Machine Learning

Assignment 2

Fatemeh Nadi 810101285

November 25, 2022

Problem 1

$$R\left(\Theta\right) = \frac{1}{2N}\left\|y - X\Theta\right\|_2^2 + \Theta^T H\Theta + \Theta^T\Theta + a^T\Theta \tag{1}$$

$$\begin{aligned} R\left(\Theta\right) &= \frac{1}{2N}\left(y - X\Theta\right)^T\left(y - X\Theta\right) + \Theta^T H\Theta + \Theta^T\Theta + a^T\Theta \\ &= \frac{1}{2N}\left(y^T - \Theta^T X^T\right)\left(y - X\Theta\right) + \Theta^T H\Theta + \Theta^T\Theta + a^T\Theta \\ &= \frac{1}{2N}\left(y^T y - y^T X\Theta - \Theta^T X^T y + \Theta^T X\Theta\right) + \Theta^T H\Theta + \Theta^T\Theta + a^T\Theta \end{aligned}$$

Minimizing 1 easily results in

$$\begin{aligned} \frac{\partial R\left(\Theta\right)}{\partial\Theta} &= 0 \\ \frac{1}{2N}\left(-2X^T y + 2X^T X\Theta\right) + 2H\Theta + 2I\Theta + a &= 0 \\ \left(\frac{X^T X}{N} + 2H + 2I\right)\Theta &= \frac{X^T y}{N} - a \end{aligned}$$

$$\rightarrow \boxed{\Theta = \left(\frac{X^T X}{N} + 2H + 2I\right)^{-1}\left(\frac{X^T y}{N} - a\right)}$$

Problem 2

A.
1-norm (also known as L1 norm) is defined by:

$$\|W\|_1 = |w_1| + |w_2| + ... + |w_N|$$

2-norm (also known as L2 norm or Euclidean norm) is defined by:

$$\|W\|_2 = \left(|w_1|^2 + |w_2|^2 + ... + |w_N|^2\right)^{\frac{1}{2}}$$

$\hat{y}$ is the predicted result such that:

$$\hat{y} = w_1 x_1 + w_2 x_2 + ... w_N x_N + b$$

The below function calculates an error without the regularization function:

$$\begin{aligned} L(x, y) &= Error(y, \hat{y}) \\ &= (\hat{y} - y)^2 \\ &= (wx + b - y)^2 \end{aligned}$$

L1 regularization also called Lasso regression, does feature selection, it adds the “absolute value of magnitude” of the coefficient as a penalty term to the loss function.
It does this by assigning insignificant input features with zero weight and useful features with a non zero weight.

$$L(x, y) = Error(y, \hat{y}) + \boxed{\lambda \sum_{i=1}^{N} |w_i|}$$

The L1 regularization term is highlighted in the red box.

It is the preferred choice when having a high number of features as it provides sparse solutions. Even, we obtain the computational advantage because features with zero coefficients can be avoided.

L2 regularization also called Ridge regression adds the “squared magnitude” of the coefficient as the penalty term to the loss function.
it forces the weights to be small but does not make them zero and does non sparse solution.

$$L\left(x, y\right) = Error\left(y, \hat{y}\right) + \boxed{\lambda \sum_{i=1}^{N} w_i^2}$$

In L2 regularization, regularization term is the sum of square of all feature weights as shown above in the blue box.

It can deal with the multicollinearity (independent variables are highly correlated) problems through constricting the coefficient and by keeping all the variables.
It can be used to estimate the significance of predictors and based on that it can penalize the insignificant predictors.

Table 1: Difference between L1 and L2 regularization

	L1	**L2**
1	penalizes sum of absolute value of weights.	penalizes sum of square weights.
2	has a sparse solution	has a non sparse solution
3	is robust to outliers	is not robust to outliers
4	generates model that are simple and interpretable but cannot learn complex patterns	is able to learn complex data patterns

so the main difference between L1 and L2 regularization is that L1 can yield sparse models while L2 doesn’t.

B.

