(\theta)\right]_i &= \frac{1}{m}\sum_{i=1}^{m} (y^{(i)})^2 x_j^{(i)} x_i^{(i)} g(\theta^T y^{(i)} x^{(i)})(1 - g(\theta^T y^{(i)} x^{(i)})) \\
&= \frac{\partial}{\partial \theta_j}\left[\nabla_\theta J(\theta)\right]_i && \text{H is symmetric}
\end{aligned}$$

Let's show that for any vector z,
$z^T Hz \geq 0$

$$\sum_i \sum_j z_i x_i x_j z_j = \sum_i z_i x_i \sum_j z_j x_j = (x^T z)(x^T z) = (x^T z)^2 \geq 0$$

$$\begin{aligned}
z^T Hz &= \sum_i z_i^T (Hz)_i = \sum_i \sum_j z_i (H_{i,j}) z_j \\
&= \sum_i \sum_j z_i \Big(\frac{1}{m} \sum_{k=1}^{m} (y^{(k)})^2 x_j^{(k)} x_i^{(k)} g(\theta^T y^{(k)} x^{(k)})(1 - g(\theta^T y^{(k)} x^{(k)}))\Big) z_j \\
&= \frac{1}{m} \sum_{k=1}^{m} \sum_i \sum_j (y^{(k)})^2 z_j x_j^{(k)} z_i x_i^{(k)} g(\theta^T y^{(k)} x^{(k)})(1 - g(\theta^T y^{(k)} x^{(k)})) \\
&= \frac{1}{m} \sum_{k=1}^{m} (y^{(k)})^2 g(\theta^T y^{(k)} x^{(k)})(1 - g(\theta^T y^{(k)} x^{(k)}))((x^{(k)})^T z)^2
\end{aligned}$$

For any vector z, $g(z) \in [0, 1]$, hence $z^T Hz \geq 0$.
This implies that H is positive semi-definite, therefore J is convex and has no local minima other than the global one.

(b) After implementing Newton's method for optimizing $J(\theta)$ and applying it to fit a logistic regression model to the data, I obtained a parameter vector:
$\theta = [-2.61847133, 0.75979248, 1.1707512]^T$.

(c)

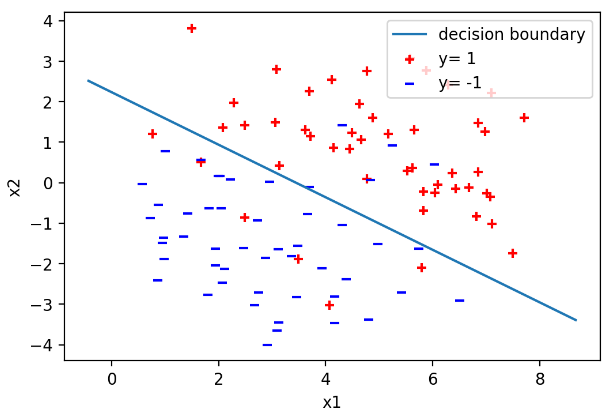

Figure 1: Training data and decision boundary fit by logistic regression

2. Poisson regression and the exponential family

(a) We consider the Poisson distribution parametrized by λ:

$$p(y;\lambda) = \frac{e^{-y}\lambda^{y}}{y!} = \frac{\exp(y\log(\lambda) - \lambda)}{y!} = b(y)(\exp(\eta^T T(y) - a(\eta))$$

The Poisson distribution is in the exponential family, with:

$$\begin{aligned} b(y) &= 1 \\ \eta &= \log(\lambda) \\ T(y) &= y \\ a(\eta) &= \lambda = e^{\eta} \end{aligned}$$

(b) We want to perform regression using a GLM model with a Poisson response variable. To construct the GLM model, we make the following assumptions: - $y|x;\theta \sim \text{ExponentialFamily}(\eta)$
- our goal is to predict the expected value of $T(y)$ given x. Because T(y)=y, this means we would like the hypothesis $h_\theta(x)$ to satisfy: $h_\theta(x) = \mathbb{E}\left[y|x\right]$
- The natural parameter η and the inputs x are related linearly $y = \theta^T x$ It follows that our hypothesis will output:

$$h_\theta(x) = \mathbb{E}\left[y|x\right] = \lambda = e^{\eta} = e^{\theta^T x}$$

Therefore, the canonical response of this family is $g(z) = h(\theta^T z) = e^z$.

(c) Our model assumes that the conditional probability of y given x is:
$p(y^{(i)}|x^{(i)};\theta) = \dfrac{\exp(y^{(i)}\theta^T x^{(i)} - e^{\theta^T x^{(i)}})}{y^{(i)}!}$
We now maximize the likelihood $L(\theta)$ of our parameter θ using gradient ascent.

$$\begin{aligned} \ell(\theta) &= \log(L(\theta)) = \log(p(y^{(i)}|x^{(i)};\theta)) = y^{(i)}\theta^T x^{(i)} - e^{\theta^T x^{(i)}} - \log(y^{(i)}!) \\ \frac{\partial \ell(\theta)}{\partial \theta_j} &= y^{(i)} x_j^{(i)} e^{\theta^T x^{(i)}} = x_j^{(i)}(y^{(i)} - e^{\theta^T x^{(i)}}) \end{aligned}$$

We obtain the following stochastic gradient ascent update rule:
$\theta_j := \theta_j + \alpha x_j^{(i)}(y^{(i)} - h_\theta(x^{(i)}))$ with $h_\theta(x) = e^{\theta^T x}$

(d) We now use GLM for any member of the exponential family for which $T(y) = y$, and the canonical response $h(x)$ for the family. From our model's assumptions,

$$\begin{aligned} p(y|X;\theta) &= b(y)(\exp(\eta^T T(y) - a(\eta)) = b(y)(\exp(\eta^T y - a(\eta)) \\ \ell(\theta) &= \log p(y|X;\theta) = \eta^T y - a(\eta) + log(b(y)) \end{aligned}$$

