Q1 Newton's method for computing least squares

Prove that if we use Newton's method to solve the least squares optimization problem, then we only need one iteration to converge to θ^*.

a. Find the Hessian of the cost function $J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(\theta^T x^{(i)} - y^{(i)})^2$.

We know that

$$\theta^T = \begin{bmatrix}\theta_0 & \theta_1 & \cdots & \theta_d\end{bmatrix} \qquad x = \begin{bmatrix} x_0^{(i)} \\ x_1^{(i)} \\ \vdots \\ x_d^{(i)} \end{bmatrix} \text{ where } \forall i, \quad x_0^i = 1$$

Then each entry at index i, j of the Hessian is:

$$\begin{aligned}
H_{ij}(J(\theta)) &= \frac{\partial}{\partial\theta_i\partial\theta_j}\frac{1}{2}\sum_{k=1}^{m}(\theta^T x^{(k)} - y^{(k)})^2 \\
&= \frac{\partial}{\partial\theta_j}\sum_{k=1}^{m}(\theta^T x^{(k)} - y^{(k)})x_i^{(k)} \\
&= \frac{\partial}{\partial\theta_j}\sum_{k=1}^{m}(\theta_0^{(k)}x_0^{(k)}x_i^{(k)} + \cdots + \theta_d^{(k)}x_d^{(k)}x_i^{(k)} - y^{(k)}x_i^{(k)}) \\
&= \sum_{k=1}^{m}(x_i^{(k)}x_j^{(k)})
\end{aligned}$$

Thus

$$H = \begin{bmatrix}
\sum_{k=1}^m x_0^{(k)}x_0^{(k)} & \sum_{k=1}^m x_0^{(k)}x_1^{(k)} & \cdots & \sum_{k=1}^m x_0^{(k)}x_d^{(k)} \\
\sum_{k=1}^m x_1^{(k)}x_0^{(k)} & \sum_{k=1}^m x_1^{(k)}x_1^{(k)} & \cdots & \sum_{k=1}^m x_1^{(k)}x_d^{(k)} \\
\vdots & \vdots & \ddots & \vdots \\
\sum_{k=1}^m x_d^{(k)}x_0^{(k)} & \sum_{k=1}^m x_d^{(k)}x_1^{(k)} & \cdots & \sum_{k=1}^m x_d^{(k)}x_d^{(k)}
\end{bmatrix} = X^T X$$

where

$$X = \begin{bmatrix} \vec{x}^{(1)} \\ \vec{x}^{(2)} \\ \vdots \\ \vec{x}^{(m)} \end{bmatrix}$$

b. Show that the first iteration of Newton's method gives us the optimal solution to least squares problem.

Newton's method: $\theta := \theta - H^{-1}\nabla_\theta J(\theta)$

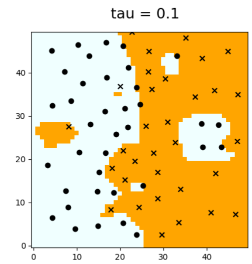
It is known that $\nabla_\theta J(\theta) = X^T X\theta - X^T\vec{y}$, then,

$$\begin{aligned}
\theta - H^{-1}\nabla_\theta J(\theta) =&\theta - (X^T X)^{-1}(X^T X\theta - X^T\vec{y}) \\
=&\theta - (X^T X)^{-1}X^T X\theta + (X^T X)^{-1}X^T\vec{y} \\
=&\theta - \theta + (X^T X)^{-1}X^T\vec{y} \\
=&(X^T X)^{-1}X^T\vec{y}
\end{aligned}$$

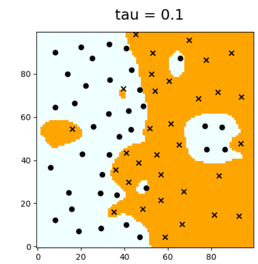
which is the optimal solution to the least squares problem.

Q2 Locally-weighted Logistic regression

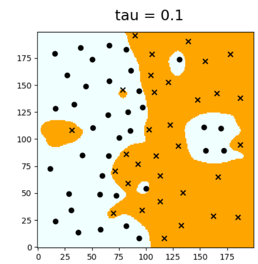
Implement the Newton-Raphson algorithm for optimizing $\ell(\theta)$, where $\ell(\theta)$ is cost function for locally weighted logistic regression, for a new query point x, and use this to predict the class of x.



