ICBS: Machine Learning and Finance **(Due: 9 June 2020)**

Coursework

Instructors: Arnaud de Servigny & Hachem Madmoun

Conventions:

- Mathematically, the probability that a random variable X takes the value x is denoted $p(X = x)$. In this document, we simply write $p(x)$ to denote the distribution evaluated for the particular value x.
- When an activation function a is applied to every element of a vector $Z \in \mathbb{R}^d$, we simple write $a(Z)$.
- For a matrix $X \in \mathbb{R}^{n \times d}$ with rows $X_1, \ldots, X_n$ and an activation function a, we simply write $a(X)$ to denote the matrix composed of the rows $a(X_1), \ldots, a(X_n)$

Problem A: Building a Language Model (60 points)

For the whole *problem A*, we will use the csv file **RedditNews.csv**[1]in the **data** folder.

1 Preprocessing the data

In the *RedditNews.csv* file are stored historical news headlines from Reddit WorldNews Channel, ranked by reddit users' votes, and only the top 25 headlines are considered for a single date.

You will find two colomns:

- The first column is for the "date".
- The second column is for the "News". As all the news are ranked from top to bottom, there are only 25 lines for each date.

Question 1: Load the data from the csv file, create a list of all the news.

Question 2: Preprocess the data by transforming the list of sentences into a list of sequences of integers, via a dictionary that maps the words to integers. (For each sentence, add the token "Start" at the beginning of the sentence and "End" at the end.)

[1]Source: Sun, J. (2016, August) Daily News for Stock Market Prediction, Version 1. Retrieved [26 may 2020] from https://www.kaggle.com/aaron7sun/stocknews.

Let x be the list of the sequences of integers and N be the length of x.

Let T_i be the length of each sequence x_i in x.

$$x \begin{cases} x_1 = x_1^1, x_1^2, \ldots, x_1^{T_1} \\ x_2 = x_2^1, x_2^2, \ldots\ldots, x_2^{T_2} \\ x_3 = x_3^1, x_3^2, \ldots\ldots\ldots, x_3^{T_3} \\ \quad\vdots \qquad\qquad \vdots \\ x_N = x_N^1, x_N^2, \ldots, x_N^{T_N} \end{cases}$$

2 Building a Language Model on Reddit News

A **Language Model** is a model that aims to compute the probabilty of a sequence.

Language Modeling is one of the most important tasks in natural language processing.

For instance, if we want to create a speech recognition system. A wave signal is transformed into a sentence. By computing the likelihood of two possible sentences, we can choose the most likely one.

The objective of this section is to build a **language model** on the corpus Reddit News. To create it, we need a corpus of documents $(x_i)_{\{1\leq i\leq N\}}$ for the training. Each document x_i of length T_i is a sequence $x_i^1, \ldots, x_i^{T_i}$. (T_i depends on i because the sentences are of different lengths).

We also need to make assumptions regarding the dependencies in each document. Let $x = x^1, \ldots, x^T$ be a sequence. We will assume a first order Markov assumption. Which means:

$$\forall t \in \{1, \ldots, T-1\} \quad p(x^{t+1}|x^1, \ldots, x^t) = p(x^{t+1}|x^t)$$

We define $l(x)$ the normalized log likelihood of a sentence $x = x^1, \ldots, x^T$ as follows

$$L(x) = \frac{1}{T} \log p(x^1, \ldots, x^T)$$

Question 3:

Show that, for each sequence $x = x^1, \ldots, x^T$

$$p(x^1) = 1$$

And deduce that:

$$L(x) = \frac{1}{T} \sum_{t=1}^{T-1} \log p(x^{t+1}|x^t) \tag{2.1}$$

The objective of the subsection 2.1 and 2.2 is to model $p(x^{t+1}|x^t)$. The subsection 2.1 will use a simple Markov Model for that and the subsection 2.2 will use a shallow neural network.

2.1 A Markov Model

Let V be the size of the vocabulary of our dataset.

For each couple $(i, j) \in \{1, \ldots, V\}$, we model the transition from the $i - th$ word to the $j - th$ word of the vocabulary using the element $Q[i, j]$ of a transition matrix Q.

$$p(i \to j) = Q[i, j]$$

Question 4: What is the expression of the matrix Q^* which maximizes the likelihood of a training corpus $(x_i)_{\{1 \leq i \leq N\}} = (x_i^1, \ldots, x_i^{T_i})_{\{1 \leq i \leq N\}}$

Question 5: From the equation 2.1 and Question 4, deduce the new expression of the normalized log likelihood for a sequence $x^1, \ldots, x^T$

Question 6: Estimate the matrix Q^* on the Reddit News data

Question 7: Compare the normalized log likelihoods of 5 sentences from the Reddit News corpus and 5 fake sentences generated randomly from the vocabulary (without carrying of their meaning). What can you conclude?