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \left(Y_i - X_i\beta\right)^2 + \lambda \left\|\beta\right\|_2^2$$

Cost function in this problem is:

$$\begin{aligned} J\left(\beta\right) &= \left(Y - A\beta\right)^T \left(Y - A\beta\right) + \lambda\left(\beta^T\beta\right) \\ &= \left(Y^T - \beta^T A^T\right)\left(Y - A\beta\right) + \lambda\left(\beta^T\beta\right) \\ &= Y^TY - Y^TA\beta - \beta^TA^TY + \beta^TA^TA\beta + \lambda\beta^T\beta \end{aligned} \tag{2}$$

Minimizing 2 easily results in:

$$\begin{aligned} \frac{\partial J\left(\beta\right)}{\partial \beta} &= 0 \\ -2A^TY + 2A^TA\beta + 2\lambda\beta &= 0 \\ \left(A^TA + \lambda I\right)\beta &= A^TY \end{aligned}$$

$$\rightarrow \boxed{\beta = \left(A^TA + \lambda I\right)^{-1}\left(A^TY\right)}$$

Problem 3

Logistic Regression is a binary classification and discriminative and linear classifier:

$$g\left(x\right)=\ln\frac{P\left(Y=1|X=x\right)}{P\left(Y=0|X=x\right)}\gtrless_{0}^{1}=0,\quad g\left(x\right)=w_0+w^Tx$$

$g\left(x\right)$ is linear function and $\ln\frac{P(Y=1|X=x)}{P(Y=0|X=x)}$ likes logarithm of the likelihood ratios.

$$\rightarrow P\left(Y=1|X=x\right)=\frac{1}{1+e^{w_0+w^Tx}}$$

If Y can take on more than two values, say k of them, we can still use logistic regression. Instead of having one set of parameters w_0, w, each class k in $[0, k-1]$ will have w_{k0} and vector w_k, and the predicted conditional probabilities will be:

$$P\left(Y=y_k|X=x\right)=\frac{e^{w_{y_k0}+w_{y_k}^Tx}}{\sum_k e^{w_{k0}+w_k^Tx}} \tag{3}$$

A.

It is assumed that:

$$P\left(Y=y_k|X\right)\propto\exp\left(w_{k0}+\sum_{i=1}^{d}w_{ki}X_i\right) for\ k=1,...,K-1$$

The sum of the probabilities in a probability distribution is always 1:

$$\sum_{i=1}^{K}P\left(Y=y_i|X\right)=1 \tag{4}$$

We should have:

$$P\left(Y=y_k|X\right)=1-\sum_{k=1}^{K-1}P\left(Y=y_k|X\right)$$

According to 3 we have:

$$log\frac{P\left(Y=y|X=x\right)}{P\left(Y=K|X=x\right)}=\left(w_{y0}-w_{K0}\right)+\left(w_y-w_K\right)^Tx=W_0+W^Tx \tag{5}$$

$$\boxed{P\left(Y=y|X=x\right)=\exp\left(W_0+W^Tx\right)P\left(Y=K|X=x\right)}$$

We will use this method to continue.

We can calculate the probability of the last class(K):

$$\begin{aligned}
P\left(Y=K|X=x\right) &\overset{4}{=} 1-\sum_{c=1}^{K-1} P\left(Y=c|X=x\right) \\
&\overset{5}{=} 1-\sum_{c=1}^{K-1} P\left(Y=K|X=x\right)\exp\left(W_{c0}+W_c^T x\right) \\
&= 1-P\left(Y=K|X=x\right)\sum_{c=1}^{K-1}\exp\left(W_{c0}+W_c^T x\right) \\
&= \frac{1}{1+\sum_{c=1}^{K-1}\exp\left(W_{c0}+W_c^T x\right)}
\end{aligned}$$

and the probability for any class k:

$$\begin{aligned}
P\left(Y=k|X=x\right) &\overset{5}{=} P\left(Y=K|X=x\right)\exp\left(W_{k0}+W_k^T x\right) \\
&= \boxed{\frac{\exp\left(W_{k0}+W_k^T x\right)}{1+\sum_{c=1}^{K-1}\exp\left(W_{c0}+W_c^T x\right)}}
\end{aligned}$$

B.

The classification rule picks the label with the highest probability:

$$y=y_k^* \; where \; k^*=\underset{k\in\{1,..,K\}}{\arg\max} P\left(Y=y_k|X=x\right)$$

Problem 4

A.

$\bar{x} = \mu = \frac{\sum x_i}{N} = \frac{100}{10} = 10$

$\bar{y} = \frac{\sum y_i}{N} = \frac{564}{10} = 56.4$

$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{(4-10)^2+(9-10)^2+(10-10)^2+(14-10)^2+(4-10)^2+(7-10)^2+(12-10)^2+(22-10)^2+(1-10)^2+(17-10)^2}{10} = \boxed{37.6}$

