Homework #2: Machine Learning

Kyler Little

February 2, 2018

Problem #1

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

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

(b) For the case $h(\boldsymbol{x}) = \theta(\boldsymbol{w}^T \boldsymbol{x})$, argue that minimizing the in-sample error in part (a) is equivalent to minimizing the one in (3.9).
For two probabilistic distributions $p, 1-p$ and $q, 1-q$ with binary outcomes, the cross-entropy (from information theory) is:

$$p \log \frac{1}{q} + (1-p) \log \frac{1}{1-q} \tag{2}$$

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

Problem #2

Recall the objective function for linear regression can be expressed as

$$E(w) = \frac{1}{N} ||Xw - y||^2 \tag{3}$$

as in Equation (3.3) of LFD. Minimizing this function with respect to w leads to the optimal w as $(XTX)^{-1}X^T y$. This solution holds only when $X^T X$ is

nonsingular. To overcome this problem, the following objective function is commonly minimized instead:

$$E_2(w) = ||Xw - y||^2 + \lambda||w||^2 \tag{4}$$

where $\lambda > 0$ is a user-specified parameter. Please do the following:
(a) Derive the optimal w that minimize $E_2(w)$.
(b) Explain how this new objective function can overcome the singularity problem of X^TX.

Problem #3

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

$$E(w) = \frac{1}{N}\sum_{n=1}^{N} \ln(1 + e^{-y_n w^T x_n}) \tag{5}$$

(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:

$$\text{Predicted class of x} = \begin{cases} 1 & \text{if } \theta(w^T x) \geq 0.5 \\ -1 & \text{if } \theta(w^T x) < 0.5 \end{cases} \tag{6}$$

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

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

$$\text{Predicted class of x} = \begin{cases} 1 & \text{if } \theta(w^T x) \geq 0.9 \\ -1 & \text{if } \theta(w^T x) < 0.9 \end{cases} \tag{7}$$

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?

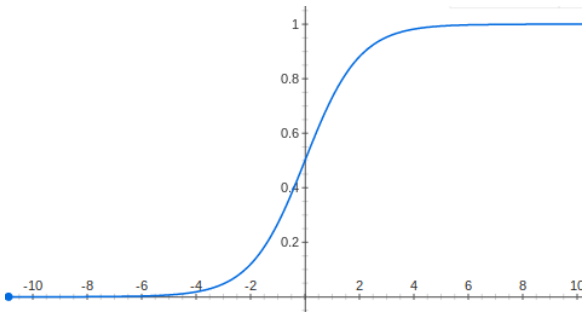

Figure 1: A basic sigmoid curve

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.