Big Data Hw2

Alex Fay

February 2020

1a. Let $\sigma(x) = 1/1 + (e^x)$ be the sigmoid function. Show that σ'(x) = σ(x)[1 σ(x)].

$\frac{d}{dx}\ \frac{1}{1+e^{-x}} = \frac{1}{1+e^{-x}}\ (1 - \frac{1}{1+e^{-x}})$

$= e^{-x}(1+e^{-x})^{-2}$ chain rule from derivative

$= ((1+e^{-x})^{-1})(\frac{e^{-x}+1-1}{1+e^{-x}})$

$= \sigma$(x) * (1 - $\frac{1}{1+e^{-x}}$)

$= \sigma(x)(1-\sigma(x))$

Thus, the proposition holds true.

1b. The given eqn for negative logistic regression is:
nll(θ) = Σ y log$(\sigma(\theta^T x) + (1-y)(log(1-(\sigma)(\theta^T * x))$

By taking the gradient of the neg log equation we get:

∇nll = $\nabla\Sigma\frac{y*\sigma'*\theta^T*x}{\sigma*\theta^T*x}$ + (1-y) $\frac{1}{1-\sigma(\theta^T*x)} - \sigma(\theta^T * x)$

From substituting in part A:
$\nabla x = \sigma y(1-\sigma(\theta^T * x)x) - (\frac{1-y}{1-\sigma(\theta^T*x))-\sigma'(\theta^T*x)})$
$= \Sigma\sigma(\sigma^T * x - y)$
$= X^T((\sigma * \theta^T * x) - y)$
$= X^T * (\mu^T y)$

1c.
Taking the gradient the proposition given in part c: $= \nabla\theta[X^T * (\mu - y)]^T$
$= \nabla X(\mu^T - y^T))$
$= X * \mu^T$
$= X^T * diag(\mu(1-\mu) * X)$
$= X^T * R * X$

Since R represents the diagonal matrix, and by definition of a diagonal matrix the eigenvalues are the diagonals, then: For all sigma(theta) is between 0 and 1
Thus $\sigma(x) * (1 - \sigma(x))$ is greater than or equal to 0.
By definition of a positive semi-definite, the above is positive semi-definite

2. Z = $\int e$-x$^2 \frac{}{2*\sigma^2 dx)}$
Squaring Z and subsituting x and y in forms of r ($x^2 + y^2 = r^2$):
$= \int_0^\infty \int_0^2 \pi e^{\frac{-r^2}{2\sigma^2}} rdrd\theta$
$= 2\pi * \int_0^\infty e^{\frac{-r^2}{2\sigma^2}} * rdr$
$= 2\pi\sigma^2$
Z = $root(Z^2) = \sigma * \sqrt{2\pi}$

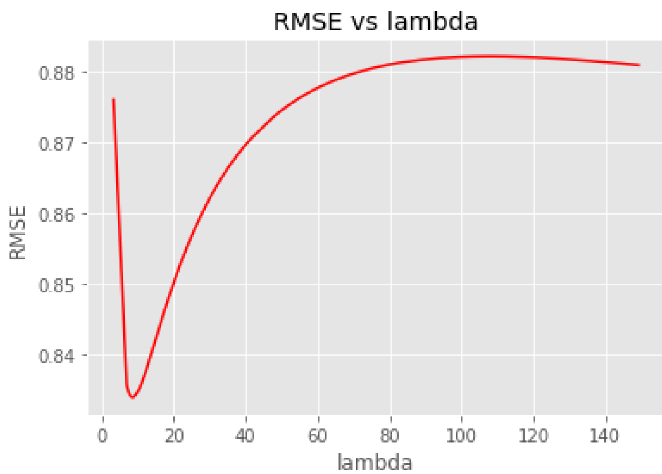
3a. N(x— μ, σ) =
arg max $\Sigma log(\frac{1}{\sigma\sqrt{2\pi}} * e^{-\frac{(-w^t - w0 + y)^2}{2\sigma^2}} + \Sigma e^{\frac{-w^2}{2T^2}} log(\frac{1}{\sigma\sqrt{2\pi}})$
$= argmax - ((N + D)log(\sigma\sqrt{2\pi}) + \sigma\frac{-w0 - w^T x - y}{2\sigma^2} + \Sigma\frac{w^2}{2T^2})$

The constants can be scaled above by $2\sigma^2$ since it does not affect the weight value.

