CS 760: Machine Learning Spring 2024

Homework 3: Logistic Regression

AUTHORS: Jed Pulley

DO NOT POLLUTE! AVOID PRINTING, OR PRINT 2-SIDED MULTIPAGE.

Problem 3.1

(a) I used good ol' fashioned guess and check. I set the iterations to 3000 and printed out the weights after every 100 iterations. Once the values started to level off, I considered that convergence

(b) It took me about 3000 iterations to converge.

(c) I performed min-max normalization to avoid running into a RuntimeWarning (when the exponential got WAY too big), so my results for $\hat{\theta}$ were:

$$[-6.75837555, 4.55726066, -14.28891882, -3.05890586, -0.50213527, 7.82165471]$$

(d) The maximum log-likelihood of $\hat{\theta}$ was: -470.15901001623024

(e) From Theorem 6.2 in the Logistic Regression notes, we can see that $\hat{\theta} \xrightarrow{\text{d}} \mathcal{N}(\theta^*, I_{\theta^*}^{-1})$ where the Fisher Information is shown as:

$$I_{\theta^*} = \sum_{i=1}^{N} \frac{e^{-\theta^{*T}\mathsf{x}_\mathsf{i}}}{(1+e^{-\theta^{*T}\mathsf{x}_\mathsf{i}})^2} \mathsf{x}_\mathsf{i}\mathsf{x}_\mathsf{i}^\mathsf{T}$$

Problem 3.2

(a) Borrowing from the Logistic Regression notes again, we can see that the MLE of the log-odds $\hat{\omega} := \hat{\theta}^\mathsf{T}\mathsf{x}$ where $\hat{\theta}$ are the true parameters, θ^*.

(b) Furthermore, the asymptotic distribution of $\hat{\omega}$ is defined as $\hat{\omega} \xrightarrow{\text{d}} \mathcal{N}(\theta^{*\mathsf{T}}\mathbf{x}, \mathbf{x}^\mathsf{T} I_{\theta^*}^{-1}\mathbf{x})$

Problem 3.3

(a) I maximized my feature vector, having my entire family on board in the cheapest class and a really low fare. With that, my feature vector looked like so:

$$[Pclass = 3, male = 0, age = 24, siblings = 7, parents = 2, fare = 8]$$

When I ran this through my model, unfortunately I did not survive.

(b) Given $\tau = \Phi_{\mathcal{N}}^{-1}(\frac{\alpha}{2}|0, \mathbf{x}^\mathsf{T} I_{\theta^*}^{-1}\mathbf{x})$, I found τ to be just around 1.

(c) Interpreting this, we fall around 1 standard deviation of the mean given a normal distribution. I would say this is more certain than not. I'm not a betting man, so I don't like my odds, but it's better than 50/50. Also, the accuracy of my model falls around the 70% range, which also isn't terrible.

Problem 3.4

(a) To find the significance of our features, we can use the generalized likelihood ratio test as found in the Logistic Regression Notes under formula 6.13:

$$(\frac{\hat{\theta}_j}{\nu_j})^2 \gtrless \phi_{\mathcal{X}}^{-1}(\alpha)$$

(b) Plugging in $\alpha = 0.05$, we find that passenger class, sex, age, and fare were significant while and number of siblings/spouses and parents/children were not significant.

(c) All things being equal, just changing sex from male to female still has me dying. However, if I change from male to female and then increase my fare, I no longer go down with the ship and make it out alive.