EE5907/EE5027 Week 4: Logistic Regression

The following questions are from Kevin Murphy's (KM) book "Machine Learning: A Probabilistic Perspective".

Exercise 8.3 Gradient and Hessian of log-likelihood for logistic regression

a. Let $\sigma(a) = \frac{1}{1+e^{-a}}$ be the sigmoid function. Show that

$$\frac{d\sigma(a)}{da} = \sigma(a)(1-\sigma(a))$$

b. Using the previous result and the chain rule of calculus, derive an expression for the gradient of the negative log likelihood (NLL)

$$g = \frac{d}{dw}NLL(w) = \sum_i (\mu_i - y_i)x_i$$

c. The Hessian can be written as $H = X^T SX$, where $S \triangleq \text{diag}(\mu_1(1-\mu_1), \cdots, \mu_n(1-\mu_n))$. Show that H is positive definite. (You may assume that $0 < \mu_i < 1$, so the elements of S will be strictly positive, and that X is full rank.)

Exercise 8.6 Elementary properties of ℓ_2 regularized logistic regression

Consider minimizing

$$J(w) = NLL(w) + \lambda||w||_2^2$$

Answer the following true/ false questions.

a. $J(w)$ has multiple locally optimal solutions: T/F?

b. Let $\hat{w} = \text{argmin}_{wJ(w)}$ be a global optimum. $\hat{w}$ is sparse (has many zero entries): T/F?

c. If the training data is linearly separable, then some weights ω_j might become infinite if $\lambda = 0$: T/F?

d. $NLL(w)$ of training set always increases as we increase λ: T/F?

e. $NLL(w)$ of test set always increases as we increase λ: T/F?

Exercise 8.7 Regularizing separate terms in 2d logistic regression

a. Consider the data in Figure 1, where we fit the model $p(y = 1|x, w) = \sigma(\omega_0 + \omega_1 x_1 + \omega_2 x_2)$. Suppose we fit the model by maximum likelihood, i.e., we minimize

$$J(w) = NLL(w)$$

where $NLL(w)$ is the negative log likelihood of the training set. Sketch a possible decision boundary corresponding to $\hat{w}$. (Copy the figure first (a rough sketch is enough), and then superimpose your answer on your copy, since you will need multiple versions of this figure). Is your answer (decision boundary) unique? How many classification errors does your method make on the training set?

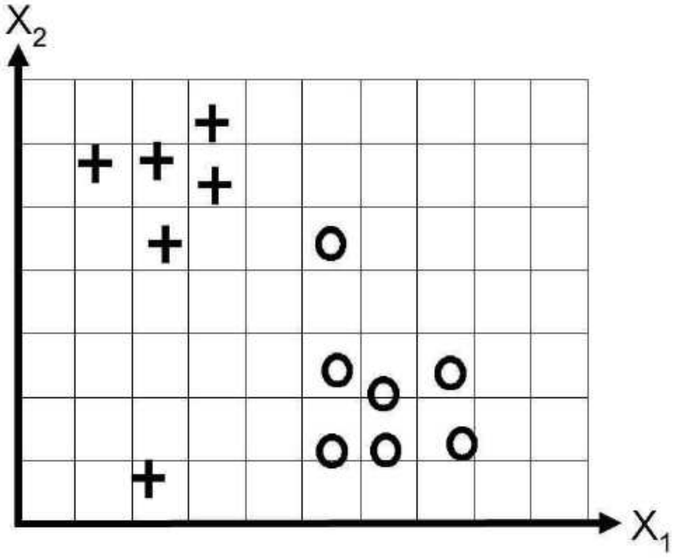

Figure 1: Data for logistic regression question

b. Now suppose we regularize only the ω_0 parameter, i.e., we minimize

$$J_0(w) = NLL(w) + \lambda\omega_0^2$$

Suppose λ is a very large number, so we regularize ω_0 all the way to 0, but all other parameters are unregularized. Sketch a possible decision boundary. How many classification errors does your method make on the training set? Hint: consider the behavior of simple linear regression, $\omega_0 + \omega_1 x_1 + \omega_2 x_2$ when $x_1 = x_2 = 0$.

c. Now suppose we heavily regularize only the ω_1 parameter, i.e., we minimize

$$J_1(w) = NLL(w) + \lambda\omega_1^2$$

Sketch a possible decision boundary. How many classification errors does your method make on the training set?

d. Now suppose we heavily regularize only the ω_2 parameter. Sketch a possible decision boundary. How many classification errors does your method make on the training set?