

# Homework #2: Machine Learning

Kyler Little

February 3, 2018

## Problem #1

(a) More generally, if we are learning from  $\pm 1$  data to predict a noisy target  $P(y|\mathbf{x})$  with candidate hypothesis  $h$ , show that the maximum likelihood method reduces to the task of finding  $h$  that minimizes:

$$E_{\text{in}}(\mathbf{w}) = \sum_{n=1}^N \mathbb{I}[y_n = +1] \ln\left(\frac{1}{h(\mathbf{x}_n)}\right) + \mathbb{I}[y_n = -1] \ln \frac{1}{1 - h(\mathbf{x}_n)}$$

Formally, we are trying to learn the target function:

$$f(\mathbf{x}) = P[y = +1|\mathbf{x}]$$

In non-mathematical terms, this target function represents the probability of getting  $y_n = +1$  from the data  $(\mathbf{x}_n)$ . The function  $f$  is generated by a noisy target function, so the data doesn't give us the value of  $f$  explicitly. Instead, the data is generated by a noisy target function  $P(y|\mathbf{x})$ . Our goal is to minimize the error between the classifications of actual training data and the predicted classifications from our hypothesis model  $h$ . We can do this by talking about the notion of likelihood. The likelihood that the target distribution  $P(y|\mathbf{x})$  is captured by our hypothesis  $h(\mathbf{x})$  is:

$$P(y|\mathbf{x}) = \begin{cases} h(\mathbf{x}) & \text{for } y = +1; \\ 1 - h(\mathbf{x}) & \text{for } y = -1. \end{cases}$$

If we assume that the data points were independently generated (a very fair assumption to make), then we can express the probability of getting all of

the  $y_n$ 's in the data set from their corresponding  $x_n$ 's as:

$$\prod_{n=1}^N P(y_n | \mathbf{x}_n)$$

The method of maximum likelihood method would select the hypothesis  $h$  which maximizes the probability above. We can actually express this maximization problem as a minimization problem by using the properties of 'ln'. The natural logarithm is a monotonically increasing function, and so taking the natural logarithm of the probability above will not affect the maximization problem. By the same token, ' $-\ln$ ' is a monotonically decreasing function, so we can take the natural logarithm of the probability above and convert the maximization problem to a minimization problem without affecting global minima of the probability distribution. Using a simple logarithm rule ( $-\ln x = \ln x^{-1}$ ), we find:

$$\prod_{n=1}^N P(y_n | \mathbf{x}_n) = \sum_{n=1}^N \ln \frac{1}{P(y_n | \mathbf{x}_n)}$$

Substituting in what  $P(y|\mathbf{x})$  actually equals, we arrive at what we wanted. We now find  $h$  that minimizes:

$$E_{\text{in}}(\mathbf{w}) = \sum_{n=1}^N [\![y_n = +1]\!] \ln \frac{1}{h(\mathbf{x}_n)} + [\![y_n = -1]\!] \ln \frac{1}{1 - h(\mathbf{x}_n)}$$

(b) For the case  $h(\mathbf{x}) = \theta(\mathbf{w}^T \mathbf{x})$ , argue that minimizing the in-sample error in part (a) is equivalent to minimizing the one in (3.9).

Substitute in  $h(\mathbf{x}) = \theta(\mathbf{w}^T \mathbf{x}) = \frac{e^{w^T \mathbf{x}}}{1 + e^{w^T \mathbf{x}}}$  to the in-sample error from part (a). This yields:

$$E_{\text{in}}(\mathbf{w}) = \sum_{n=1}^N [\![y_n = +1]\!] \ln \frac{1 + e^{w^T \mathbf{x}}}{e^{w^T \mathbf{x}}} + [\![y_n = -1]\!] \ln \frac{1 - e^{w^T \mathbf{x}}}{1 - 2e^{w^T \mathbf{x}}}$$

Then we do some more math...

**Side Note (from LFD):**

For two probabilistic distributions  $p, 1 - p$  and  $q, 1 - q$  with binary outcomes, the cross-entropy (from information theory) is:

$$p \log \frac{1}{q} + (1 - p) \log \frac{1}{1 - q}$$

The in-sample error in part (a) corresponds to a cross-entropy error measure on the data point  $(\mathbf{x}_n, y_n)$ , with  $p = \mathbb{I}[y_n = +1]$  and  $q = h(\mathbf{x}_n)$ .

## Problem #2

Recall the objective function for linear regression can be expressed as

$$E(w) = \frac{1}{N} \|Xw - y\|^2$$

as in Equation (3.3) of LFD. Minimizing this function with respect to  $w$  leads to the optimal  $w$  as  $(X^T X)^{-1} X^T y$ . This solution holds only when  $X^T X$  is nonsingular. To overcome this problem, the following objective function is commonly minimized instead:

$$E_2(w) = \|Xw - y\|^2 + \lambda \|w\|^2$$

where  $\lambda > 0$  is a user-specified parameter. Please do the following:

(a) Derive the optimal  $w$  that minimize  $E_2(w)$ .

This part is fairly easy because we went over it in class, albeit without the  $\lambda$  term. All I do now is add this in.

(b) Explain how this new objective function can overcome the singularity problem of  $X^T X$ .

Now, just prove it's nonsingular by showing it's positive definite.

## Problem #3

In logistic regression, the objective function can be written as:

$$E(w) = \frac{1}{N} \sum_{n=1}^N \ln(1 + e^{-y_n w^T x_n})$$

(a) (10 points) Compute the first-order derivative  $\nabla E(w)$ . You will need to provide the intermediate steps of derivation.

(b) (10 points) Once the optimal  $w$  is obtained, it will be used to make predictions as follows:

$$\text{Predicted class of } \mathbf{x} = \begin{cases} 1 & \text{if } \theta(w^T \mathbf{x}) \geq 0.5 \\ -1 & \text{if } \theta(w^T \mathbf{x}) < 0.5 \end{cases}$$

where the function  $\theta(z) = \frac{1}{1+e^{-z}}$  looks like

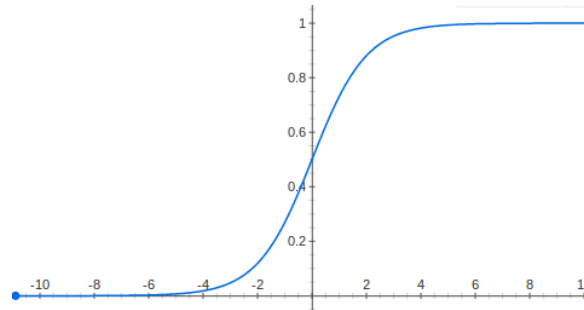


Figure 1: A basic sigmoid curve

Explain why the decision boundary of logistic regression is still linear, though the linear signal  $w^T x$  is passed through a nonlinear function  $\theta$  to compute the outcome of prediction. Is the decision boundary still linear if the prediction rule is changed to the following? Justify briefly.

$$\text{Predicted class of } x = \begin{cases} 1 & \text{if } \theta(w^T x) \geq 0.9 \\ -1 & \text{if } \theta(w^T x) < 0.9 \end{cases}$$

In light of your answers to the above two questions, what is the essential property of logistic regression that results in the linear decision boundary?

## Problem #4

As your final deliverable to a customer, would you use the linear model with or without the 3rd order polynomial transform? Briefly explain your reasoning.