Machine Learning Methodologies and Applications (AI6012) Individual Assignment

Leon Sun (G2204908A)

29 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 arrive 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.

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

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 $\mathrm{P(y_i \mid x_i; w_c)}$ be f(z). Using chain rule:

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

Applying the chain rule logic, we can obtain the derivative 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:

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

We will apply quotient rule to obtain the first order derivative:

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

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

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

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

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

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

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

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)

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

2 Question 2 (5 marks)

2.2. Answer:

C=0.01	C=0.05	C=0.1	C=0.5	C=1
0.84402	0.84610	0.84644	0.84693	0.84721

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	0.75919	0.81991	0.81985	0.75919	0.75919
C=0.05	0.83121	0.83575	0.83425	0.78916	0.75919
C=0.1	0.83772	0.83965	0.83876	0.80612	0.76199
C=0.5	0.84297	0.84577	0.84681	0.83216	0.78975
C=1	0.84442	0.84675	0.84742	0.83661	0.79829

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:

	kernel=RBF, C=1, gamma=0.1
Accuracy of SVMs	0.84614

Table 3: Classification accuracy on running rbf kernel SVM on 3-fold cross-validation using test set with C=1 and gamma=0.1

3 Question 3 (5 marks)

Linear soft margin SVMs:

$$
\min_{\boldsymbol{w}, b, \xi_i} \frac{\|\boldsymbol{w}\|_2^2}{2} + C \left(\sum_{i=1}^{N} \xi_i \right) \tag{1}
$$

Empirical Structural Risk Minimisation

$$
\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} \ell \left(f\left(\boldsymbol{x}_i; \boldsymbol{\theta}\right), y_i \right) + \lambda \Omega(\boldsymbol{\theta}) \tag{2}
$$

To reformulate the optimisation of linear non SVMs as an instance of empirical structural risk minimisation, we will leverage on hinge loss.

$$\text{if } (w \cdot x_i + b)_{y_i} \geqslant 1,$$

$$\varepsilon_i^*(w,b) = 0 \tag{3}$$

$$\text{if } (w, x; +b) y_i < 1,$$

$$\varepsilon_i^*(w,b) = 1 - (w, x_i + b)\, y_i \tag{4}$$

$$\therefore \varepsilon_i = \max\left(0, 1 - (w_i x_i + b)_{y_i}\right) \tag{5}$$

We can then derive the empirical structural risk minimisation:

$$\sum_{i=1}^{N} \varepsilon_i = \sum_{i=1}^{N} \max\left(0, 1 - (w_i \cdot x_i + b)_{y_i}\right) \tag{6}$$

Substituting them back into the objective will give us the hinge loss reformulation of the linear non SVM.

$$\min_{w,b} \|w\|_2^2 + C \sum_{i=1}^{N} \max\left(0, 1 - (w_i x_i + b)\, y_i\right) \tag{7}$$

4 Question 4 (5 marks)

The regularised linear regression can be represented as an optimisation problem given by this formula:

$$\hat{w} = \arg\min_{w} \frac{1}{2} \sum_{i=1}^{N} (w \cdot x_i - y_i)^2 + \frac{\lambda}{2} \|w\|_2^2 \tag{1}$$

Using the kernel trick, we can map the current dimensional space into a feature space to extend this regularised linear regression for solving non linear problems.

$$\hat{w} = \arg\min_{w} \frac{1}{2} \sum_{i=1}^{N} \quad (w \cdot \Psi(x_i) - y_i)^2 + \frac{\lambda}{2} \|w\|_2^2 \tag{2}$$

To obtain a closed form solution, the derivative of the objective w.r.t. w will be set to 0. Solving the resultant equation:

$$\begin{aligned}
&\frac{\partial \left(\frac{1}{2} \sum_{i=1}^{N} (w \cdot \psi(x_i) - y_i)^2 + \frac{\lambda}{2} |w|_2^2\right)}{\partial w} = 0 \\
&\frac{\partial \left(\frac{1}{2} \sum_{i=1}^{N} (w \cdot \psi(x_i) - y_i)^2\right)}{\partial w} + \frac{\partial \left(\frac{\lambda}{2} |w|_2^2\right)}{\partial w} = 0 \\
&\left(\sum_{i=1}^{N} (\psi(x_i)) \left(\psi(x_i^T)\right)\right) w - \sum_{i=1}^{N} y_i \left(\psi(x_i)\right) + \lambda w = 0
\end{aligned} \tag{3}$$

Let $\sum_{i=1}^{N} y_i \left(\psi(x_i)\right) = K$, where K represents the sum of inner product between mapped instances. Subbing this back to (3) gives us:

$$\begin{aligned}
&\mathbf{K}\mathbf{K}^{\mathbf{T}}\mathbf{w} - \mathbf{K}\mathbf{y} + \lambda \mathbf{I}\mathbf{w} = 0 \\
&(\mathbf{K}\mathbf{K}^{\mathbf{T}} + \lambda \mathbf{I})\mathbf{w} - \mathbf{K}\mathbf{y} = 0 \\
&(\mathbf{K}\mathbf{K}^{\mathbf{T}} + \lambda \mathbf{I})\mathbf{w} = \mathbf{K}\mathbf{y} \\
&\mathbf{w} = (\mathbf{K}\mathbf{K}^{\mathbf{T}} + \lambda \mathbf{I})^{-1}\mathbf{K}\mathbf{y}
\end{aligned} \tag{4}$$