Let's start by defining a few things:

define the line of best fit as: $\hat{y_i} = \beta_1 x_i + \beta_0$

error function: $S = \sum_{i=1}^{n} (y_i - \hat{y_i})^2 = \sum_{i=1}^{n} (y_i - \beta_1 x_i - \beta_0)^2$

find β_0 by : $\frac{\partial S}{\partial \beta_0} = 0$

$$\begin{aligned} \frac{\partial s}{\partial \beta_0} &= 0 \\ -2\sum_{i=1}^{n}(y_i - \beta_1 x_i - \beta_0) &= 0 \\ \sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \beta_0 &= 0 \\ \sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i - n\beta_0 &= 0 \\ \\ \rightarrow \beta_0 &= \frac{\sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i}{n} \\ &= \boxed{\bar{y} - \beta_1 \bar{x}} \end{aligned} \tag{6}$$

find β_1 by : $\frac{\partial S}{\partial \beta_1} = 0$

$$\begin{aligned} \frac{\partial s}{\partial \beta_0} &= 0 \\ -2 \sum_{i=1}^{n} x_i \left(y_i - \beta_1 x_i - \beta_0 \right) &= 0 \\ \sum_{i=1}^{n} \left(x_i y_i - \beta_1 x_i^2 - \beta_0 x_i \right) &= 0 \\ 6 \sum_{i=1}^{n} \left(x_i y_i - \beta_1 x_i^2 - \left(\bar{y} - \beta_1 \bar{x} \right) x_i \right) &= 0 \\ \sum_{i=1}^{n} \left(x_i y_i - \bar{y} x_i \right) - B \sum_{i=1}^{n} \left(\bar{x} x_i - x_i^2 \right) &= 0 \end{aligned}$$

$$\rightarrow \boxed{\beta_1 = \frac{\sum_{i=1}^{n} x_i y_i - \bar{y} x_i}{\sum_{i=1}^{n} \left(x_i^2 - \bar{x} x_i \right)}} \tag{7}$$

$\beta_1 =$
$\frac{(4\times31-56.4\times4)+(9\times58-56.4\times9)+(10\times65-56.4\times10)+(14\times73-56.4\times14)+(4\times37-56.4\times4)+(7\times44-56.4\times7)+(12\times60-56.4\times12)+(22\times91-56.4\times22)+(1\times21-56.4\times1)+(17\times84-56.4\times17)}{(4^2-10\times4)+(9^2-10\times9)+(10^2-10\times10)+(14^2-10\times14)+(4^2-10\times4)+(7^2-10\times7)+(12^2-10\times12)+(22^2-10\times22)+(1^2-10\times1)+(17^2-10\times17)+}$
$\simeq \boxed{3.47}$

$\beta_0 = 56.4 - 3.47 \times 10 \simeq \boxed{21.7}$

Simple Linear Regression:

$$\boxed{Y = 3.47X + 21.7}$$

B.

$\beta_1 = \frac{\sum(x_i-\bar{x})y_i}{\sum(x_i-\bar{x})^2}$

proof:

$$\frac{\sum\left(x_i-\bar{x}\right)y_i}{\sum\left(x_i-\bar{x}\right)^2} = \frac{\sum x_i y_i - \sum \bar{x} y_i}{\sum x_i^2 - 2\sum x_i \bar{x} + \sum \bar{x}^2}$$

$$= \frac{\sum x_i y_i - \bar{x}\sum y_i}{\sum x_i^2 - 2\bar{x}\sum x_i + n\bar{x}^2}$$

$$= \frac{\sum x_i y_i - \frac{\sum x_i}{n}\sum y_i}{\sum x_i^2 - 2\frac{\sum x_i}{n}\sum x_i + n\bar{x}\frac{\sum x_i}{n}}$$

$$= \frac{\sum x_i y_i - \frac{\sum y_i}{n}\sum x_i}{\sum x_i^2 - \frac{2(\sum x_i)^2}{n} + n\left(\frac{\sum x_i}{n}\right)^2}$$

$$= \frac{\sum x_i y_i - \bar{y}\sum x_i}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}$$

$$= \boxed{\frac{\sum x_i y_i - \bar{y}\sum x_i}{\sum x_i^2 - \bar{x}\sum x_i}}7$$

$$\begin{aligned} Var\left(\beta_1\right) &= \frac{\sum\left(x_i-\bar{x}\right)^2 Var\left(y_i\right)}{\left[\sum\left(x_i-\bar{x}\right)^2\right]^2} \\ &= \frac{\sigma^2}{\sum\left(x_i-\bar{x}\right)^2} \\ &= \frac{37.6}{376} \\ &= \boxed{0.1} \end{aligned}$$

$$\begin{aligned} Var\left(\bar{y}\right) &= Var\left(\frac{1}{n}\sum y_i\right) \\ &= \frac{1}{n^2}\sum Var\left(y_i\right) \\ &= \frac{\sigma^2}{n} \end{aligned} \tag{8}$$