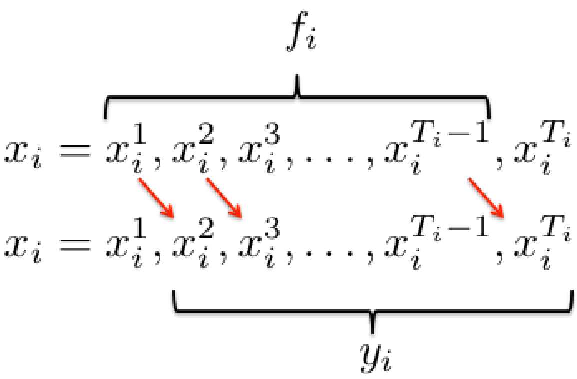
2.2 A Shallow Neural Network

In this subsection, we are going to model $p(x^{t+1}|x^t)$ using a shallow neural network.

The *Reddit News* corpus is composed of N sequences $(x_i)_{1 \leq i \leq N}$.

Let's describe the forward and the backward propagation for a specific batch of features, associated with the sequence $x_i = (x_i^1, \ldots, x_i^{T_i})$.

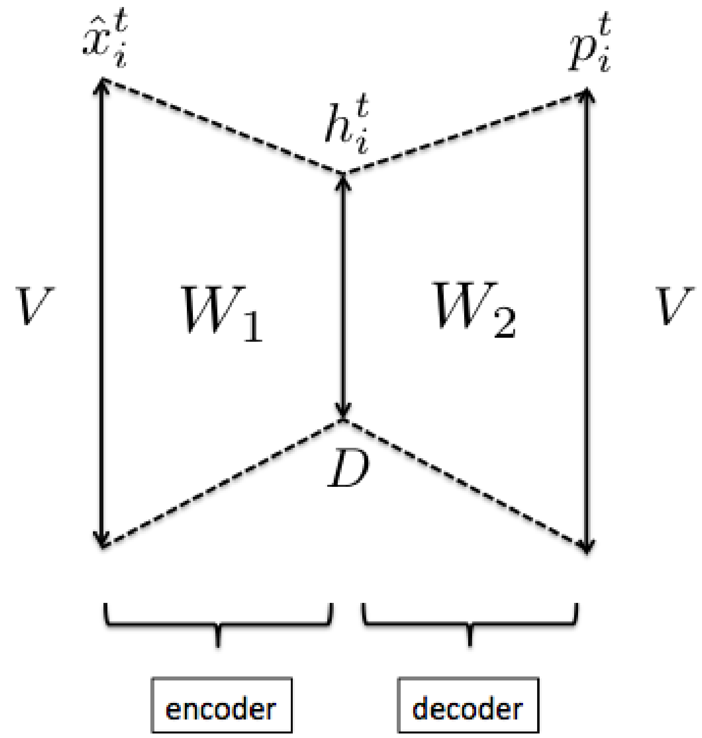
We want the model to learn to predict x_i^{t+1} given x_i^t for all $t \in \{1, \ldots, T_i - 1\}$ as reprensented in the following figure:



For each sequence x_i, we have a batch of $T_i - 1$ features, which are the elements $f_i = (f_i^1, \dots, f_i^{T_i-1}) = (x_i^1, \dots, x_i^{T_i-1})$. These $T_i - 1$ features are associated with the $T_i - 1$ targets $y_i = (y_i^1, \dots, y_i^{T_i-1}) = (x_i^2, \dots, x_i^{T_i})$. Each target is an integer in the vocabulary of size V. Hence, our problem is a multiclass classification problem with V possible classes.

Let $\hat{x}_i^t$ represents the V-dimensional one hot vector associated with the integer x_i^t.

The forward probagation for a specific feature vector $\hat{x}_i^t$ is described in the following figure:



- First, the V-dimensional one hot vector $\hat{x}_i^t$ is feeded to the neural network.
- The first transformation maps the vector $\hat{x}_i^t$ to the low D-dimensial vector h_i^t throught the W_1 matrix and the *tanh* activation function:
$$h_i^t = \tanh(W_1^T \hat{x}_i^t)$$
- The second transformation maps the hidden vector h_i^t to the discrete probability distribution p_i^t via the matrix W_2 as follows:
$$p_i^t = \text{softmax}(W_2^T h_i^t)$$

Matrix Notation:

Instead of performing the forward propagation described before on a specific vector $\hat{x}_i^t$, we can combine all the $T_i - 1$ vectors $\hat{x}_i^t$ for all $t \in \{1, \dots, T_i - 1\}$ as the rows of a feature matrix F_i of shape $(T_i - 1, V)$ as follows:

$$\forall t \in \{1, \dots, T_i - 1\} \ \forall v \in \{1, \dots, V\} \quad F_i[t, v] = \hat{x}_i^t[v]$$

The feature matrix:

