S&DS 365 Midterm Solutions 2021

Yale University, Department of Statistics

March 31, 2021

Problem 1

$$g(A, B, x) = B^T(Ax)^3$$

For the first derivative,

$$\nabla_B g(A, B, x) = (Ax)^3$$

Let $z = Ax$, $g(A, B, x) = B^T z^3$ then $\nabla_Z g = 3z^2 \circ B$ (note $\circ$ means element wise multiplication).

$$\begin{aligned}\nabla_A g(A, B, x) &= (3z^2 \circ B)x^T \\ &= (3(Ax)^2 \circ B)x^T\end{aligned}$$

$$\nabla_x g = A^T(3z^2 \circ B) = A^T(3(Ax)^3 \circ B)$$

Problem 2

$$\nabla L(\theta) = \sum_i d_i \left(\frac{e^{x_i^T \theta}}{1 + e^{x_i^T \theta}} - y_i \right) x_i$$

Problem 3

$$\begin{aligned}P(y_1, \cdots, y_n | (x_i, \sigma_i)_{i=1}^n, \theta) &= \prod_{i=1}^n p(y_i | x_i, \sigma_i, \theta) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma_i^2}(y_i - x_i^T\theta)^2}\end{aligned}$$

$$l(\theta) = \sum_{i=1}^n -\frac{1}{2} \log 2\pi\sigma_i^2 - \left(\frac{(y_i - x_i^T\theta)^2}{2\sigma_i^2} \right)$$

Problem 4

$$\theta_{k+1} = \theta_k - \eta_k(\nabla f(\theta_k) - \lambda\theta)$$

Problem 5

a) For each flower, we compute the euclidean distance in terms of the sum of square differences of the features. For the 5 closest data points, we classify the data point as the the most common species.

b) Overfit, each point is assigned single nearest neighbour. Will have low bias but high variance resulting in low training error but poor test error.

c) Underfit, each point is assigned mode of trained set. Will have high bias and low variance resulting in poor test error.

Problem 6

By chain rule,

$$\nabla l(x_i^T\theta, y_i) = l'(x_i^T\theta, y_i)x_i$$

we then have for each individual x_i

$$\nabla l_i(\theta) = \begin{cases} x_i & \text{if } x_i^T\theta > y_i + 1 \\ (x_i^T\theta - y_i\theta)x_i & \text{if } |x_i^T\theta - y_i| \leq 1 \\ -x_i & \text{if } x_i^T\theta < y_i - 1 \end{cases}$$

and for the whole data set,

$$\nabla l(\theta) = \frac{1}{n}\sum_i \nabla l_i(\theta)$$

and the gradient update is

$$\theta_{k+1} = \theta_k - \eta_k \nabla l(\theta_k)$$

Problem 7

This likelihood is only non zero when each $y_i \geq x_i^T\theta$

$$P(y_1, \cdots, y_n | x_1, \cdots, x_n, \theta) = \prod_{i=1}^n p(y_i|x_i, \theta) = \prod_{i=1}^n 1_{y_i \geq x_i^T\theta} e^{-(y_i - x_i^T\theta)}$$

$$l(\theta) = \sum_{i=1}^n -(y_i - x_i^T\theta) + \log 1_{y_i \geq x_i^T\theta}$$

Problem 8

Log likelihood is

$$l(\sigma^2) = \sum_i -\frac{1}{2}\log 2\pi\sigma^2 - \frac{x_i^2}{2\sigma^2}$$

we will work with the variable σ^2 as our deriving variable so let $u = \sigma^2$

$$\begin{aligned} l'(u) &= \sum_i -\frac{1}{2u} + \frac{x_i^2}{2u^2} = 0 \\ -\frac{n}{2u} &+ \frac{\sum_i x_i^2}{2u^2} = 0 \\ -n &+ \frac{\sum_i x_i^2}{u} = 0 \\ u &= \frac{\sum_i x_i^2}{n} \end{aligned}$$

could also derive for σ directly

$$\begin{aligned} \frac{\partial l(\sigma^2)}{\partial \sigma} &= \sum_i -\frac{1}{\sigma} + \frac{x_i^2}{\sigma^3} \\ -n &+ \frac{\sum_i x_i^2}{\sigma^2} = 0 \\ \sigma &= \sqrt{\frac{\sum_i x_i^2}{n}} \end{aligned}$$

and therefore

$$\hat{\sigma}^2 = \frac{\sum_i x_i^2}{n}$$

Problem 9

Each y_i has k entries, let $y_i(j)$ be the entry which is one and all other entries are zero. There are k possible values y_i can take each with probability p_j, so we have

$$\begin{aligned} P(y_1, \cdots, y_n) &= \prod_{i=1}^{n} p_{y_i(j)} \\ &= \prod_{i=1}^{n} \frac{e^{\theta_{y_i(j)}}}{\sum_{s=1}^{k} e^{\theta_s}} \end{aligned}$$

log likelihood is then

$$l(\theta) = \sum_{i=1}^{n} \theta_{y_i(j)} - \log\left(\sum_{s=1}^{k} e^{\theta_s}\right)$$

let $n_j = \sum_{i=1}^{n} 1_{y_i(j)=1}$ count how many observations have the j entry as 1, and $\sum_{j=1}^{k} n_j = n$ counts all observations. We have

$$l(\theta) = \sum_{j=1}^{k} \theta_j n_j - n \log\left(\sum_{s=1}^{k} e^{\theta_s}\right)$$

$$\frac{\partial l}{\partial \theta_j}(\theta) = n_j - n\frac{e^{\theta_j}}{\sum_{s=1}^{k} e^{\theta_s}} = 0$$
$$\frac{e^{\theta_j}}{\sum_{s=1}^{k} e^{\theta_s}} = \frac{n_j}{n}$$

Thus, the estimates $\widehat{\theta}_j$ must satisfy the above equality. We will make the assumption that $\sum_{s=1}^{k} e^{\widehat{\theta}_s} = 1$ and solve. Thus, the above equality becomes

$$e^{\widehat{\theta}_j} = \frac{n_j}{n}$$

Therefore,

$$\widehat{\theta}_j = \log \frac{n_j}{n}$$

We can now verify that indeed our choice does satisfy the assumption we made above.

$$\sum_{s=1}^{k} e^{\widehat{\theta}_s} = \sum_{s=1}^{k} \frac{n_s}{n}$$
$$= 1$$

Thus, the solution $\widehat{\theta}_j = \log \frac{n_j}{n}$ satisfies the required properties. Note that the solution is not unique. Any choice of vector $\bar{\theta} = \widehat{\theta} + c\mathbf{1}$ for $c \in \mathbb{R}$ and $\mathbf{1}$ the all ones vector is also a valid solution.

Problem 10

h is the NN classifier trained previously. We must standardize our new input and we have prediction

$$\hat{y} = h\left(\frac{x - m}{s}\right)$$