For a single parameter θ_i,

$$\frac{\partial \ell(\theta)}{\partial \theta_i} = \frac{\partial}{\partial \theta_i}(\theta^T x)^T y - \frac{\partial}{\partial \theta_i} a(\theta^T x)$$

To determine $a(\eta)$, we use the fact that for $p(y|X;\theta)$ to be a pdf, it must integrate to 1.

$$\begin{aligned}
\int_y p(y|X;\theta)\,dy &= 1 \\
\int_y b(y)(\exp(\eta^T T(y) - a(\eta))\,dy &= 1 \\
e^{a(\eta)} &= \int_y b(y)\exp(\eta^T y))\,dy \\
a(\eta) &= log \int_y b(y)\exp(\eta^T y))\,dy
\end{aligned}$$

Let f be a differentiable function such that $a(\eta) = \log f(\eta)$. Using the chain rule, $\frac{\partial a(\eta)}{\partial \eta} = \frac{\partial \log f(\eta)}{\partial \eta} = \frac{\partial f(\eta)}{\partial \eta}\frac{1}{f(\eta)}$. Hence,

$$\begin{aligned}
\frac{\partial a(\eta)}{\partial \eta} &= \frac{1}{\int_y b(y)\exp(\eta^T y))\,dy} \int_y b(y)\frac{\partial \exp(\eta^T y)}{\partial \eta}\,dy \\
&= \frac{1}{\int_y b(y)\exp(\eta^T y)\,dy} \int_y b(y)\exp(\eta^T y)\frac{\partial \eta^T y}{\partial \eta}\,dy \\
\frac{\partial a(\theta^T x)}{\partial \theta_i} &= \frac{1}{\int_y b(y)\exp(\eta^T y))\,dy} \int_y b(y)\exp(\eta^T y)\frac{\partial x^T \theta y}{\partial \theta_i}\,dy \\
&= \frac{1}{\int_y b(y)\exp(\eta^T y))\,dy} \int_y b(y)\exp(\eta^T y) x_i y\,dy \\
&= \int_y \frac{b(y)\exp(\eta^T y)\,dy}{\int_y b(y)\exp(\eta^T y)\,dy} x_i y\,dy \\
&= \int_y b(y)\frac{\exp(\eta^T y)}{\exp(a(\eta))} x_i y\,dy \\
&= \int_y p(y|X;\theta) x_i\,dy \\
&= xi \int_y p(y|X;\theta)\,dy = x_i \mathbb{E}\left[y|x;\theta\right] = x_i h_\theta(x)
\end{aligned}$$

It follows that:

$$\begin{aligned}
\frac{\partial \ell(\theta)}{\partial \theta_i} &= \frac{\partial}{\partial \theta_i}(\theta^T x)^T y - \frac{\partial}{\partial \theta_i} a(\theta^T x) \\
&= x_i y - x_i h_\theta(x) = x_i(y - h_\theta(x))
\end{aligned}$$

Therefore, the stochastic gradient ascent on the log likelihood of $p(y|X;\theta)$ results in the update rule:

$$\theta_i := \theta_i - \alpha(h_\theta(x) - y)x_i$$

5. Regression for denoising quasar spectra

(a) Locally weighted linear regression
We want to minimize

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m} w^{(i)}(\theta^T x^{(i)} - y^{(i)})^2$$

where $w^{(i)}$ is the weight for a training example (i).
Let X be the m-by-$d+1$ design matrix that contains the training examples' input values in its rows and y be an m-dimensional vector containing all the target values from the training set: $X = \begin{bmatrix} — (x^{(1)})^T — \\ — (x^{(2)})^T — \\ \vdots \\ — (x^{(m)})^T — \end{bmatrix}$; $y = \begin{bmatrix} — y^{(1)} — \\ — y^{(2)} — \\ \vdots \\ — y^{(m)} — \end{bmatrix}$

(i)

$$\begin{aligned} (X\theta - y)_j &= (x^{(j)})^T\theta - y^{(j)} \\ [W(X\theta - y)]_i &= W_i(X\theta - y) = \sum_{i=1}^{m} W_{i,j}(x^{(j)})^T\theta - y^{(j)} \\ (X\theta - y)_i^T &= (x^{(i)})^T\theta - y^{(i)} \end{aligned}$$

$$\begin{aligned} (X\theta - y)^T W(X\theta - y) &= \sum_{i=1}^{m}(X\theta - y)_i^T\, [W(X\theta - y)]_i \\ &= \sum_{i=1}^{m}\big((x^{(i)})^T\theta - y^{(i)}\big)\left(\sum_{i=1}^{m} W_{i,j}(x^{(j)})^T\theta - y^{(j)}\right) \end{aligned}$$

Let

$$W = \frac{1}{2}\begin{bmatrix} w^{(n)} & & \dots(0) \\ \vdots & \ddots & \\ (0) & & w^{(m)} \end{bmatrix}$$

Then,

$$W_{i,j} = \begin{cases} \frac{w^{(i)}}{2} & i = j \\ 0 & i \neq j \end{cases}$$

Hence,

$$\begin{aligned}
(X\theta - y)^T W (X\theta - y) &= \sum_{i=1}^{m} \big((x^{(i)})^T\theta - y^{(i)}\big)\big(\frac{w^{(i)}}{2}((x^{(i)})^T\theta - y^{(i)})\big) \\
&= \frac{1}{2}\sum_{i=1}^{m} w^{(i)}((x^{(i)})^T\theta - y^{(i)})\big)^2 \\
&= J(\theta)
\end{aligned}$$