Web Content Mining – Classification

Exercise sheet

Keti Korini & Alexander Brinkmann

In this exercise, we will use a list of user reviews to do sentiment analysis.

Dataset

We start with the infos below to show the book reviews of "The Da Vinci Code" from 4 users and their corresponding sentiment labels: Negative, Positive.

Review ID	Reviews	Label
R1	The plot is predictable without excitement. I won't recommend it.	Negative
R2	The plot is novel and the story is interesting.	Positive
R3	The plot is fast-paced and the story is filled with excitement. I would recommend it.	Positive
R4	The story is unconvincing and the plot is predictable.	Negative

We want to use the above infos and build a classifier to classify the new user reviews which we haven't read yet.

Task 1: Using TF-IDF Vectors and Logistic Regression for Sentiment Classification

Before we represent each user view as the tf-idf vector, we first extract the distinct terms from the metadata information. Please consider that for matter of simplicity we do lowercase, remove stopwords, and remove the following punctuations: period (.) and comma (,).

(a) We first compute the tf-idf weight of a term t for a document d (which corresponds to a user review in our example) as the product of its tf weight and its idf:

$$w_{t,d} = (1 + \log_{10} tf_{t,d}) \times \log_{10} \frac{N}{df_t}$$

where tf is the frequency of t in d, N is the total number of user reviews and df is the total number of user reviews in which the term t appeared at least once.

Fill in the table below with the tf-idf weights of each term according to our corpus.

	R1	R2	R3	R4
plot				
predictable				
without				
excitement				
won't				
recommend				
novel				
story				
interesting				
fast-paced				
filled				
unconvincing				

(b) Use Logistic Regression as a binary classifier to classify the sentiments.

Logistic regression is given with the following (parametrized) function:

$$h(\mathbf{x}|\mathbf{w}) = \sigma(\mathbf{w}^T\mathbf{x}) = \frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x})}.$$

where x is the input text representation as a vector, parametrized by w, and apply with sigmoid function σ to map the values between 0 and 1. The binary classification is then obtained by thresholding on the value of 0.5, i.e., if $h(\mathbf{x}|\mathbf{w}) > 0.5$ then $\mathbf{x}$ belongs to class 1 (i.e. *yes*), otherwise to class 0 (or *no*). In our case, class 1 is identical to *positive* sentiment while class 0 is *negative*.

You are given the vocabulary as above list {*plot*, *predictable*, *without*, *excitement*, *won't*, *recomment*, *novel*, *story*, *interesting*, $fast-paced$, $filled$, *unconvincing*}. Assuming that your current value of the parameter vector w is given as

$$\mathbf{w} = [0.5, -0.2, -0.3, 1.5, -0.3, 1.1, 0.9, 0.8, 1.3, 1.7, 0.1, -0.4],$$

Make the binary classification prediction for four user reviews {*R1, R2, R3, R4*} (use the tf-idf features to represent each of the review).

Review ID	**Sigmoid Score (Pred Score)**	**Pred Label**
R1		
R2		
R3		
R4		

(c) The predictions made by a machine learning model are compared with true labels (classes) of instances in order to measure the prediction error. The prediction errors are determined by the model's loss function. The loss function of the logistic regression is the so-called (binary) cross-entropy loss, which is, for a single input instance, given as follows:

$$ce(\mathbf{x}, \mathbf{w}, y) = -\left(y \ln h(\mathbf{x}|\mathbf{w}) + (1-y)\ln(1-h(\mathbf{x}|\mathbf{w}))\right).$$

Compute the **average cross-entropy loss** for our user reviews {*R1, R2, R3, R4*} using the same parameter vector $\mathbf{w}$ as in (b), assuming that the sentiment labels for {*R1, R2, R3, R4*} is {*0, 1, 1, 0*} (i.e. *Negative: 0, Positive: 1*)

(d) Suppose we have a new user review: *"The story is filled with excitement, I will recommend it."*. Please first calculate the tf-idf representation based on (a) and use the same weight w in (b) to calculate the sigmoid score and the predicted sentiment label.

Task 2: Using Dense Vector Representation and Logistic Regression for Sentiment Classification

Similarly as Task 1, instead we use dense vector representation as following to represent each of the user review:

$$d_1 = [0.58, -0.48, 0.65, -1.35, 1.85]$$
$$d_2 = [-0.61, 1.97, -1.06, 0.34, -0.37]$$
$$d_3 = [-1.46, 0.18, 0.07, 1.07, 1.74]$$
$$d_4 = [-1.64, -1.22, 1.98, -1.06, -1.04]$$

(a) Use Logistic Regression as a binary classifier to classify the sentiments. Assuming that your current value of the parameter vector w is given as

$$\mathbf{w} = [-0.3, 1.8, 0.9, 1.5, -0.1],$$

Make the binary classification prediction for four user reviews {*R1, R2, R3, R4*} (use the dense vector representation to represent each of the review).

Review ID	**Sigmoid Score (Pred Score)**	**Pred Label**
R1		
R2		
R3		
R4		

(b) Compute the **average cross-entropy loss** for our user reviews {*R1, R2, R3, R4*} using the same parameter vector **w** defined in (a), assuming that the sentiment labels for {*R1, R2, R3, R4*} is {*0, 1, 1, 0*} (i.e. *Negative: 0, Positive: 1*)

(c) Suppose we have a new user review: *"The plot is unconvincing, I won't recommend it."*, with the dense vector representation as:

$$d_{new} = [0.52, 0.94, 0.75, -1.88, 1.61]$$

Please use the same weight w in (a) to calculate the sigmoid score and the predicted sentiment label.

Task 3: Comparison

What are the differences between using tf-idf representation and dense vector representation?