STAT 535 Homework 3
Out October 22, 2020
Due October 29, 2020

mmp@stat.washington.edu

Problem 1 – Logit loss and backpropagation - NOT GRADED
The *logit loss*

$$L_{logit}(w) \;=\; \ln(1 + e^{-yw^T x}),\; x, w \in \mathbb{R}^n,\; y = \pm 1 \tag{1}$$

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 $\frac{\partial L_{logit}}{\partial w_i}$, $\frac{\partial L_{logit}}{\partial x_i}$ for L_{logit} in (1) can be rewritten as

$$\frac{\partial L_{logit}}{\partial w_i} \;=\; -(1 - P(y|x, w))yx_i \tag{2}$$

$$\frac{\partial L_{logit}}{\partial x_i} \;=\; -(1 - P(y|x, w))yw_i. \tag{3}$$

The elegant formulas above hold for a larger class of statistical models, called Generalized Linear Models.

Problem 2 – Logit loss Hessian
Assume now that you have a data set $\mathcal{D} = \{(x^i, y^i), i = 1 : N\}, x^i, w \in \mathbb{R}^n$.

a. Calculate the expression of $\nabla^2 L_{logit}$ for a single data point x. Simplify your result using $\phi(yw^T x)$ conveniently.

b. Show that the gradient of $L_{logit}(w; \mathcal{D})$ is a linear combination of the x^i vectors.

c. Show that if $N < n$ the Hessian of $L_{logit}(w; \mathcal{D})$ has at least one 0 eigenvalue, and conclude that $L_{logit}(w; \mathcal{D})$ is not strongly convex in this case.

d. – Optional, extra credit If $||x^i|| \leq R$, find a constant M sufficiently large so that $\nabla^2 L_{logit}(w; \mathcal{D}) \prec MI_n$.

Problem 3 – 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

$$
\begin{aligned}
f(x) &= \beta_0 + \sum_k \beta_k z_k && (4)\\
z_k &= \phi(\sum_{j=0}^{2} w_{jk} x_j),\ \text{for}\, k = 1:K && (5)\\
x &= \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix} && (6)\\
W &= [w_{jk}] = \begin{bmatrix} 1 & 0 & 2 & 2 & 2 \\ 1 & 1 & 0 & -1 & -0.5 \\ -1 & 1 & -1 & 0 & 1 \end{bmatrix} \times 20 && (7)\\
\phi(u) &= \frac{1}{1+e^{-u}}\ \text{the sigmoid function} && (8)\\
\beta_0 &= -4.9,\ \beta_{1:5} = 1 && (9)
\end{aligned}
$$

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

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

Problem 4 – Ridge regression
In this problem you will perform ridge regression on the function $f^*(x) = x^2 + 1$ on $[0, 1]$. In the file `hw3_rr.dat` you will find a set of N (x^i, y^i) values with $y^i = f^*(x^i)$.

a Let $f(x) = \beta_0 + \beta_1 x$ be the predictor of y; β_0, β_1 will be estimated by Ridge Regression with regularization parameter λ. Denote $\beta_{0,1}(\lambda)$ the result of this estimation. Let the data matrix be the row vector $X = [x^1 \ldots x^N]$, and define the column vector $y = [y^1 \ldots y^N]^T$

Write the expressions of $\beta_0(\lambda)$, $\beta_1(\lambda)$ as functions of X, y, λ.

b Now choose a set of λ values including 0 and N. Calculate $\beta_{0,1}(\lambda)$, $\hat{l}_{LS}(\lambda)$ and $J(\lambda)$. Plot on the same graph $\beta_{0,1}(\lambda)$ vs λ.

c Plot on the same graph $\hat{l}_{LS}(\lambda)$ and $J(\lambda)$ vs λ. Comment on what you observe in the graphs of b, c.