UNIVERSITY OF CALIFORNIA, BERKELEY

HW 4: CS 189

Oscar Ortega

July 16, 2021

I worked with Yuchen Hua and Ryan Chan: I certify that all solutions are in entirely in my own words, and that I have not looked at another students solutions. I have given credit to all external sources I consulted.
-Oscar Ortega

1 LOGISTIC REGRESSION WITH NEWTON'S METHOD

a:

$$J(w) = \lambda \| w \|_2^2 - \sum_{i=1}^{n} y_i ln(s_i) + (1 - y_i) ln(1 - s_i)$$

$$\nabla J_w = (\lambda \| w \|_2^{2\prime} - \sum_{i=1}^{n} y_i ln'(s_i) + (1 - y_i) ln'(1 - s_i))$$

$$= (2\lambda w - \sum_{i=1}^{n} \frac{y_i}{s_i} s_i' + \frac{1 - y_i}{1 - s_i}(-s_i'))$$

$$= 2\lambda w - \sum_{i=1}^{n} \frac{y_i}{s_i}(s(X_i^T w))(1 - s(X_i^T w) - \frac{1 - y_i}{1 - s_i}(s(X_i^T w))(1 - s(X_i^T w)^T$$

$$= (2\lambda w - \sum_{i=1}^{n} (s_i)(1 - s_i)(X_i^T w)'(\frac{y_i}{s_i} - \frac{1 - y_i}{1 - s_i}))$$

$$= (2\lambda w - \sum_{i=1}^{n} (s_i)(1 - s_i)(X_i^T)(\frac{y_i}{s_i} - \frac{1 - y_i}{1 - s_i}))$$

$$2\lambda w - \sum_{i=1}^{n} X_i^T (1 - s_i)(y_i) - s_i(1 - y_i)$$

$$= 2\lambda w - \sum_{i=1}^{n} X_i^T (y_i - s_i)$$

$$= 2\lambda w - X^T (y - s(Xw))$$

b:

$$\nabla^2 J_w = (2\lambda - X^T (y - s(Xw)))'$$

$$= 2\lambda - (X^T y + X^T s(Xw))'$$

$$= (2\lambda I) - X^T \text{diag}(s(Xw)(1 - s(Xw)))X$$

c: The update equation for newtons algorithm is:

$$W^{new} = W^{old} - H^{-1} \nabla_w$$

$$w^{new} = w^{old} - ((2\lambda I) - X^T \text{diag}(s(Xw)(1 - s(Xw)))X)^{-1} (2\lambda w - X^T (y - s(Xw))$$

d: This portion is on the iPython notebook

2 $L_1 and L_2 : regularization$

a:

$$J(w) = g(y) + \sum_{i=1}^{d} f(X_* i, w_i, y, \lambda)$$

$$= \langle Xw, Xw \rangle - 2\langle Xw, y \rangle + \langle y, y \rangle + \lambda \sum_{i=1}^{d} |w_i|$$

$$w^T X^T X w - 2w^T X^T y + \lambda \sum_{i=1}^{d} |w_i| + y^T y$$

$$nw^T w - 2w^T X^T y + \lambda \sum_{i=1}^{d} |w_i| + y^T y$$

$$g(y) = y^T y$$

$$f(X_* i, w_i, y, \lambda) = nw_i^2 - 2(X_{*i} w_i)^T y + \lambda |w_i|$$

b:
If we let $w_i^* > 0$

$$\frac{\partial f}{\partial w_i} = 2nw_i - 2{X_*}_i^T y_i + \lambda > 0$$

$$w_i^* = \frac{-\lambda}{2n} + \frac{{X_*}_i^T y}{n}$$

c:
If we let $w_i^* < 0$

$$\frac{\partial f}{\partial w_i} = 2nw_i - 2{X_*}_i^T y_i - \lambda = 0$$

$$w_i^* = \frac{\lambda}{2n} + \frac{{X_*}_i^T y}{n}$$

d:
We know that $w_i^* = 0\ 2{X_*}_i^T y < \lambda$ by verifying that the w_i is neither greater than nor less than 0.
Setting the inequality from part b to be false:

$$\frac{-\lambda}{2n} + \frac{{X_*}_i^T y}{n} < 0$$

$$2{X_*}_i^T y < \lambda$$

Setting the inequality from part c to be false:

$$\frac{\lambda}{2n} + \frac{{X_*}_i^T y}{n} > 0$$

$$2{X_*}_i^T y < \lambda$$

e:

$$\begin{aligned} f(X_* i, w_i, y, \lambda) &= n w_i^2 - 2(X_{*i} w_i)^T y + \lambda w_i^2 \\ &= (n-\lambda) w_i^2 - 2(X_* w_i)^T y \\ \frac{\partial f}{\partial w_i} &= 2(n-\lambda) w_i - 2(X_{*i})^T y = 0 \\ w_i^* &= \frac{(X_{*i})^T y}{n-\lambda} \end{aligned}$$

