Machine Learning Methodologies and Applications (AI6012) Individual Assignment

Leon Sun (G2204908A)

30 September 2023

1 Question 1 (10 marks)

Multi-class classification, or Multinomial Logistic Regression, can be approached using softmax regression. The softmax function is generally defined as:

$$P(y=c \mid x)=\frac{\exp \left(w^{(c)} x\right)}{\sum_{i=1}^{C} \exp \left(w^{(i)} x\right)} \quad or \quad \frac{1}{1+\sum_{i \neq C}^{C} \exp \left(w^{i}\right) . x} \tag{1}$$

Since the sum of all the conditional probabilities for the softmax is 1, we can summarise the probabilities for all classes to:

$$\sum_{c=0}^{C} \quad P(y=c \mid x)=1 \tag{2}$$

By introducing the set of logits into we can arrived at the following parametric equations for multinomial logistic regression. Suppose there are C classes, 0, 1, ..., C-1:

$$For\ c>0: \quad P(y=c \mid x)=\frac{\exp \left(-w^{(c)^{T}} x\right)}{1+\sum_{c=1}^{C-1} \exp \left(-w^{(c)^{T}} x\right)}=\hat{y}_{c} \tag{3}$$

$$For\ c>0: \quad P(y=0 \mid x)=\frac{1}{1+\sum_{c=1}^{C-1} \exp \left(-w^{(c)^{T}} x\right)}=\hat{y}_{0} \tag{4}$$

Given a set of N training input-output pairs like $\mathrm{x_i}$, $\mathrm{y_i}$, i = 1,..., N, which are i.i.d, we can define the likehood as the product of likelihoods of each individual pairs.

$$\mathcal{L}(\boldsymbol{w_c})=\prod_{i=1}^{N} l\left(\boldsymbol{w_c} \mid\left\{\boldsymbol{x}_{i}, y_{i}\right\}\right)=\prod_{i=1}^{N} P\left(y_{i} \mid \boldsymbol{x}_{i} ; \boldsymbol{w_c}\right) \tag{5}$$

Hence the maximum likelihood estimation can be represented in the following ln function which converts the product into a sum.

$$\hat{\boldsymbol{w}_{c}}=\underset{\boldsymbol{w}_{c}}{\operatorname{argmax}} \prod_{i=1}^{N} P\left(y_{i} \mid \boldsymbol{x}_{i} ; \boldsymbol{w}_{c}\right)=\underset{\boldsymbol{w}_{c}}{\operatorname{argmax}} \sum_{i=1}^{N} \sum_{c=0}^{C-1}\left(y_{i} \ln \left(g\left(\boldsymbol{x}_{i} ; \boldsymbol{w}_{c}\right)\right)\right) \tag{6}$$

$$\ln \hat{w}=-\frac{1}{N} \sum_{i=1}^{N} \sum_{C=1}^{C-1} y_{i} \ln \Big(\quad P(y_{|} x_{j} ; w_{c})\Big) \tag{7}$$

We will now derive the learning procedure for the multinomial logistic classification using Gradient Descent optimisation method.

$$\boldsymbol{w}_{t+1}=\boldsymbol{w}_{t}-\rho \frac{\partial E(\boldsymbol{w})}{\partial \boldsymbol{w}} \tag{8}$$

$$\begin{aligned}
\frac{\partial E(\boldsymbol{w})}{\partial \boldsymbol{w}} &= \frac{\partial\left(-\sum_{i=1}^{N} \sum_{c=1}^{c} y_{i} \cdot \ln \left(P\left(y_{i} \mid x_{i} ; w_{c}\right)\right)\right)}{\partial w_{c}} \\
&= -\sum_{i=1}^{N} \sum_{c=1}^{c} \frac{\partial\left(y_{i} \cdot \ln \left(p\left(y_{i} \mid x_{i} ; w_{c}\right)\right)\right.}{\partial w_{c}}
\end{aligned}$$

Let P(yi | xi; wc) be f(z). Using chain rule:

$$\begin{aligned}\frac{\partial \ln f(z)}{\partial z} &= \frac{\partial \ln f(z)}{\partial f(z)} \cdot \frac{\partial f(z)}{\partial z} \\ &= \frac{1}{f(z)} \cdot \frac{df(z)}{\partial z}\end{aligned} \tag{9}$$

We can differentiate using:

$$\frac{\partial \left(y_i \cdot \ln \left(P \left(y_i \mid x_i; w_c\right)\right)\right.}{\partial w_c} = y_i \cdot \frac{1}{P \left(y_i \mid x_i w_c\right)} \cdot \frac{\partial \left(P \left(y_i \mid x_i; w_c\right)\right.}{\partial w_c} \tag{10}$$

If y = c, subbing equations (3) into (10) gives us:

$$\frac{\partial \left(P \left(y_i \mid x_i; w_c\right)\right.}{\partial w_c} = \frac{\partial}{\partial w_c} \left(\frac{\exp\left(-w_c \cdot x_i\right)}{1 + \sum_{c=1}^{c-1} \exp\left(-w_c \cdot x_i\right)} \right)$$

