1 Newton's method for computing least squares

a) The Hessian of a function $J(\theta)$ is a matrix H such that $H_{i,j} = \frac{\partial J(\theta)}{\partial \theta_i \theta_j}$. Taking the first partial derivate of $J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(\theta^T x^{(i)} - y^{(i)})^2$, we get

$$\frac{\partial J(\theta)}{\partial \theta_k} = \sum_{i=1}^{m}(\theta^T x^{(i)} - y^{(i)})x_k^{(i)} \tag{1}$$

To see where the term $x_k^{(i)}$ in equation (1) comes from, note that $\theta^T x^{(i)} = \sum_{j=1}^{n} \theta_j x_j^{(i)}$, so $x_k^{(i)}$ is the only element of $x^{(i)}$ that remains after taking the derivate with respect to θ_k.

Taking the second partial derivative, we get

$$\frac{\partial J(\theta)}{\partial \theta_k \theta_l} = \sum_{i=1}^{m} x_l^{(i)} x_k^{(i)} = X^T X, \tag{2}$$

so $H = X^T X$. The sum in (2) shows that the l, kth entry in H is the dot product of columns l and k of X.

b) According to Newton's method, $\theta_t := \theta_{t-1} - H^{-1}\nabla_\theta l(\theta)$. Equation (1) is the kth element of $\nabla_\theta l(\theta)$, and (2) is H. We only need to translate (1) into matrix notation.

The term $\theta^T x^{(i)}$ is the dot product of θ with row i of X. In matrix notation, this becomes $X\theta$. We then subtract Y from this, giving $X\theta - Y$. The kth element of $\nabla_\theta l(\theta)$ is then the dot product of the kth column of X with $X - \theta Y$. In matrix notation, this is $X^T(X\theta - Y)$. Multiplying this by H^{-1} gives $(X^T X)^{-1} X^T (X\theta - Y)$. Distributing and cancelling, we get $\theta - (X^T X)^{-1} X^T Y$.

If we let θ_0 and θ_1 be the values of theta on the first and second iteration of Newton's algorithm, we get $\theta_1 := \theta_0 - (\theta_0 - (X^T X)^{-1} X^T Y = (X^T X)^{-1} X^T Y$. The rightside of the this equation is the normal equations for linear regression. So, Newton's method converges to the correct value for theta in one iteration.

2 Locally-weighted logistic regression

Before we implement the Newton-Raphson algorithm to perform locally-weighted regression, we'll derive the formulas that are given in the homework problem. For reference,

$$l(\theta) = -\frac{\lambda}{2}\theta^T\theta + \sum_{i=1}^{m} w^{(i)} \left[y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)})) \right] \tag{1}$$

First we'll derive

$$\nabla_\theta l(\theta) = X^T z - \lambda\theta. \tag{2}$$

Going from left to right, the regularization term, $-\frac{\lambda}{2}\theta^T\theta$, is the dot-product of θ with itself. Letting $f(\theta) = \sum_i^m \theta_i\theta_i$,we see that $\frac{\partial f}{\partial \theta_j} = 2\theta_j$, so $\frac{1}{2}\lambda\nabla_\theta\theta^T\theta = \lambda\theta$.

From page 18 of lecture notes 1, we know that the partial derivative of the summation with respect to θ_j is $w^{(i)}(y^{(i)} - h(x^{(i)})x_j^{(i)}$. Letting $z^{(i)} = w^{(i)}(y^{(i)} - h(x^{(i)})$, this becomes $z^{(i)}x_j^{(i)}$. We've seen this pattern before: the dot product of the ith row of z with the jth column of X. In matrix notation, this is X^Tz. Combining, we get equation 2.

