Machine Learning Fall 2020 ——— Homework 4

學號：B07902037 系級： 資工三 姓名：蔡沛勳

1. (0.5%) 請說明你實作之 RNN 模型架構及使用的 word embedding 方法，回報模型的正確率並繪出訓練曲線*。

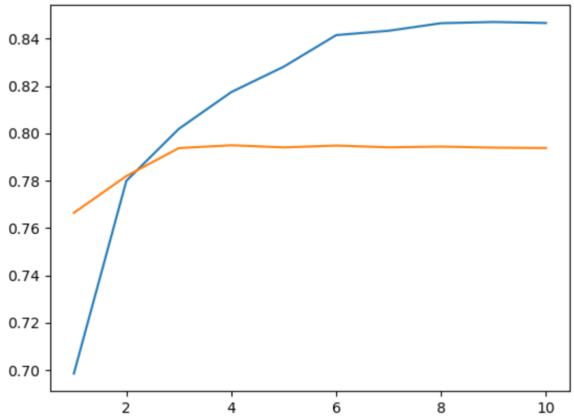
RNN 主要透過 torch.nn 中的 lstm 來實作，具體程式如下:

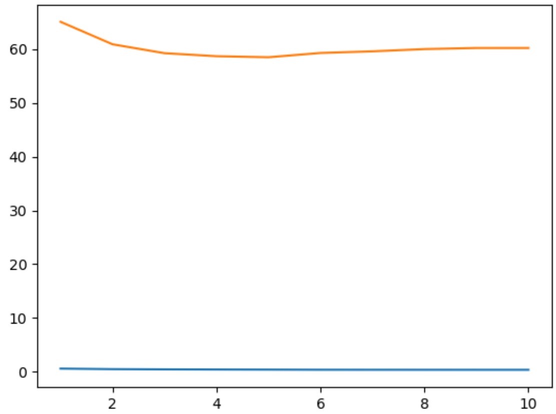
```
class RNN(nn.Module):
        def __init__(self, embedding, hidden_size, bidirectional):
                super(RNN, self).__init__()
                self.embedding = nn.Embedding(embedding.size(0),embedding.size(1))
                self.embedding.weight = nn.Parameter(embedding)
                self.embedding.weight.requires_grad = True

                self.lstm = nn.LSTM(input_size  = 250,
                                    hidden_size = hidden_size,
                                    num_layers  = 5,
                                    batch_first = True,
                                    dropout = 0.5,
                                    bidirectional = bidirectional
                                   )

                self.fc = nn.Sequential(
                        nn.Dropout(p = 0.5),
                        nn.Linear(in_features = 2 * hidden_size if bidirectional else hidden_size,
                                out_features = 1),
                )
                self.sigmoid = nn.Sigmoid()

                def forward(self, inputs):
                        x = self.embedding(inputs)
                        out, _ = self.lstm(x, None)
                        out = out[:,-1,:]
                        outputs = self.fc(out)
                        return self.sigmoid(outputs)
```

Word embedding 則使用 gensim.models.word2vec 的 Word2Vec 來實作，參數如下

```
Word2Vec(data, size = 250, iter = 3, window = 5,  min_count = 3, sg = 1)
```

此 model 的 training / validation history 如下，其中左圖為 accuracy history，右圖為 loss history。 (藍線為 training data，橘線為 validation data，比例為 0.8 : 0.2)。





