| | | |
|---|---|---|
| **Name, Vorname:** Wang, Jiahui | **Fach:** B.Sc.Mathematik | **Martriknr.**2992080 |
| **Name, Vorname:** Zhao, Jiaqi | **Fach:** M.Sc.Autonome System | **Martriknr.** 3470190 |
| **Name, Vorname:** Wang, Ziyin | **Fach:** M.Sc.Informatik | **Martriknr.** 3435397 |

# Assignment 4

## Task 1: Classification with Linear Regression

**1.** In this model we define the following matrixs:

$$X = \begin{pmatrix} \bar{x_1}^\top \\ \vdots \\ \bar{x_n}^\top \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} 1 & -1.0 \\ 1 & -2.0 \\ 1 & 0.3 \\ 1 & 0.6 \\ 1 & 3.0 \\ 1 & 6.0 \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

which $X$ represents the input variables matrix, $Y$ represents the $6x2$ indicator response matrix and in matrix $\beta$ that $\beta_0$ is for class label 0, $\beta_1$ is for class label 1.
Substitute $X, Y$ into formula $\beta = (X^\top X)^{-1} X^\top Y$ we can get:

$$\beta \approx \begin{pmatrix} 0.6937 & 0.3063 \\ -0.0235 & 0.0235 \end{pmatrix}$$

**2.** In this task we substitute the above $\beta$ into the formula $\hat{Y} = X\beta$ to calculate the predict output, then we can get :

$$\hat{Y} = X\beta = \begin{pmatrix} 1 & -1.0 \\ 1 & -2.0 \\ 1 & 0.3 \\ 1 & 0.6 \\ 1 & 3.0 \\ 1 & 6.0 \end{pmatrix} \begin{pmatrix} 0.6937 & 0.3063 \\ -0.0235 & 0.0235 \end{pmatrix} \approx \begin{pmatrix} 0.7172 & 0.2828 \\ 0.7408 & 0.2592 \\ 0.6867 & 0.3133 \\ 0.6796 & 0.3204 \\ 0.6231 & 0.3769 \\ 0.5526 & 0.4474 \end{pmatrix}$$

so we can see that the values of each row in $\hat{Y}$ add to 1. And in each row, the left value is larger than the right value, then we can get the predict label output $[1, 1, 0, 0, 0, 0]$ by calculating the fitted value function $\hat{f}(x)$, and it does not equal to $y$, finally we can use the argmax function to classify the label.

**3.** Classification with linear regression has the advantage, it satisfies the conditional expectation, and fitted values sum to 1. And the disadvantages are, the individual fitted value could be possibly larger than 1 or less than 1.If the individual value is larger than 1, that will make no sense. Also, for classification the linear regression "Masking"problem. Thay means, if we do more than 2 classes by using linear regression, e.g. 3 classes, the middle class will never dominate.

## Task 2: Log-likelihood gradient and Hessian

1. Compute $\frac{\partial}{\partial\beta}L(\beta)$:
   Let

$$p_i = p(y_i = 1|x_i) = p(x_i) \rightsquigarrow L(\beta) = -\sum_{i=1}^{n}[y_i \log p_i + (1-y_i)\log(1-p_i)]$$

Hint: $\frac{\partial}{\partial z}\sigma(z) = \sigma(z)(1-\sigma(z))$

$$\begin{aligned}\frac{\partial}{\partial\beta}\log p_i &= \frac{\partial \log p_i}{\partial p_i}\cdot\frac{\partial p_i}{\partial f(x)}\cdot\frac{\partial f(x)}{\partial\beta}\\ &= \frac{1}{p_i}\cdot p_i(1-p_i)\cdot\phi^\top(x_i)\\ &= (1-p_i)\phi^\top(x_i)\end{aligned}$$

where $\frac{\partial p_i}{\partial f(x)} := \frac{\partial}{\partial z}\sigma(z)$ with $z = f(x)$.

$$\begin{aligned}\frac{\partial}{\partial\beta}\log(1-p_i) &= \frac{\partial \log(1-p_i)}{\partial p_i}\cdot\frac{\partial p_i}{\partial f(x)}\cdot\frac{\partial f(x)}{\partial\beta}\\ &= -\frac{1}{1-p_i}\cdot p_i(1-p_i)\cdot\phi^\top(x_i)\\ &= -p_i\phi^\top(x_i)\end{aligned}$$

so

$$\begin{aligned}\frac{\partial}{\partial\beta}L(\beta) &= -\sum_{i=1}^{n}[y_i\cdot(1-p_i)\phi^\top(x_i) + (1-y_i)\cdot(-p_i)\phi^\top(x_i)]\\ &= -\sum_{i=1}^{n}[(y_i-p_i)\phi^\top(x_i)]\\ &= \sum_{i=1}^{n}[(p_i-y_i)\phi^\top(x_i)]\end{aligned}$$

2. Compute $\frac{\partial^2}{\partial\beta^2}L(\beta)$:

$$\frac{\partial}{\partial\beta}y_i\phi^\top(x_i) = 0$$

so

$$\begin{aligned}\frac{\partial^2}{\partial\beta^2}L(\beta) &= \frac{\partial}{\partial\beta}\sum_{i=1}^{n}[p_i\phi^\top(x_i)]\\ &= \sum_{i=1}^{n}[\phi(x_i)\cdot p_i(1-p_i)\phi^\top(x_i)]\end{aligned}$$