This is equal to zero when $X_{*i} \perp y$. This differs from the expression for the l_1 norm because we have an equality rather than an inequality that depends on λ.

credit to the following-notes on min-norm solutions and least squares regularization. a:

$$\nabla |w|^4 = \nabla (w^T w)(w^T w)$$

$$= 2w\|w\|_2^2 + 2w\|w\|_2^2$$

$$= 4(w^T w)w$$

$$\nabla_w \|Xw - y\|_2^4$$

$$= \nabla_w \langle Xw - y, Xw - y\rangle^2$$

$$= 2\langle Xw - y, Xw - y\rangle \langle Xw - y, Xw - y\rangle'$$

$$= 4(Xw - y)^T (Xw - y)(X^T Xw - X^T y)$$

b: Consider once again, the following expression:

$$\|Xw - y\|^4 + \lambda \|w\|_2^2$$

We know that if f is a strictly convex function, and that if g is a convex function, then their sum f + g is a strictly convex function. We also know that if f and g are both strictly convex functions then sum f + g is also a strictly convex function.

$$f(x) = x^2 \text{ is a strictly convex function}$$

Proof:

$$f(\lambda x_1 + (1-\lambda)x_2) = (\lambda x_1 + (1-\lambda)x_2)^2$$

$$= \lambda^2 x_1^2 + 2\lambda(1-\lambda)x_1 x_2 + (1-\lambda)^2 x_2^2$$

$$\leq \lambda^2 x_1^2 + (1-\lambda)^2 x_2^2$$

$$< \lambda x_1^2 + (1-\lambda)x_1^2$$

$$< \lambda f(x_1) + (1-\lambda) f(x_2)$$

$$\|w\|_2^2 = \sum_{i=1}^{n} w_i^2$$

is strictly convex. This leaves us to show that $\|Xw - y\|^4$ is convex.
Proof:
Starting from the gradient:

$$\nabla w = 4\langle Xw - y, Xw - y\rangle (X^T Xw - X^T y)$$

Let $a = \langle Xw - y, Xw - y\rangle$
Let $b = X^T Xw - X^T y = a'$

$$\nabla^2 w = 4(ab)' = 4(a'b + ab')$$

$$= 4(b^T b + ab')$$

$$= 4((X^T Xw - X^T y)^T (X^T Xw - X^T y) + \langle Xw - y, Xw - y\rangle X^T X)$$

$$= \|X^T (Xw - y)\|^2 I + \|Xw - y\|^2 X^T X = H$$

$$z^T Hz = z^T \|X^T (Xw - y)\|^2 I + \|Xw - y\|^2 X^T Xz$$

$$= z^T \|X^T (Xw - y)\|^2 Iz + z^T \|Xw - y\|^2 X^T Xz$$

$$= \|X^T (Xw - y)\| \|z\|^2 + \|Xw - y\| \|Xz\|^2 \geq 0$$

Therefore, we know w^* is unique.
We can find the optimal value of this function by setting the gradient equal to zero.

$$\nabla_w \|Xw - y\|^4 + \lambda \|w\|_2^2$$

$$= 4\|Xw - y\|_2^2 (X^T Xw - X^T y) + 2\lambda w = 0$$

$$\frac{\lambda}{2\|Xw - y\|_2^2} w + X^T Xw = X^T y$$

$$\frac{\lambda}{2\|Xw - y\|_2^2} w = X^T y - X^T Xw$$

$$\frac{\lambda}{2\|Xw - y\|_2^2} w = X^T (y - Xw)$$

$$w = X^T v : w \in \mathscr{R}(X^T)$$

$$w^* = \frac{X^T (y - Xw) 2\|Xw - y\|_2^2}{\lambda}$$

$$\sum_{i=1}^{n} \frac{2\|Xw - y\|_2^2 (y - Xw)_i}{\lambda} X_i$$

$$\text{Where } a_i = \frac{2\|Xw - y\|_2^2 (y - Xw)_i}{\lambda}$$

Consider once again, the following expression:

$$\|Xw - y\|^4 + \lambda \|w\|_2^2$$

We know that if f is a strictly convex function, and that if g is a convex function, then their sum f + g is a strictly convex function. We also know that if f and g are both strictly convex functions then sum f + g is also a strictly convex function.

$$f(x) = x^2 \text{ is a strictly convex function}$$

Proof:

$$f(\lambda x_1 + (1 - \lambda) x_2) = (\lambda x_1 + (1 - \lambda) x_2)^2$$

$$= \lambda^2 x_1^2 + 2\lambda(1 - \lambda) x_1 x_2 + (1 - \lambda)^2 x_2^2$$

$$\leq \lambda^2 x_1^2 + (1-\lambda)^2 x_2^2$$
$$< \lambda x_1^2 + (1-\lambda) x_1^2$$
$$< \lambda f(x_1) + (1-\lambda) f(x_2)$$
$$\|w\|_2^2 = \sum_{i=1}^{n} w_i^2$$

is strictly convex. This leaves us to show that $\|Xw - y\|^4$ is convex.
Proof:
Starting from the gradient:

$$\nabla w = 4\langle Xw - y, Xw - y\rangle (X^T Xw - X^T y)$$

Let $a = \langle Xw - y, Xw - y\rangle$
Let $b = X^T Xw - X^T y = a'$

$$\nabla^2 w = 4(ab)' = 4(a'b + ab')$$
$$= 4(b^T b + ab')$$
$$= 4((X^T Xw - X^T y)^T (X^T Xw - X^T y) + \langle Xw - y, Xw - y\rangle X^T X)$$
$$= \|X^T (Xw - y)\|^2 I + \|Xw - y\|^2 X^T X = H$$
$$z^T Hz = z^T \|X^T (Xw - y)\|^2 I + \|Xw - y\|^2 X^T Xz$$
$$= z^T \|X^T (Xw - y)\|^2 Iz + z^T \|Xw - y\|^2 X^T Xz$$
$$= \|X^T (Xw - y)\| \|z\|^2 + \|Xw - y\| \|Xz\|^2 \geq 0$$

c:
Proof:
Assume for the sake of contraction that

$$p^* = \text{argmin}_{x \in \mathbf{R}^d} = \sum_i a_i X_i + \epsilon : \epsilon \perp X_i, i = 1, ..., n$$

$$\rightarrow \frac{1}{n} \sum_{i=1}^{n} L(p^T x_i, y_i) + \lambda \|w\|^2 \frac{1}{n} = \sum_{i=1}^{n} L((\sum_i (a_i X_i + \epsilon)^T X_i, y_i) + \lambda \| \sum_{i=1}^{n} a_i X_i + \epsilon\|^2$$
$$= \frac{1}{n} \sum_{i=1}^{n} L((\sum_i (a_i X_i + \epsilon)^T X_i, y_i) + \lambda \| \sum_{i=1}^{n} a_i X_i + \epsilon\|^2$$
$$= \frac{1}{n} \sum_{i=1}^{n} L(\sum_i (a_i X_i^T X_i) + \epsilon^T X_i, y_i) + \lambda \langle \sum_{i=1}^{n} a_i X_i + \epsilon, \sum_{i=1}^{n} a_i X_i + \epsilon\rangle$$
$$= \frac{1}{n} \sum_{i=1}^{n} L(\sum_i (a_i X_i^T X_i), y_i) + \lambda \| \sum_{i=1}^{n} a_i X_i\|^2 + \lambda \|\epsilon\|^2$$

$\rightarrow p^*$ is not optimal because you can remove the ϵ to lower the cost w

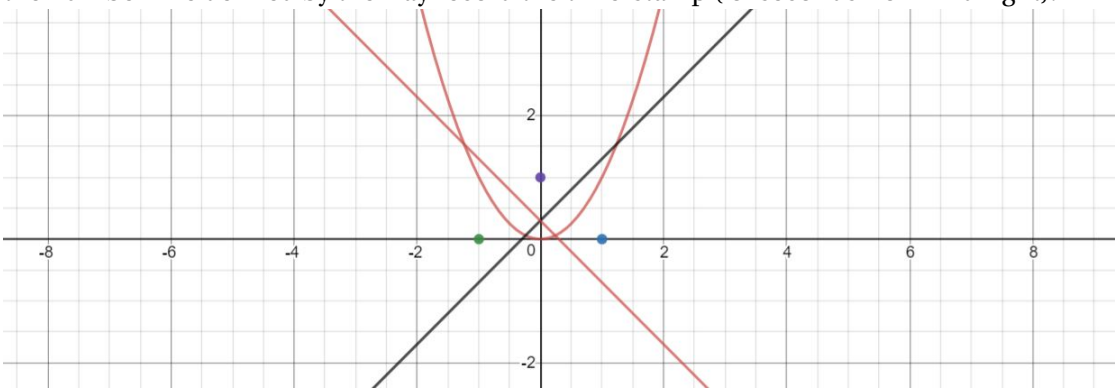
Because the argument above did not take into consideration the convexity of the loss function, this implies that for any loss function, p^* must be a strict linear combination of the data points to be a minimizer and cannot be an affine transformation.

4 Wine Classification with Logistic Regression

This section is in the iPython notebook.

5 Real World Spam-Classification

Given two classes, 'spam' and 'ham', and given that we know that around midnight our emails tend to be more 'spammy', a linear decision boundary, i.e a two dimensional hyperplane would not be able to optimally classify the features because the features are not linearly separable, because the spikes of spam at that occur at around midnight are on opposite end of the number line defined by the way record the time-stamp (of seconds from midnight).



In the graphic, I am representing the ham as the points with x-coordinate = 1, and spam as the points with x-coordinate equal to 0. As we can see any linear separator of the data cannot accurately classify 'spam' from 'ham', on the other hand, if we now had a quadratic kernel, separability is feasible as we can now define a quadratic decision boundary that accurately separates the features. Keep in mind however, that the lower than expected returns in accuracy might not improve due to noise in the data and the biases in the model by the features already included in our SVM.