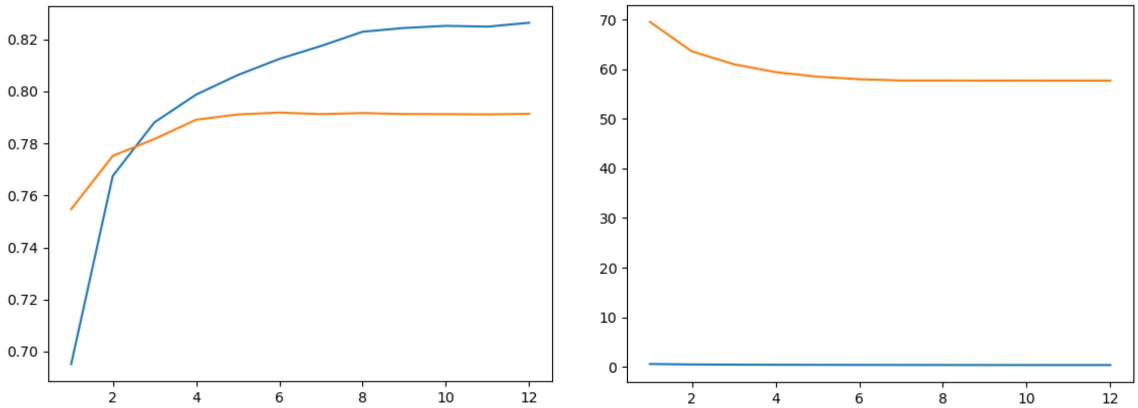
在 kaggle 上的最高分為 0.78340。

2. (0.5%) 請實作 BOW+DNN 模型，敘述你的模型架構，回報模型的正確率並繪出訓練曲線*。

我實作的 DNN 包含兩層 Linear 架構，具體程式如下

```
class DNN(nn.Module):
    def __init__(self, wnum):
        super(DNN, self).__init__()
        self.wnum = wnum    # words number

        self.fc = nn.Sequential(
            nn.Linear(in_features = wnum, out_features = 64),
            nn.LeakyReLU(negative_slope = 0.1),
            nn.Dropout(p = 0.5),
            nn.Linear(in_features = 64, out_features = 1)
        )
        self.sigmoid = nn.Sigmoid()

    def forward(self, inputs):
        x = inputs.float()
        outputs = self.fc(x)
        return self.sigmoid(outputs)
```

此 model 的 training / validation history 如下，其中左圖為 accuracy history，右圖為 loss history。 (藍線為 training data，橘線為 validation data，比例為 0.8 : 0.2)



在 kaggle 上的最高分為 0.78480。

預期中 BOW+DNN 表現應比 RNN(LSTM)要差，因為 BOW+DNN 忽略的語序的影響。而我的結果不如預期的可能原因為 RNN 沒訓練好或語序對本次測資的影響較低，所以兩者表現相差無幾。

3. (0.5%) 請敘述你如何 improve performance（preprocess, embedding, 架構等），並解釋為何這些做法可以使模型進步。

preprocess: 我對設定 Word2Vec 的參數 min_count 增加到 7，所以可以過濾掉部分錯字或罕見單字。調整 windows 大小使一些距離較遠但依舊具有關聯性的單字可以被訓練到。

embedding: 調整句子長度使 model 的判斷資料增加，但也可能導致部分句子中<PAD>太多導致結果變差。

lstm model: biderectional 增加反向序列的測資以增進其表現。

4. (0.5%) 請比較 RNN 與 BOW 兩種不同 model 對於 "Today is hot, but I am happy" 與 "I am happy, but today is hot" 這兩句話的分數（model output），並討論造成差異的原因。

	today is hot , but i am happy	i am happy , but today is hot
RNN	0.71982	0.34323
BOW+DNN	0.66404	0.66404

預期的狀況為"Today is hot, but I am happy"會大於 0.5 而"I am happy, but today is hot"會小於 0.5。而對 BOW+DNN model 兩句話的輸入是相同的，故輸出也會相同。RNN model 則會考量語序而產生出與預期較接近的結果。

5. (3%) Refer to math problem

1.

(a).

設 10 個點依序為$x_1 , x_2 , \ldots , x_{10}$，可算出 μ、Σ 為

$$\mu = \frac{1}{10}\sum_{i=1}^{10} x_i = (5.4\,, 8\,, 4.8)^T$$

$$\Sigma = \frac{1}{10}\sum_{i=1}^{10}(x_i-\mu)(x_i-\mu)^T = \begin{pmatrix} 12.04 & 0.5 & 3.28 \\ 0.5 & 12.2 & 2.9 \\ 3.28 & 2.9 & 8.16 \end{pmatrix}$$

Σ 可被正交對角化得到 $\Sigma = U\Lambda U^T$，可算出 U、Λ 為

$$U = \begin{pmatrix} 0.616596 & 0.678179 & -0.399856 \\ 0.58815 & -0.73439 & -0.337589 \\ 0.522596 & 0.027286 & 0.852144 \end{pmatrix}$$

$$\Lambda = \begin{pmatrix} 15.29744 & 0 & 0 \\ 0 & 11.63052 & 0 \\ 0 & 0 & 5.47203 \end{pmatrix}$$

得到 principle axes 依 eigenvalue 由大到小依序為

$$v_1 = \begin{pmatrix} 0.616596 \\ 0.58815 \\ 0.522596 \end{pmatrix}, v_2 = \begin{pmatrix} 0.678179 \\ -0.73439 \\ 0.027286 \end{pmatrix}, v_3 = \begin{pmatrix} -0.399856 \\ -0.337589 \\ 0.852144 \end{pmatrix}$$

(b).