Next we derive the equation for the Hessian,

$$H = X^TDX - \lambda I, \tag{3}$$

where D is a diagonal matrix with

$$D_{ii} = -w^{(i)}h_\theta(x^{(i)})(1 - h_\theta(x^{(i)})).$$

From 2, we can read off the jth component of $\nabla_\theta l(\theta)$ as $\sum_{i=1}^m w^{(i)}(y^{(i)} - h_\theta(x^{(i)}))x_j^{(i)} - \lambda\theta_j$. Taking the second-partial derivative with respect to k, we get

$$\begin{aligned}\frac{\partial\nabla_\theta l}{\partial\theta_j\partial\theta_k} &= -\sum_{i=1}^m w^{(i)}h_\theta(x^{(i)})x_j^{(i)} \\ &= -\sum_{i=1}^m w^{(i)}h_\theta(x^{(i)})(1 - h_\theta(x^{(i)}))x_j^{(i)}x_k^{(i)} - \lambda\theta_{k=j},\end{aligned}$$

where the expansion of $h_\theta(x^{(i)})$ in the second equation comes from page 17 of lecture notes 1.

Written in matrix notation, the sum is $X^TD_{ii}X$, and the second term is λI. Putting it together, we get equation 3.

The R code to implement locally-weighted is located in ps1_q2.R.

3 Multivariate least squares

a) The Frobenious norm of a matrix A is given by

$$\|A\|_F = \sqrt{\sum_{i=1}^m\sum_{j=1}^n A_{ij}^2} = \sqrt{tr(A^TA)}.$$

From this we can see that $J(\Theta)$ is the (squared) Frobenious norm with $A = X\Theta - Y$, so in matrix notation, $J(\Theta) = tr((X\Theta - Y)^T(X\Theta - Y))$.

b) We'll use various properties of the trace and it's derivative to derive the normal equations for theta in the multivariate context. First, expand the expression inside the trace,

$$\begin{aligned}
\nabla_\Theta tr((X\Theta - Y)^T(X\Theta - Y)) &= \nabla_\Theta tr((X\Theta - Y)^T(X\Theta - Y)) \\
&= \nabla_\Theta tr((\Theta^T X^T X\Theta - Y^T X\Theta - \Theta^T X^T Y + Y^T Y)) \\
&= \nabla_\Theta (tr(\Theta^T X^T X\Theta) - tr(Y^T X\Theta) - tr(\Theta^T X^T Y) + tr(Y^T Y)) \\
&= \nabla_\Theta (tr(\Theta^T X^T X\Theta) - tr(\Theta Y^T X) - tr(Y^T X\Theta) + tr(Y^T Y)) \\
&= \nabla_\Theta tr(\Theta^T X^T X\Theta) - \nabla_\Theta tr(\Theta Y^T X) - \nabla_\Theta tr(\Theta Y^T X) + \nabla_\Theta tr(Y^T Y) \\
&= 2X^T X\Theta - 2X^T Y \\
\Theta &= (X^T X)^{-1} X^T Y
\end{aligned}$$

c) Treating the problem as multiple, independent least-square problems will not change the parameter values, because the matrix Y acts on $(X^T X)^{-1} X^T$ columnwise. In other words, the ith column of Θ is the product of $(X^T X)^{-1} X^T$ with the the ith column of Y, the exact same formula we derived in the univariate regression setting.

4 Naive Bayes

a) To find the joint likelihood function of $l(\phi) = \log \prod_{i=1}^m p(x^{(i)}, y^{(i)}; \phi)$, first use Bayes Theorem to factor the joint probabilities to get

$$p(x^{(i)}, y^{(i)}; \phi) = p(x^{(i)} \mid y^{(i)}; \phi) p(y^{(i)} \mid \phi),$$

and then distribute the log to get

$$\log \prod_{i=1}^m p(x^{(i)}, y^{(i)}; \phi) = \sum_{i=1}^m \log p(x^{(i)} \mid y^{(i)}; \phi) + \log p(y^{(i)} \mid \phi). \qquad (1)$$