We can also combine all the D-dimensional vectors h_i^t as rows of a matrix H_i. Let $H_i[1], \dots, H_i[T_i - 1]$ be the rows oh H_i.

$$F_i \left\{ \begin{array}{lcl} & & \overset{V}{\longleftrightarrow} \\ \hat{x}_i^1 & = & \hat{x}_i^1[1]\ , \quad \dots \quad , \hat{x}_i^1[V] \\ \hat{x}_i^2 & = & \hat{x}_i^2[1]\ , \quad \dots \quad , \hat{x}_i^2[V] \\ & & \\ \hat{x}_i^{T_i-1} & = & \hat{x}_i^{T_i-1}[1]\ , \quad \dots \quad , \hat{x}_i^{T_i-1}[V] \end{array} \right. \quad \Big\updownarrow \; T_i - 1$$

Then, the **hidden matrix** H_i is defined as follows:

$$\forall t \in \{1, \dots, T_i - 1\} \quad H_i[t] = h_i^t$$

Finally, we combine all the final vectors p_i^t as rows of a matrix P_i. Let $P_i[1], \dots, P_i[T_i - 1]$ be the rows of P_i.

Then, the **prediction matrix** P_i is defined as follows;

$$\forall t \in \{1, \dots, T_i - 1\} \quad P_i[t] = p_i^t$$

Question 8: What are the shapes of the matrics H_i and P_i ?

Question 9: Show that:

$$H_i = \tanh(FW_1)$$
$$P_i = \text{softmax}(HW_2)$$

We also need to one hot encode the targets $y_i = (y_i^1, \dots, y_i^{T_i-1}) = (x_i^2, \dots, x_i^{T_i})$.

Each target y_i^t for $t \in \{1, \dots, T_i - 1\}$ is one hot encoded into a V-dimensional vector $\hat{y}_i^k$.

We combine all the final vectors $\hat{y}_i^t$ as rows of a matrix Y_i. Let $Y_i[1], \dots, Y_i[T_i - 1]$ be the rows of Y_i.

Then, the **target matrix** Y_i is defined as follows;

$$\forall t \in \{1, \dots, T_i - 1\} \quad Y_i[t] = \hat{y}_i^t$$

The **Target Matrix**:

$$\begin{array}{cc} & \overset{V}{\longleftrightarrow} \\ Y_i & \left[\begin{array}{lcl} \hat{y}_i^1 & = & \hat{y}_i^1[1]\,, \quad \cdots \quad , \hat{y}_i^1[V] \\ \hat{y}_i^2 & = & \hat{y}_i^2[1]\,, \quad \cdots \quad , \hat{y}_i^2[V] \\ & & \\ \hat{y}_i^{T_i-1} & = & \hat{y}_i^{T_i-1}[1]\,, \quad \cdots \quad , \hat{y}_i^{T_i-1}[V] \end{array}\right] \end{array} \; \updownarrow T_i - 1$$

Question 10: Explain why the loss function J_i for the batch associated with the sequence x_i is :

$$J_i = -\frac{1}{T_i - 1} \sum_{t=1}^{T_i-1} \sum_{v=1}^{V} \hat{y}_i^t[v] \log p_i^t[v]$$

For the learning process, we need to calculate the gradient of the loss function J_i with respect to the matrices W_1 and W_2.

Let's consider the two matrices:

$$G_i^1 = H_i^T (P_i - Y_i)$$
$$G_i^2 = F_i^T (P_i - Y_i) W_2^T$$

Question 11: Based on their shapes, determine which of these two matrices G_i^1 and G_i^2 represents $\nabla_{W_1} J_i$ and which one represents $\nabla_{W_2} J_i$

Question 12: Train the neural network for one epoch using the gradient descent as follows:

- **Shuffle the sequences.**
- **For each sequence x_i in the training corpus:**
  - **Create the feature matrix F_i ,the hidden matrix H_i, the prediction matrix P_i and the target matrix Y_i.**
  - **Calculate the loss function associated with this batch and store it in a list**
  - **Perform one step of Gradient Descent to update the weights matrices W_1 and W_2.**

Question 13: Using an EWMA, plot a smooth version of the list of losses associated with each update of the Gradient Descent.

Question 14: How can we create an embedding matrix from the trained neural network?

Problem B: Building a Sentiment Analysis Model (40 points)

The data we will use for the *problem B* is the **TweetSentiment.csv**[2] file in the **data** folder.

In the *TweetSentiment.csv* file are stored several tweets. Each tweet is associated with a target representing its sentiment. There are three possible sentiments:

- $+1$ is for a positive sentiment.
- -1 is for a negative sentiment.
- 0 is for a neutral sentiment.

Question 15: Load the data from the csv file, split it into the train and the test datasets. Learn a sentiment analysis model of your choice on the training dataset and evaluate it on the test dataset.

[2] Source: https://www.kaggle.com/vivekrathi055/sentiment-analysis-on-financial-tweets