Subtask B - MLE for Logistic regression

1. Find the form of the log-likelihood.
2. Compute the gradient of logistic function with respect to β

Introduction

In the logistic regression model, the output variable y_i is a Bernoulli random variable (it can take only two values, either 1 or 0) and

$$\mathrm{P}\left(y_i=1 \mid x_i\right)=S\left(x_i \beta\right)$$

where

$$S(t)=\frac{1}{1+\exp (-t)}$$

is the logistic function, x_i is a $1 \times K$ vector of inputs and β is a $K \times 1$ vector of coefficients. Furthermore,

$$\mathrm{P}\left(y_i=0 \mid x_i\right)=1-S\left(x_i \beta\right)$$

The vector of coefficients β is the parameter to be estimated by maximum likelihood. We assume that the estimation is carried out with an IID sample comprising N data points

$$\left(y_i, x_i\right) \text { for } i=1, \ldots, N$$

1. Finding form of the log-likelihood

Y takes value of either 0 or 1. We can merge equations from the introduction into

$$P(y_i \mid x_i)=[S(x_i \beta)]^{y_i}\left[1-S(x_i \beta)\right]^{1-y_i}$$

The log-likelihood function, which is the logarithm of the likelihood is then:

$$\log L(\beta)=\sum_{i=1}^N\left[y_i \ln \left(S(x_i \beta)\right)+\left(1-y_i\right) \ln \left(1-S(x_i \beta)\right)\right]$$

This is the form of the log-likelihood for logistic regression (so first, simpler part of the task is complete)

2. Compute the gradient of logistic function with respect to β

To compute the gradient of the log-likelihood with respect to $\boldsymbol{\beta}$, we need to differentiate the log-likelihood with respect to the vector of parameters $\boldsymbol{\beta}$.
Lets start by restating equation we found in first part of the task.

$$\log L(\boldsymbol{\beta})=\sum_{i=1}^N\left[y_i \ln \left(S(\boldsymbol{x}_i \boldsymbol{\beta})\right)+\left(1-y_i\right) \ln \left(1-S(\boldsymbol{x}_i \boldsymbol{\beta})\right)\right]$$

Now, differentiate this with respect to $\boldsymbol{\beta}$:

$$\frac{\partial \log L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}=\sum_{i=1}^N\left[\frac{\partial}{\partial \boldsymbol{\beta}}\left(y_i \ln \left(S(\boldsymbol{x}_i \boldsymbol{\beta})\right)\right)+\frac{\partial}{\partial \boldsymbol{\beta}}\left(\left(1-y_i\right) \ln \left(1-S(\boldsymbol{x}_i \boldsymbol{\beta})\right)\right)\right]$$

For the first term, apply the chain rule:

$$\frac{\partial}{\partial \boldsymbol{\beta}} \left(y_i \ln \left(S(\boldsymbol{x}_i \boldsymbol{\beta}) \right) \right) = y_i \frac{1}{S(\boldsymbol{x}_i \boldsymbol{\beta})} \cdot \frac{\partial}{\partial \boldsymbol{\beta}} S(\boldsymbol{x}_i \boldsymbol{\beta})$$

Now, using the derivative of the logistic function:

$$S'(t) = S(t)(1 - S(t))$$

We get:

$$\frac{\partial}{\partial \boldsymbol{\beta}} S(\boldsymbol{x}_i \boldsymbol{\beta}) = S(\boldsymbol{x}_i \boldsymbol{\beta}) \left(1 - S(\boldsymbol{x}_i \boldsymbol{\beta})\right) \boldsymbol{x}_i$$

So, the gradient of the first term is:

$$y_i \frac{1}{S(\boldsymbol{x}_i \boldsymbol{\beta})} \cdot S(\boldsymbol{x}_i \boldsymbol{\beta}) \left(1 - S(\boldsymbol{x}_i \boldsymbol{\beta})\right) \boldsymbol{x}_i = y_i \left(1 - S(\boldsymbol{x}_i \boldsymbol{\beta})\right) \boldsymbol{x}_i$$

For the second term:

$$\frac{\partial}{\partial \boldsymbol{\beta}} \left((1 - y_i) \ln \left(1 - S(\boldsymbol{x}_i \boldsymbol{\beta})\right) \right) = -(1 - y_i) \frac{1}{1 - S(\boldsymbol{x}_i \boldsymbol{\beta})} \cdot \left(-S(\boldsymbol{x}_i \boldsymbol{\beta})(1 - S(\boldsymbol{x}_i \boldsymbol{\beta}))\boldsymbol{x}_i \right)$$

This simplifies to:

$$-(1 - y_i) S(\boldsymbol{x}_i \boldsymbol{\beta}) \boldsymbol{x}_i$$

Putting it all together, the gradient of the log-likelihood with respect to $\boldsymbol{\beta}$ is:

$$\frac{\partial \log L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} \left[y_i \left(1 - S(\boldsymbol{x}_i \boldsymbol{\beta})\right) \boldsymbol{x}_i - (1 - y_i) S(\boldsymbol{x}_i \boldsymbol{\beta}) \boldsymbol{x}_i \right]$$

This can be simplified to:

$$\frac{\partial \log L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} \left(\left(y_i - S(\boldsymbol{x}_i \boldsymbol{\beta}) \right) \boldsymbol{x}_i \right)$$

This is the gradient of the log-likelihood for logistic regression.