由上題得到 $W = [v_1\ ,\ v_2]^T = \begin{bmatrix} 0.616596 & 0.58815 & 0.522596 \\ 0.678179 & -0.73439 & 0.027286 \end{bmatrix}$

代入 $\hat{x} = Wx$ 得到

$$\widehat{x_1} = \begin{bmatrix} 3.36068 \\ -0.70873 \end{bmatrix}\ \widehat{x_2} = \begin{bmatrix} 9.78456 \\ -3.02597 \end{bmatrix}\ \widehat{x_3} = \begin{bmatrix} 13.6110 \\ -6.53257 \end{bmatrix}\ \widehat{x_4} = \begin{bmatrix} 7.93478 \\ -5.06051 \end{bmatrix}\ \widehat{x_5} = \begin{bmatrix} 12.3623 \\ -6.83599 \end{bmatrix}$$

$$\widehat{x_6} = \begin{bmatrix} 7.19137 \\ 1.83698 \end{bmatrix}\ \widehat{x_7} = \begin{bmatrix} 14.9579 \\ 0.474065 \end{bmatrix}\ \widehat{x_8} = \begin{bmatrix} 7.07758 \\ -3.81330 \end{bmatrix}\ \widehat{x_9} = \begin{bmatrix} 12.8589 \\ 3.95174 \end{bmatrix}\ \widehat{x_{10}} = \begin{bmatrix} 16.2938 \\ -1.10550 \end{bmatrix}$$

(c).

$$average\ reconstruction\ error = \frac{1}{10}\sum_{i=1}^{10}\|x_i - W^T(Wx_i)\|^2$$

$= 1/10(2.192311 + 0.001839 + 5.835359 + 1.341905 + 25.240965 + 10.882219 + 1.867559 + 9.300916 + 0.95343 + 3.065159\ ) = 6.0681662$

2.

(a).

Symmetric:

$$(AA^T)^T = (A^T)^T\, A^T = AA^T$$

$$(A^TA)^T = A^T\, (A^T)^T = A^TA$$

positive semi-definite:

$$x^TAA^Tx = (x^TA)(A^Tx) = (A^Tx)^T(A^Tx) = \|A^Tx\|^2 \geq 0$$

$$x^TA^TAx = (x^TA^T)(Ax) = (Ax)^T(Ax) = \|Ax\|^2 \geq 0$$

share the same eigenvalues:

設 λ_1 為 (AA^T) 的 eigenvalue，x_1 為對應的 eigenvector，得

$$(AA^T)x_1 = \lambda_1 x_1$$

兩邊同乘 A^T 得

$$A^T(AA^T)x_1 = A^T\lambda_1 x_1$$

$$(A^TA)(A^Tx_1) = \lambda_1(A^Tx_1)$$

可看出 λ_1 為 (A^TA) 的 eigenvalue，(A^Tx_1) 為對應的 eigenvector

設 λ_2 為 (A^Ta) 的 eigenvalue，x_2 為對應的 eigenvector，得

$$(A^TA)x_2 = \lambda_2 x_2$$

兩邊同乘 A 得

$$A(A^TA)x_2 = A\lambda_2 x_2$$

$$(AA^T)(Ax_2) = \lambda_2(Ax_2)$$

可看出 λ_2 為 (AA^T) 的 eigenvalue，(Ax_2) 為對應的 eigenvector

(b).

已知 Σ 為半正定對稱矩陣，故可由 Cholesky decomposition 得

$$\Sigma = LL^T, L \in R^{n*n}$$

設 $z_1 \dots z_{2n}$ 為

$$\begin{bmatrix} \sqrt{n} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{bmatrix} -\sqrt{n} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ \sqrt{n} \\ \vdots \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ -\sqrt{n} \\ \vdots \\ 0 \end{bmatrix} \dots \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \sqrt{n} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ -\sqrt{n} \end{bmatrix}$$

可堆得 mean 及 covariance 為

$$\frac{1}{2n}\sum_{i=1}^{2n} z_i = 0$$

$$\frac{1}{2n}\sum_{i=1}^{2n}(z_i - 0)(z_i - 0)^T = \frac{1}{2n}\sum_{i=1}^{2n} z_i z_i^T = I_n$$

故我們可設$x_i = Lz_i + \mu$，並推得其 mean 及 covariance 為

$$\frac{1}{2n}\sum_{i=1}^{2n} x_i = \frac{1}{2n}\sum_{i=1}^{2n}(Lz_i + \mu) = L\left(\frac{1}{2n}\sum_{i=1}^{2n} z_i\right) + \mu = \mu$$

$$\frac{1}{2n}\sum_{i=1}^{2n}(x_i - \mu)(x_i - \mu)^T = \frac{1}{2n}\sum_{i=1}^{2n}(Lz_i)(Lz_i)^T$$

$$= \frac{1}{2n}\sum_{i=1}^{2n} Lz_i z_i^T L^T = L\left(\frac{1}{2n}\sum_{i=1}^{2n} z_i z_i^T\right)L^T = LI_nL^T = \Sigma$$

(c).