We can then factor each expression in the sum separately, and plug them back into 1. Going from left to right,

$$\begin{aligned}
\log p(x^{(i)} \mid y^{(i)}; \phi) &= \log \prod_{j=1}^n (\phi_{j|y=y^{(i)}})^{x_j^{(i)}} (1 - \phi_{j|y=y^{(i)}})^{1-x_j^{(i)}} \\
&= \sum_{j=1}^n x_j^{(i)} \log \phi_{j|y=y^{(i)}} + (1 - x_j^{(i)}) \log(1 - \phi_{y=y^{(i)}})
\end{aligned}$$

and

$$\log p(y \mid \phi) = y \log \phi_y + (1 - y) \log \phi_y$$

Substituting these two expressions into 1, we get

$$\log \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}; \phi) = \sum_{i=1}^{m} \left[\sum_{j=1}^{n} x_j^{(i)} \log(\phi_{j|y=y^{(i)}}) + (1 - x_j^{(i)}) \log(1 - \phi_{j|y=y^{(i)}}) \right. \\ \left. + y^{(i)} \log \phi_y + (1 - y^{(i)}) \log(1 - \phi_y) \right] \quad (2)$$

Note that the second line in this equation is also within the sum $\sum_{i=1}^{m}$.

b) We'll start with the parameters $\phi_{j|y=y^{(i)}}$. Taking partial derivatives, setting equal to zero, and then cross-multiplying, we get

$$\begin{aligned}
\frac{\partial \log \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}; \phi)}{\partial \phi_{j|y=y^{(i)}}} &= \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{x_j^{(i)}}{\phi_{j|y=y^{(i)}}} - \frac{(1 - x_j^{(i)})}{(1 - \phi_{j|y=y^{(i)}})} \\
&= \sum_{i=1}^{m} \sum_{j=1}^{n} x_j^{(i)} (1 - \phi_{j|y=y^{(i)}}) - \phi_{j|y=y^{(i)}} (1 - x_j^{(i)}) \\
&= \sum_{i=1}^{m} \sum_{j=1}^{n} x_j^{(i)} - \phi_{j|y=y^{(i)}}
\end{aligned}$$

To simplify notation, we'll let $y = 0$ and $j = k$. The results will immediately generalize to all values of y and j. This gives

$$\begin{aligned}
\sum_{i=1}^{m} x_k^{(i)} - \phi_{k|y=0} &= \sum_{i=1}^{m} 1 \left\{ x_k^{(i)} = 1 \wedge y = 0 \right\} - \phi_{k|y=0} \sum_{i=1}^{m} 1 \{ y = 0 \} \\
\phi_{k|y=0} &= \frac{\sum_{i=1}^{m} 1 \left\{ x_k^{(i)} = 1 \wedge y = 0 \right\}}{\sum_{i=1}^{m} 1 \{ y = 0 \}}.
\end{aligned}$$

Substitute j for k to get the general result. The derivation for $y = 1$ is identical.

The derivation of the parameters ϕ_y is similar: take partial derivatives, set equal to zero, cross-multiply, cancel, profit. Skipping straight to the partial derivative, we get,

$$\begin{aligned}
\sum_{i=1}^{m} \frac{y^{(i)}}{\phi_y} - \frac{1 - y^{(i)}}{1 - \phi_y} &= 0 \\
&= \sum_{i=1}^{m} (1 - \phi_y) y^{(i)} - (1 - y^{(i)}) \phi_y \\
&= \sum_{i=1}^{m} y^{(i)} - \phi_y
\end{aligned}$$

Setting $y^{(i)} = 0$, we get

$$\begin{aligned}\sum_{i=1}^{m} 1\{y = 0\} - \phi_0 \sum_{i=1}^{m} 1 = 0 \\ \phi_0 = \frac{\sum_{i=1}^{m} 1\{y = 0\}}{m}\end{aligned}$$

The derivation is idential for $y^{(i)} = 1$.

c) First write $p(y = 1 \mid x) > p(y = 0 \mid x)$ in terms of Bayes Theorem, to get

$$\begin{aligned}\frac{p(x \mid y = 1)p(y = 1)}{p(x)} &> \frac{p(x \mid y = 0)p(y = 0)}{p(x)} \\ p(x \mid y = 1)p(y = 1) &> p(x \mid y = 0)p(y = 0)\end{aligned}$$