(a) Resolution = 50

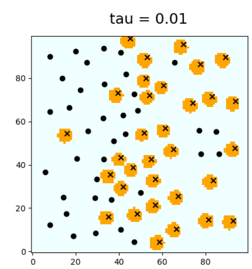


(b) Resolution = 100

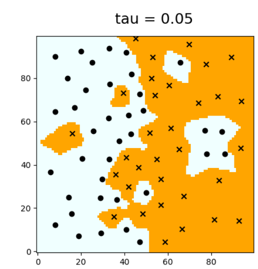


(c) Resolution = 200

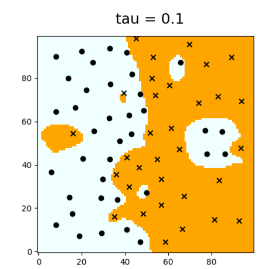
Figure 1: Results when varying the resolution



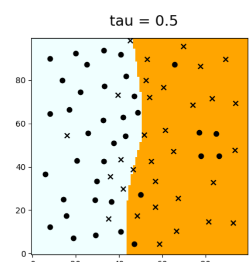
(a) Resolution = 100; $\tau = 0.01$



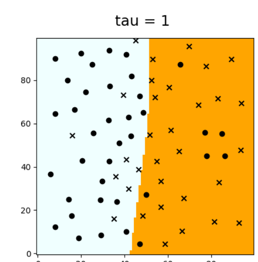
(b) Resolution = 100; $\tau = 0.05$



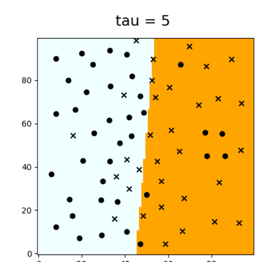
(c) Resolution = 100; $\tau = 0.1$



(d) Resolution = 100; $\tau = 0.5$



(e) Resolution = 100; $\tau = 1$



(f) Resolution = 100; $\tau = 5$

Figure 2: Varying $\tau = 0.01, 0.05, 0.1, 0.5, 1, 5$

As we increase the bandwidth parameter τ, the decision boundary approaches a straight line. As $\tau \to \infty$, the model approaches an unweighted logistic regression model.

3. Multivariate least squares

a. Simplify to matrix-vector notation:

$$J(\Theta) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{p}((\Theta^T x^{(i)})_j - y_j^{(i)})^2$$

where $\Theta \in \mathbb{R}^{n \times p}$.

Let $X = \begin{bmatrix} x^{(1)^T} \\ x^{(2)^T} \\ \vdots \\ x^{(m)^T} \end{bmatrix}$, $x^{(i)} \in \mathbb{R}^n$, and $Y = \begin{bmatrix} y^{(1)^T} \\ y^{(2)^T} \\ \vdots \\ y^{(m)^T} \end{bmatrix}$, $y^{(i)} \in \mathbb{R}^p$.

Let $\Theta = \begin{bmatrix} \theta_1 & \theta_2 & \cdots & \theta_p \end{bmatrix}$, where $\theta_i \in \mathbb{R}^n$.

Then

$$\begin{aligned} X\Theta - Y &= \begin{bmatrix} x_{(1)} \cdot \theta_1 - y_1^{(1)} & x_{(1)} \cdot \theta_2 - y_2^{(1)} & \cdots & x_{(1)} \cdot \theta_p - y_p^{(1)} \\ x_{(2)} \cdot \theta_1 - y_1^{(2)} & x_{(2)} \cdot \theta_2 - y_2^{(2)} & \cdots & x_{(2)} \cdot \theta_p - y_p^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{(n)} \cdot \theta_1 - y_1^{(m)} & x_{(m)} \cdot \theta_2 - y_2^{(m)} & \cdots & x_{(m)} \cdot \theta_p - y_p^{(m)} \end{bmatrix} \\ &= \begin{bmatrix} z_1 & z_2 & \cdots & z_p \end{bmatrix} \end{aligned}$$

Notice that $\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{p}((\Theta^T x^{(i)})_j - y_j^{(i)})^2$ is equal to the sum of the square of every entry in $X\Theta - Y$. We can calculate that by taking the dot product of each column of $X\Theta - Y$ by itself, then sum them up. First,

$$(X\Theta - Y)^T(X\Theta - Y) = \begin{bmatrix} z_1 \cdot z_1 & z_1 \cdot z_2 & \cdots & z_1 \cdot z_p \\ z_2 \cdot z_1 & z_2 \cdot z_2 & \cdots & z_2 \cdot z_p \\ \vdots & \vdots & \ddots & \vdots \\ z_p \cdot z_1 & z_p \cdot z_2 & \cdots & z_p \cdot z_p \end{bmatrix}$$

Thus

$$J(\Theta) = \frac{1}{2}\operatorname{tr}\{(X\Theta - Y)^T(X\Theta - Y)\}$$

Or

$$J(\Theta) = \frac{1}{2}\operatorname{tr}\{(X\Theta - Y)(X\Theta - Y)^T\}$$

b. Find the closed form solution for Θ which minimizes $J(\Theta)$.