$= argmax - 1\Sigma(-w0 - w^T + y)^2 + \frac{\sigma^2}{T^2} * \sigma w^2$
Since the function is maximizing the negative function, it is the same as minimizing its positive counterpart.
From substitution of Lamda = sigma$^2/T^2, weget:$
argmin $\Sigma(y - w0 - w^T * x)^2 + \lambda||w||$

3b. $\nabla f(x) = \nabla x[(Ax-b)^T * (Ax-b) + (Tx^T) * (Tx)]$
$= \nabla x(x^T * A^T - b^T)(Ax-b) + x^T * T^T * Tx$
$= \nabla x(x^T * AT * Ax - 2x^T * A^T * b + b^T b + x^T * T^T * Tx$
$0 = 2A^T Ax - 2A^T b + 2TT * Tx$
$A^T b = A^T Ax + T^T * Tx$
$x = (A^T * A + T^T * T)^{-1} * A^T * b$
T $=$ I$\sqrt{\lambda}$
$x = (A^T * A + \lambda * I)^- 1 * A^T * b$

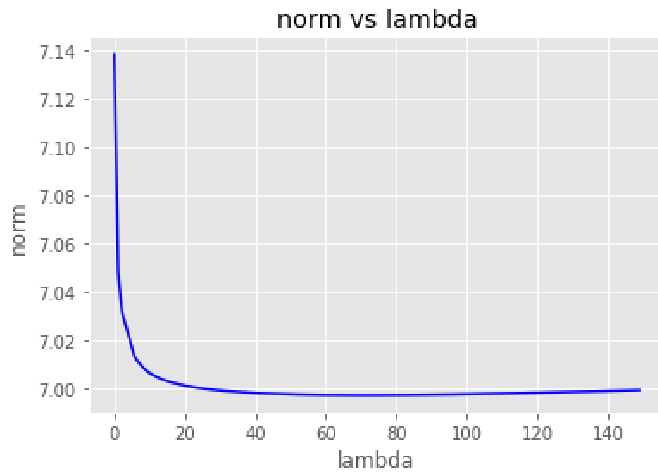
3c.



3d. f(x) $= ||Ax+b-y|| + ||Tx||$
$= (Ax+b-y)^T(Ax+b-y) + (Tx)^T * (Tx)$
$= x^T * A^T * AX + 2b^T Ax - 2y^T Ax - 2b^T y + nb^2 + y^T y + x^T T^T Tx$
$\nabla f(x) = 2A^T Ax + 2bA^T - 2A^T y + 2T^T Tx = 0$
$0 = 2Ax - 2y + 2bm$
$b = \frac{1^T(y-Ax)}{n}$

Plugging b back in:
$(A^T A + T^T T)x + bA^T - A^T y = 0$
$(A^T A + T^T T - \frac{1^T A^T A}{n})x = A^T y - \frac{A^T(1^T)y}{n}$
$(A^T(I - \frac{1^T}{n}))A + T^T T)x = A^T(I - \frac{1^T}{n})$
$x = [A^T(I - \frac{1^T}{n}) + T^T T]^{-1} * A^T(I - 1^T/n)y$

3e.
The results compared to part C are the the same. (Very tiny, negligible error less than 1 percent)



From the code attached we can see:
The optimal regularization parameter is 8.5736.
The RMSE on the validation set with the optimal regularization parameter is 0.8340.
The RMSE on the test set with the optimal regularization parameter is 0.8628.