Taking the log of $p(x \mid y = k)$, we get

$$\begin{aligned}\log p(x \mid y = k) &= \log \prod_{j=1}^{n} (\phi_{j|y=1})^{x_j} (1 - \phi_{j|y=1})^{1-x_j} \phi_k^k (1 - \phi_k)^{1-k} \\ &= \sum_{j=1}^{m} x_j \log \phi_{j|y=1} + (1 - x_j) \log(1 - \phi_{j|y=1}) + \log(\phi_k^k (1 - \phi_k)^{1-k}) \\ &= \sum_{j=1}^{m} x_j (\log \phi_{j|y=1} - \log(1 - \phi_{j|y=1})) + \log(\phi_k^k (1 - \phi_k)^{1-k}) \\ &= \sum_{j=1}^{m} x_j \log \frac{\phi_{j|y=1}}{1 - \phi_{j|y=1}} + \log(1 - \phi_{j|y=1}) + \log(\phi_k^k (1 - \phi_k)^{1-k})\end{aligned}$$

Define θ_k as the vector whose jth component is $\log \frac{\phi_{j|y=k}}{1-\phi_{j|y=k}}$ for $j > 0$, and $\sum_{j=1}^{m} \log(1 - \phi_{j|j=k}) + \log(\phi_k^k (1 - \phi_k)^{1-k})$ for $j = 0$. Then $p(x \mid y = 1) = \theta_1^T \begin{bmatrix} 1 \\ x \end{bmatrix}$. Define θ_0 in an analogous way. Then $\theta = \theta_1 - \theta_0$ is the vector such that $\theta^T \begin{bmatrix} 1 \\ x \end{bmatrix} > 0 \iff p(x \mid y = 1) > p(x \mid y = 0)$.

5 Exponential family and geometric distribution

a) Use the take-logs-then-exponentiate trick:

$$\begin{aligned}\exp \log p(y; \phi) &= \exp \log((1 - \phi)^{y-1} \phi) \\ &= \exp((y - 1) \log(1 - \phi) + \log \phi) \\ &= \exp(\log(1 - \phi) y + \log(\frac{\phi}{1 - \phi}))\end{aligned}$$

So,

$$\begin{aligned} b(y) &= 1 \\ \eta &= \log(1-\phi) \\ T(y) &= y \\ a(\eta) &= \log(\frac{\phi}{1-\phi}). \end{aligned}$$

b) We can write the mean of the geometric distribution as a function of η as

$$\frac{1}{\phi} = \frac{1}{1-\exp\eta}.$$

c) The log-likelihood is

$$\begin{aligned} \log \ell(\theta) &= \log \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}; \theta) \\ &= \sum_{i=1}^{m} \theta^T x^{(i)} y + \log(\frac{\phi}{1-\phi})) \\ &= \sum_{i=1}^{m} \theta^T x^{(i)} y + \log(\frac{1-\exp(\theta^T x^{(i)})}{-\exp(\theta^T x^{(i)})}) \\ &= \sum_{i=1}^{m} \theta^T x^{(i)} (y+1) + \log(1-\exp(\theta^T x^{(i)})) \end{aligned}$$

The jth partial derivative is then $=$

$$\begin{aligned} \frac{\partial \log \ell(\theta)}{\partial \theta_j} &= \sum_{i=1}^{m} x_j^{(i)} (y^{(i)}+1) + \frac{x_j^{(i)} \exp(\theta^T x^{(i)})}{1-\exp(\theta^T x^{(i)}} \\ &= \sum_{i=1}^{m} x_j^{(i)} \left(y^{(i)} + \frac{1}{1-\exp(\theta^T x^{(i)})} \right) \end{aligned}$$

The update rule for stochastic gradient descent is th

$$\theta_j := \theta_j + \alpha x_j^{(i)} \left(y^{(i)} + \frac{1}{1-\exp(\theta^T x^{(i)}} \right)$$