Using quotient rule:

$$\frac{\partial}{\partial w_c} \left(\frac{\exp\left(-w_c \cdot x_i\right)}{1 + \sum_{c=1}^{c-1} \exp\left(-w_c \cdot x_i\right)} \right)$$

$$= \frac{\left(\frac{\partial}{\partial w_c} \exp\left(-w_c \cdot x_i\right)\right) \left(1 + \sum_{c=1}^{c-1} \exp\left(-w_c \cdot x_i\right)\right) + \left(\frac{\partial}{\partial w_c} \left(1 + \sum_{c=1}^{c-1} \exp\left(-w_c \cdot x_i\right)\right)\right) \left(\exp\left(-w_c \cdot x_i\right)\right)}{\left(1 + \sum_{c=1}^{c-1} \exp\left(-w_c \cdot x_i\right)\right)^2}$$

$$= \frac{\left(x_i \cdot \exp\left(-w_c \cdot x_i\right)\right) \left(1 + \sum_{i=1}^{c-1} \exp\left(-w_c \cdot x_i\right)\right) + x_i \cdot \exp\left(-w_c \cdot x_i\right) \cdot \left(\exp\left(-w_c \cdot x_i\right)\right)}{\left(1 + \sum_{c=1}^{c-1} \exp\left(-w_c \cdot x_i\right)\right)^2}$$

$$= \frac{x_i \exp\left(-W_c \cdot x_i\right) \left(1 + \sum_{c=1}^{c-1} \exp\left(-W_c \cdot x_i\right) - \exp\left(-W_c \cdot x_i\right)\right)}{\left(1 + \sum_{c=1}^{c-1} \exp\left(-W_c \cdot x_i\right)\right)^2}$$

$$= \frac{x_i \cdot \exp\left(-w_c \cdot x_i\right)}{1 + \sum_{c=1}^{c-1} \operatorname{cop}\left(-w_c \cdot x_i\right)} \cdot \frac{1 + \sum_{c=1}^{c-1} \exp\left(-w_c \cdot x_i\right) - \exp\left(-w_c \cdot x_i\right)}{1 + \sum_{c=1}^{c-1} \exp\left(-w_c \cdot x_i\right)}$$

$$= \frac{x_i \cdot \exp\left(-w_c \cdot x_i\right)}{1 + \sum_{c=1}^{c-1} \exp\left(-w_c, x_i\right)} \cdot \left(\frac{1 + \sum_{c=1}^{c-1} \exp\left(-w_c, x_i\right)}{1 + \sum_{c=1}^{c-1} \exp\left(-w_c, x_i\right)} - \frac{\exp\left(-w_c \cdot x_i\right)}{1 + \sum_{c=1}^{c-1} \exp\left(-w_{ci} x_i\right)} \right)$$

$$= x_i \cdot \hat{y}_c (1 - \hat{y}_c) \tag{11}$$

Subbing (11) back into (10):

$$\begin{aligned}&\frac{\partial \left(P \left(y_i \mid x_i; w_c\right)\right.}{\partial w_c} \\ &= y_i \cdot \frac{1}{P \left(y_i \mid x_i w_c\right)} \cdot \frac{\partial \left(P \left(y_i \mid x_i; w_c\right)\right.}{\partial w_c} \\ &= y_i \cdot \frac{1}{\hat{y}_c} \cdot \left(x_i \cdot \hat{y}_c (1 - \hat{y}_c)\right) \\ &= x_i (y_i - y_i \cdot \hat{y}_c)\end{aligned} \tag{12}$$

Putting them back together, we will sub (12) into the gradient descent rule (8)

$$
\begin{aligned}
\boldsymbol{w}_{t+1} &= \boldsymbol{w}_t - \rho\frac{\partial E(\boldsymbol{w})}{\partial \boldsymbol{w}} \\
&= \boldsymbol{w}_t - \rho(-\sum_{i=1}^{N}\sum_{c=1}^{C}\frac{\partial\left(y_i \cdot \ln\left(p\left(y_i \mid x_i; w_c\right)\right)\right.}{\partial w_c} - \lambda \boldsymbol{w}) \\
&= \boldsymbol{w}_t + \rho(\sum_{i=1}^{N}(y_i - y_i \cdot \hat{y}_c)x_i - \lambda \boldsymbol{w})
\end{aligned} \tag{13}
$$

2 Question 2 (5 marks)

2.2. Answer:

C=0.01	C=0.05	C=0.1	C=0.5	C=1
0.84958	0.85038	0.85038	0.85050	0.85032

Table 1: Classification accuracy on running linear kernel SVM on 3-fold cross-validation using training set with different values of the parameter C in {0.01, 0.05, 0.1, 0.5, 1}

2.3. Answer:

	g=0.01	g=0.05	g=0.1	g=0.5	g=1
C=0.01	0.76377	0.76433	0.77096	0.76377	0.76377
C=0.05	0.78871	0.83293	0.83029	0.76961	0.76377
C=0.1	0.83226	0.83754	0.83717	0.79092	0.76377
C=0.5	0.84411	0.84546	0.84559	0.82507	0.77772
C=1	0.84682	0.84589	0.84651	0.82956	0.78668

Table 2: Classification accuracy on running rbf kernel SVM on 3-fold cross-validation using training set with parameter gamma in {0.01, 0.05, 0.1, 0.5, 1} and different values of the parameter C in {0.01, 0.05, 0.1, 0.5, 1}

2.4. Answer:

	kernal=Linear, C=1
Accuracy of SVMs	0.84733

Table 3: Classification accuracy on running linear kernel SVM on 3-fold cross-validation using test set with different values with C=1

3 Question 3 (5 marks)

4 Question 4 (5 marks)

	kernal=Linear, C=1
Accuracy of SVMs	0.84733

Table 4: Classification accuracy on running linear kernel SVM on 3-fold cross-validation using test set with different values with C=1