$$\begin{aligned} Cov\left(\bar{y},\beta_1\right) &= Cov\left[\frac{1}{n}\sum y_i, \frac{\sum\left(x_i-\bar{x}\right)y_i}{\sum\left(x_i-\bar{x}\right)^2}\right] \\ &= \frac{1}{n}\,\frac{1}{\sum\left(x_i-\bar{x}\right)^2}\,Cov\left[\sum y_i, \sum\left(x_i-\bar{x}\right)y_i\right] \\ &= \frac{1}{n}\,\frac{1}{\sum\left(x_i-\bar{x}\right)^2}\,\boxed{\sum\left(x_i-\bar{x}\right)}\sum Cov\left(y_i,y_i\right) \\ &= 0 \end{aligned} \tag{9}$$

$\beta_0 = \bar{y} - \beta_1\bar{x}$

$$\begin{aligned} Var\left(\beta_0\right) &= Var\left(\bar{y}-\beta_1\bar{x}\right) \\ &= Var\left(\beta_0\right) = Var\left(\beta_0\right) = Var\left(\bar{y}\right) - \bar{x}^2 Var\left(\beta_1\right) - 2\bar{x}Cov\left(\bar{y},\beta_1\right) \\ &\overset{9,8}{=} \frac{\sigma^2}{n} - \frac{\sigma^2\bar{x}^2}{\sum\left(x_i-\bar{x}\right)^2} - 0 \\ &= \frac{37.6}{10} - \frac{37.6\times 100}{376} \\ &= \boxed{-6.24} \end{aligned}$$

C.

$$\begin{aligned} Cov\left(\beta_0,\beta_1\right) &= Cov\left(\bar{y}-\beta_1\bar{x},\beta_1\right) \\ &= Cov\left(\bar{y},\beta_1\right) - \bar{x}Cov\left(\beta_1,\beta_1\right) \\ &\overset{9}{=} 0 - \bar{x}Var\left(\beta_1\right) \\ &= -10\times 0.1 \\ &= -1 \end{aligned} \tag{10}$$

$$\begin{aligned} Correlation\left(\beta_0,\beta_1\right) &= \frac{Cov\left(\beta_0,\beta_1\right)}{\sigma_{\beta_0}\sigma_{\beta_1}} \\ &\overset{10}{=} \frac{-1}{\sqrt{13.76}\sqrt{0.1}} \\ &\simeq -0.85 \end{aligned}$$

Problem 5:

Please see this file: ”Q5.ipynb”
Code and explanation are provided.

A.
You can see those in this file: ”Q5.ipynb” + conclusion.

B.
One-vs-all is a strategy that involves training N distinct binary classifiers, each designed to recognize a specific class. After that we collectively use those N classifiers to predict the correct class. How do we do it in code? By considering one class as 1 and rest all as 0, we train the model and get the requisite wights. We store the value of weights in a dictionary format for each classifiers. Then by the help of Sigmoid Function we calculate the probability. Highest probability takes a presidency and we classify that data to corresponding classifier.

Table 2: Report some scores from my Logistic Regression

	precision	**recall**	**f1-score**	**jaccard**	**support**
Adelie	0.93	1	0.96	0.93	27
Chinstrap	1	0.92	0.96	0.92	12
Gentoo	1	0.96	0.98	0.96	27
accuracy			$\boxed{0.97}$		66

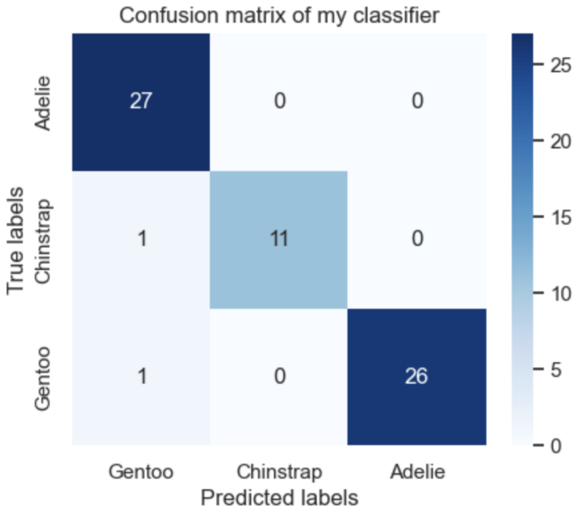


Figure 1: Q5-Confusion matrix of my Logistic Regression model

C.