因為 $\Phi\Phi^T$ 為 symmetric，故可被正規對角化。設 $\Phi\Phi^T$ 的 eigenvalue 為 $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_m$。設 Φ 為 $[\Phi_1, \Phi_2 \ldots \Phi_k], \Phi_i \in R^{m*1}$，由 $\Phi^T\Phi = I_k$ 得到 $\Phi_1, \Phi_2 \ldots \Phi_k$ 為一組正交向量。設$w_1, w_2 \ldots w_{m-k} \in R^{m*1}$ 可使得 $\Phi_1 \ldots \Phi_m$, $w_1 \ldots w_m$ 為一組正交基底。

由 $\Phi^T\Phi_\mathrm{i} = e_i, (\Phi\Phi^\mathrm{T})\Phi_i = \Phi(\Phi^\mathrm{T}\Phi_i) = \Phi e_i = 1\Phi_i$得到 $1, \Phi_i$ 為一組對應的 eigenvalue 及 eigenvector。而由 $\Phi^T w_i = 0_k, (\Phi\Phi^\mathrm{T})w_i = \Phi(\Phi^\mathrm{T}\mathrm{w}_i) = \Phi 0_k = 0w_i$得到 $0, w_i$ 為一組對應的 eigenvalue 及 eigenvector。

設 $\mu_i = \begin{matrix} 1, & 1 \geq i > k \\ 0, & k+1 \geq i \geq m \end{matrix}$。By Von Neumann's Inequality，得到

$$Trace(\Phi^T\Sigma\Phi) = Trace(\Sigma\Phi\Phi^T) = Trace\big(\Sigma(\Phi\Phi^T)\big) \geq \sum_{i=1}^{m}\lambda_i\mu_{m-i+1}$$

$$\sum_{i=1}^{m}\lambda_i\mu_{m-i+1} = \sum_{i=1}^{m-k}\lambda_i\mu_{m-i+1} + \sum_{i=m-k+1}^{m}\lambda_i\mu_{m-i+1}$$

$$= 0 + \sum_{i=m-k+1}^{m}\lambda_i = \sum_{i=m-k+1}^{m}\lambda_i$$

因此得 lower bound 為

$$Trace(\Phi^T\Sigma\Phi) \geq \sum_{i=m-k+1}^{m} \lambda_i$$

因為 Σ 為 symmetric，故可被正規對角化為 $\Sigma = \mathrm{Q}\Lambda Q^T$，其中 $\mathrm{Q} = [\mathrm{\mu}_1, \mathrm{\mu}_2 \dots \mathrm{\mu}_\mathrm{m}]$，$\Lambda = \mathrm{diag}(\lambda_1, \lambda_2 \dots \lambda_\mathrm{m})$。設 $\Phi = [\mu_{m-k+1} \dots \mu_m]$ 且 $\Phi^T\Phi = I_k$。可推得

$$Trace(\Phi^T\Sigma\Phi) = Trace\big(\Phi^T(\Sigma\Phi)\big)$$

$$= Trace\left(\begin{pmatrix} \mu_{m-k+1}^T \\ \mu_{m-k+2}^T \\ \vdots \\ \mu_m^T \end{pmatrix} \cdot \Sigma(\mu_{m-k+1}\ \ \mu_{m-k+2} \dots \mu_m)\right)$$

$$= Trace\left(\begin{pmatrix} \mu_{m-k+1}^T \\ \mu_{m-k+2}^T \\ \vdots \\ \mu_m^T \end{pmatrix} \cdot (\Sigma\mu_{m-k+1}\ \ \Sigma\mu_{m-k+2} \dots \Sigma\mu_m)\right)$$

$$= Trace\left(\begin{pmatrix} \mu_{m-k+1}^T \\ \mu_{m-k+2}^T \\ \vdots \\ \mu_m^T \end{pmatrix} \cdot (\lambda_{m-k+1}\mu_{m-k+1}\ \ \lambda_{m-k+1}\mu_{m-k+2} \dots \lambda_{m-k+1}\mu_m)\right)$$

$$= Trace\begin{pmatrix} \lambda_{m-k+1}\|\mu_{m-k+1}\|^2 & & & \\ & \lambda_{m-k+2}\|\mu_{m-k+2}\|^2 & & \\ & & \ddots & \\ & & & \lambda_m\|\mu_m\|^2 \end{pmatrix}$$

$$= \lambda_{m-k+1}\|\mu_{m-k+1}\|^2 + \lambda_{m-k+2}\|\mu_{m-k+2}\|^2 + \cdots + \lambda_m\|\mu_m\|^2$$

$$= \lambda_{m-k+1} + \lambda_{m-k+2} + \cdots + \lambda_m$$

而 $\lambda_{m-k+1} + \lambda_{m-k+2} + \cdots + \lambda_m$為$Trace(\Phi^T\Sigma\Phi)$ 的最小值，故 $\Phi = [\mu_{m-k+1} \dots \mu_m]$ 即為本題解。