MDSC - 301

Sri Sathya Sai Institute of Higher Learning

(Deemed to be University)

Department of Mathematics and Computer Science

Muddenahalli Campus

Course: **M.Sc. Data Science and Computing** Date: **September 10, 2022**

Subject: **Machine Learning** Module : **Linear Models for Regression**

Answer the following:

(1) Consider the sample points $\{\mathbf{x}_i, y_i\}_{i=1}^n$ satisfies the following model:

$$Y_i = \mathbf{w}^T\mathbf{x}_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2),$$

where Y_i assumes the value y_i. Answer the following:

(a) Prove that each $Y_i \sim N(\mathbf{w}^T\mathbf{x}_i, \sigma^2),\ i = 1, 2, \ldots, n$

(b) Define the likelihood function $\mathcal{L}(\mathbf{w}|\mathbf{x}, y, \sigma^2)$.

(c) Prove that $\widehat{\mathbf{w}} = (X^TX)^{-1}X^TY$ is the maximum likelihood estimation of $\mathbf{w}$, where X and Y is given below:

$$X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}, \quad Y = (Y_1, \ldots, Y_n)^T$$

(d) Prove that the maximum likelihood estimator for σ^2 is

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{w}^T\mathbf{x}_i)^2$$

(e) Prove that $Cov(\widehat{\mathbf{w}}) = \sigma^2(X^TX)^{-1}$

(2) Assume that a dataset of n binary values, $x_1, \ldots, x_n$, was sampled from a Bernoulli. Compute the maximum likelihood estimate for the Bernoulli parameter.

(3) Prove that minimizing the function $\tilde{\mathcal{L}}(\mathbf{w}, \mathbf{x}) = \mathcal{L}(\mathbf{w}, \mathbf{x}) + \frac{\lambda}{2}\|\mathbf{w}\|_p^p$ is equivalent to minimizing function $\mathcal{L}(\mathbf{w}, \mathbf{x})$ subject to the constraint $\sum_{j=1}^{k}|w_j|^p \leq \eta$ for an appropriate value of the parameter η. Discuss the relation between the parameters η and λ. (Hint: Use Lagrange Multipliers)

(4) Show that the *tanh* and the logistic sigmoid function are related by $\tanh(a) = 2\sigma(2a) - 1$. Hence show that a general linear combination of logistic sigmoid functions of the form

$$y(x, w) = w_0 + \sum_{j=1}^{k} w_j\sigma\Big(\frac{x - \mu_j}{s}\Big)$$

is equivalent to a linear combination of tanh functions of the form

$$y(x, u) = u_0 + \sum_{j=1}^{k} u_j \tanh\Big(\frac{x - \mu_j}{2s}\Big)$$

and find expressions to relative the new parameters $\{u_0, \ldots, u_k\}$ to the original parameters $\{w_0, \ldots, w_k\}$.

(5) Show that the matrix $H = X(X^T X)^{-1} X^T$ takes any vector v and projects it onto the space spanned by the columns of X.

(6) Find the linear regression model for the following data:

x	y
1	4.8
3	11.3
5	17.2

(7) Suppose we have collection of (n = 100 observations) data points containing a single predictor and a quantitative response. If we fit a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = w_0 + w_1 X + w_2 X^2 + w_3 X^3 + \epsilon$

(a) Suppose that the true relationship between X and Y is linear, i.e. $Y = w_0 + w_1 X + \epsilon$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

(b) Answer (a) using test rather than training RSS.

(c) Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

(d) Answer (c) using test rather than training RSS.

(8) Prove that in the case of simple regression of Y on X, the R^2 statistics is equivalent to the square of $Corr(X, Y)$.

* * * * *