$$
\begin{aligned}
J(\Theta) &= \frac{1}{2}\operatorname{tr}[(X\Theta - Y)(X\Theta - Y)^T] \\
&= \frac{1}{2}\operatorname{tr}[(X\Theta - Y)(\Theta^T X^T - Y^T)] \\
&= \frac{1}{2}\operatorname{tr}[X\Theta\Theta^{\mathrm{T}}X^{\mathrm{T}} - Y\Theta^{\mathrm{T}}X - X\Theta Y^{\mathrm{T}} + YY^{\mathrm{T}}] \\
&= \frac{1}{2}\{\operatorname{tr}[X\Theta\Theta^{\mathrm{T}}X^{\mathrm{T}} - \operatorname{tr}[Y\Theta^{\mathrm{T}}X^{\mathrm{T}}] - \operatorname{tr}[X\Theta Y^{\mathrm{T}}] + \operatorname{tr}[YY^{\mathrm{T}}]\}
\end{aligned}
$$

1.) Consider $f = \operatorname{tr}[AXB]$

$$
\begin{aligned}
f = \sum_i [AXB]_{ii} &= \sum_i \sum_j A_{ij}[XB]_{ji} \\
&= \sum_i \sum_j A_{ij} \sum_k X_{jk}B_{ki} \\
&= \sum_i \sum_j \sum_k A_{ij}X_{jk}B_{ki} \\
\therefore \frac{\partial f}{\partial X_{jk}} &= \sum_i A_{ij}B_{ki} = \sum_i B_{ki}A_{ij} = [BA]_{kj} \\
\therefore \frac{\partial \operatorname{tr}[AXB]}{\partial X} &= (BA)^{\mathrm{T}} = A^{\mathrm{T}}B^{\mathrm{T}} \qquad (1)
\end{aligned}
$$

2.) Consider $f = \operatorname{tr}[AX^{\mathrm{T}}B]$

$$
\begin{aligned}
f &= \sum_i \sum_j \sum_k A_{ij}X_{kj}B_{ki} \\
\frac{\partial f}{\partial X_{kj}} &= \sum_i A_{ij}B_{ki} = [BA]_{kj} \\
\therefore \frac{\partial \operatorname{tr}[AX^{\mathrm{T}}B]}{\partial X} &= BA \qquad (2)
\end{aligned}
$$

Now taking the partial derivatives with respect to X in each term of $J(\Theta)$,

$$
\begin{aligned}
\frac{\partial\,\mathrm{tr}[X\Theta\Theta^{\mathrm{T}}X^{\mathrm{T}}]}{\partial\Theta} &= \frac{\partial\,\mathrm{tr}[X\Theta A]}{\partial\Theta} + \frac{\partial\,\mathrm{tr}[B\Theta^{\mathrm{T}}X^{\mathrm{T}}]}{\partial X} \\
&\text{where } A = \Theta^T X^T \text{ and } \Theta \text{ is constant, and } B = X\Theta \text{ where } \Theta \text{ is constant.} \\
&= (AX)^{\mathrm{T}} + X^{\mathrm{T}}B && \text{by (1) and (2)} \\
&= X^{\mathrm{T}}X\Theta + X^{\mathrm{T}}X\Theta \\
&= 2X^{\mathrm{T}}X\Theta \\
\frac{\partial\,\mathrm{tr}[Y\Theta^{\mathrm{T}}X^{\mathrm{T}}]}{\partial\Theta} &= X^{\mathrm{T}}Y && \text{by (2)} \\
\frac{\partial\,\mathrm{tr}[X\Theta Y^{\mathrm{T}}]}{\partial\Theta} &= (Y^{\mathrm{T}}X)^{\mathrm{T}} = X^{\mathrm{T}}Y && \text{by (1)} \\
\frac{\partial\,\mathrm{tr}[YY^{\mathrm{T}}]}{\partial\Theta} &= 0 \\
\therefore \frac{\partial J(\Theta)}{\partial\Theta} &= \frac{1}{2}\{2X^{\mathrm{T}}X\Theta - X^{\mathrm{T}}Y - X^{\mathrm{T}}Y + 0\} \\
&= X^{\mathrm{T}}X\Theta - X^{\mathrm{T}}Y
\end{aligned}
$$

Setting $J(\Theta) = 0$, then we have

$$\Theta = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$$

which looks like the solution for regular least squares:

$$\theta = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\vec{y}$$

c. Suppose instead of considering the multivariate vectors $y^{(i)}$ all at once we instead compute each variable $y_j^{(i)}$ separately for each $j = 1, \ldots, p$. In this case we have p individual linear models of the form

$$y_j^{(i)} = \theta_j^{\mathrm{T}} x^{(i)}, j = 1, \ldots, p.$$

How do the parameters from these p independent least squares problem compare to the multivariate solution?

The optimum solutions from each independent linear model from $j = 1, \ldots, p$ will be column vectors from the multivariate solution Θ.

Q5 Exponential family and the geometric distribution