Table 3: Report some scores from Sklearm's Logistic Regression

	precision	**recall**	**f1-score**	**jaccard**	**support**
Adelie	0.97	1	0.99	0.97	36
Chinstrap	1	0.75	0.86	0.75	12
Gentoo	1	1	1	1	21
accuracy			$\boxed{0.957}$		69

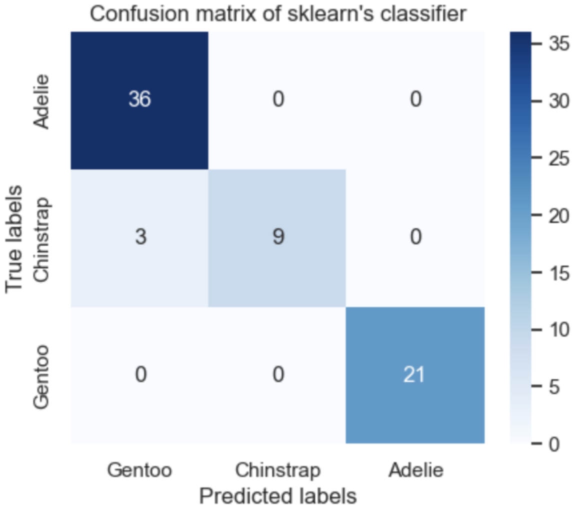
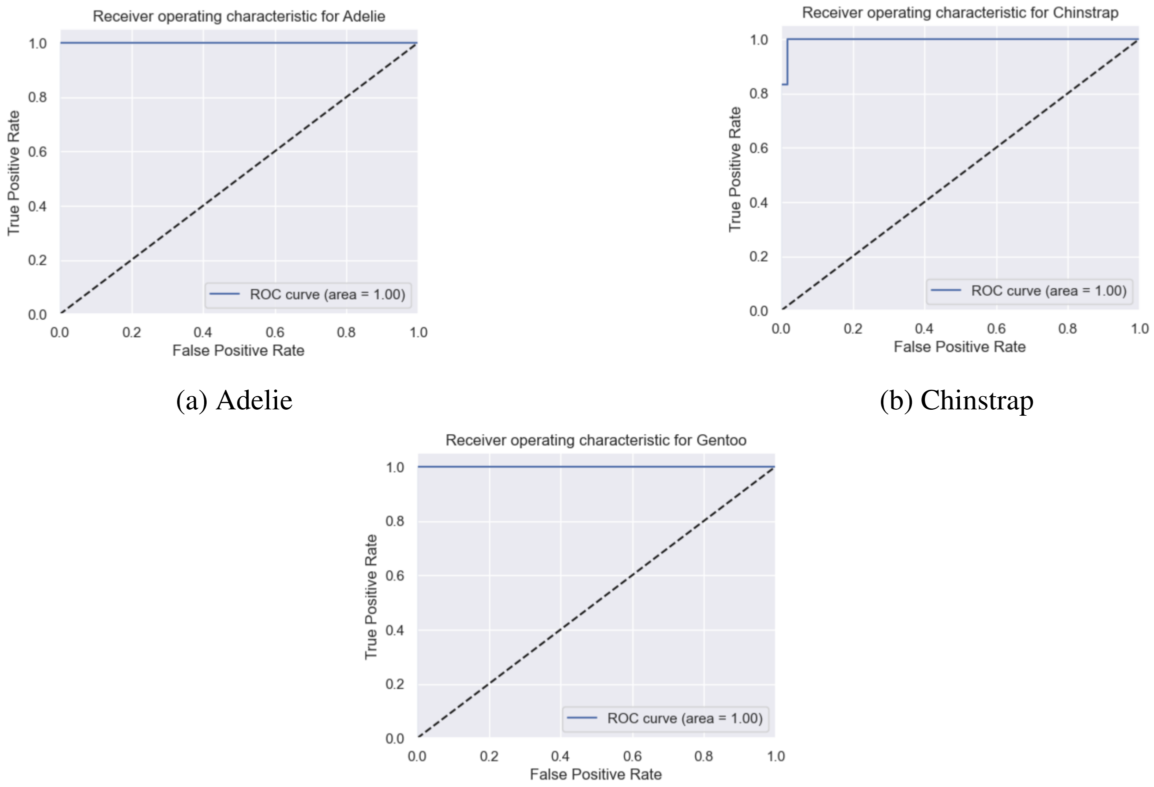


Figure 2: Q5-Confusion matrix of Slearn's Logistic Regression model

ROC curve



(a) Adelie

(b) Chinstrap

(c) Gentoo

Figure 3: Three ROC for each class

Area Under the ROC Curve is close to one. so this method and features are good for separating classes.

Problem 6

Please see this file: ”Q6.ipynb”
Code and explanation are provided.