STAT 535 Homework 4
Out October 26, 2023
Due November 2, 2023

mmp@stat.washington.edu

Problem 1 – Decision regions for the neural network
In this problem, the inputs are of the form $[x_1\ x_2]^T \in \mathbb{R}^2$ and if necessary we introduce the dummy variable $x_0 \equiv 1$.

a. Consider the following two-layer neural network

$$f(x) = \beta_0 + \sum_k \beta_k z_k \tag{1}$$

$$z_k = \phi(\sum_{j=0}^{2} w_{jk} x_j), \text{ for } k = 1 : K \tag{2}$$

$$x = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix} \tag{3}$$

$$W = [w_{jk}] = \begin{bmatrix} 1 & 0 & 2 & 2 & 2 \\ 1 & 1 & 0 & -1 & -0.5 \\ -1 & 1 & -1 & 0 & 1 \end{bmatrix} \times 20 \tag{4}$$

$$\phi(u) = \frac{1}{1+e^{-u}} \text{ the sigmoid function} \tag{5}$$

$$\beta_0 = -4.9,\ \beta_{1:5} = 1 \tag{6}$$

Plot the decision regions of this neural network, i.e the regions $D_\pm = \{x \,|\, f(x) \lessgtr 0\}$ and the decision boundary $\{x \,|\, f(x) = 0\}$.

b. Repeat the plots for $\beta_0 = -3.9$.

Problem 2 – The sigmoid function, logistic regression, and the logit function – NOT GRADED

These are simple exercises to familiarize yourself with the relationships between these functions. It's quite likely that you have encountered them in other courses

$$\text{Sigmoid}\phi(u) = \frac{1}{1+e^{-u}} \quad \text{Logit}\psi(\theta) = \ln(1+e^\theta) \quad \text{Logistic regression}\frac{P[Y=1|X=x]}{P[Y=0|X=x]} = \theta(x) \equiv \beta^T x \tag{7}$$

Here $Y \in \{0, 1\}$ (this simplifies some expressions).

a. Show that $\phi(-u) = 1 - \phi(u)$, $\phi'(-u) = \phi'(u)$, $\operatorname{argmax}_{u\in\mathbb{R}} \phi'(u) = 0$, $\max_u \phi'(u) = 1/2$.

b. Check also that $\phi(u) = e^{u/2}/(e^{u/2} + e^{-u/2}) = e^u/(1 + e^u)$.

c. Let $P_\theta(y) = e^{\theta y - \psi_0(\theta)}$ where $\psi_0(\theta) = \ln Z(\theta)$ and $Z(\theta)$ is the normalization constant for P_θ. Show that $\psi_0 = \psi$ the logit function.

d. Calculate $E_\theta[Y]$ and identify it with one of the three functions above.

Problem 3 – Logit loss gradient and Hessian - NOT GRADED
Note that here we switched to $y = \pm 1$ The *logit loss*

$$\hat{L}_{\text{logit}}(w) \;=\; \ln(1 + e^{-yw^T x}),\; x, w \in \mathbb{R}^d,\; y = \pm 1 \tag{8}$$

is the negative log-likelihood of observation (x, y) under the logistic regression model $P(y = 1|x, w) = \phi(w^T x)$ where ϕ is the logistic function.

a. Show that the partial derivatives $\frac{\partial \hat{L}_{\text{logit}}}{\partial w_j}$, $\frac{\partial \hat{L}_{\text{logit}}}{\partial x_j}$ for $\hat{L}_{\text{logit}}$ in (8) can be rewritten as

$$\frac{\partial \hat{L}_{\text{logit}}}{\partial w_j} \;=\; -(1 - P(y|x, w))yx_j \tag{9}$$

$$\frac{\partial \hat{L}_{\text{logit}}}{\partial x_j} \;=\; -(1 - P(y|x, w))yw_j. \tag{10}$$

Here x_j, w_j denote the j-th componet of vector $x, w \in \mathbb{R}^d$. *The elegant formulas above hold for a larger class of statistical models, called* <u>*Generalized Linear Models*</u>*, as shown in Lecture II*

b. Assume now that you have a data set $\mathcal{D} = \{(x^i, y^i), i = 1 : n\}, x^i, w \in \mathbb{R}^d$. Show that the gradient of $\hat{L}_{\text{logit}}(w; \mathcal{D})$ is a linear combination of the x^i vectors.

c. Calculate the expression of $\nabla^2 \hat{L}_{\text{logit}}$, the Hessian of $_{\text{logit}}$, for a single data point x. Simplify your result using $\phi(yw^T x)$ and its derivatives conveniently.