Recitation 5
Logistic Regression

10-601: Introduction to Machine Learning

10/01/2021

This recitation consists of 3 parts: In part 1, we will go over how to **represent text data using two different feature extraction methods**. Part 2 will go over the **negative log likelihood** and **gradient derivations** for **binary logistic regression**, as well as a small toy example. Part 3 will focus on **multinomial logistic regression**. The materials were designed to help you with Homework 4.

1 Feature Representation for Sentiment Classification

In many machine learning problems, we will want to find appropriate representations for the inputs of the algorithm we are developing. In Homework 4, we will work on using logistic regression for a sentiment classification task, where our algorithm takes a paragraph of movie review as the input and outputs a binary value denoting whether the review is positive or not. To build an appropriate representation for the input (aka. the review text), we consider two different representations – (1) a bag-of-word representation and (2) a representation built on top of Word2vec word embeddings.

In this section, consider a scenario where we are interested in representing the following text:

```
a hot dog is not a sandwich because it is not square
```
(1)

We consider the following dictionary (denoted below as **Vocab**) as the set of vocabulary that we will consider. Note that the vocabulary dictionary might not contain all words in the text shown above.

```
dictionary = {
    "the": 0,
    "square": 1,
    "hot": 2,
    "is": 3,
    "not": 4,
    "a": 5,
    "happy": 6,
    "sandwich": 7
}
```

1. **Bag-of-words Representation**

   A bag-of-words representation $\boldsymbol{\phi}_1(\mathbf{x})$ of text $\mathbf{x}$ is defined by $\boldsymbol{\phi}_1(\mathbf{x}) = \mathbf{1}_{\mathbf{occur}}(\mathbf{x}^{(\mathbf{i})}, \mathbf{Vocab})$, indicating which words in vocabulary **Vocab** of the dictionary occur at least once in the movie review example $\mathbf{x}^{(i)}$. Let $\mathbf{x}$ be the sample text defined above. Write the bag-of-words representation of $\mathbf{x}$. `[0, 1, 1, 1, 1, 1, 0, 1]`

2. **Word Embedding Based Representation**

   (a) Word embeddings are reduced dimension vector representations (features) of words. Given a single word in the dictionary, word embeddings can convert it to a vector of fixed dimension. In Homework 4, we will provide a dictionary file specifying pre-computed mappings between every word in **Vocab** and their corresponding word embeddings. To facilitate better understanding towards word embeddings, we produce a plot showing the spatial relationship between several sample words from the vocabulary used in Homework 4:

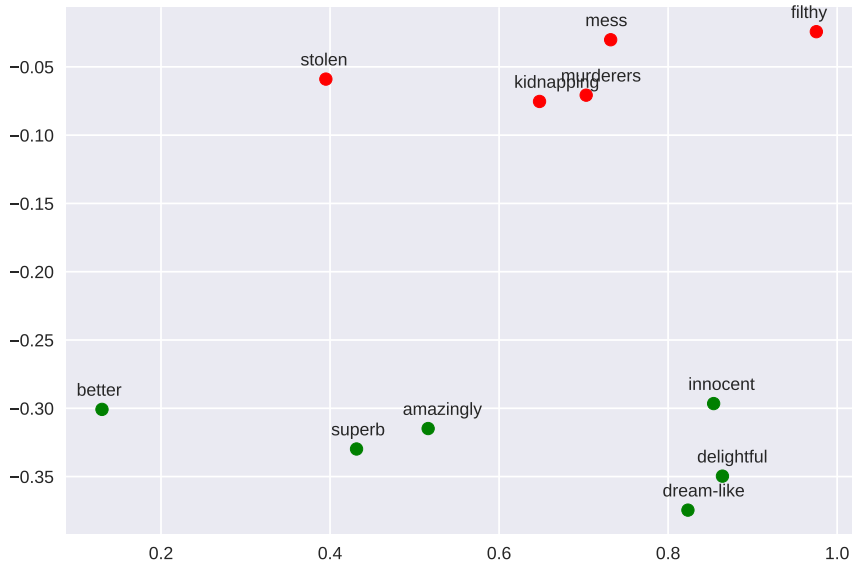

Figure 1: Visualization of word embeddings. We select a few positive words (shown in green) and a few negative words (shown in red). To make the plot, we map the high-dimensional word representations of these words to 2D space using PCA and then visualize them in the scatter plot above.

Please comment on your observations and findings based on this plot. Closer-related

words are located closer in the representation space, while farther-related words are located farther from each other.

(b) One approach to build a representation for text is to average out the vector representation of every word in the text that are in the dictionary. For example, given text "`a hot dog flies like a sandwich`", we can find the representation for this text by taking the average of the vector representation of the words "`a`", "`hot`", "`a`", and "`sandwich`".

Now suppose we have the following word embedding dictionary for building vector representation of text (this is a toy example used for illustrative purposes; actual word embeddinngs will have higher dimensions than this example):

```
dictionary = {
    "the": [0.2, 0.3],
    "square": [0.8, 0.9],
    "hot": [0.1, -0.2],
    "is": [0.1, 0.1],
    "not": [-0.2, -0.3],
    "a": [0.0, 0.0],
    "happy": [0.4, 0.4],
    "sandwich": [0.2, -0.3]
}
```

Write the word embedding based representation of the sample text define above.

$$\begin{aligned}\phi_2(\mathbf{x}) &= \frac{1}{9}\big(f(\text{square}) + f(\text{hot}) + 2\cdot f(\text{is}) + 2\cdot f(\text{not}) + 2\cdot f(\text{a}) + f(\text{sandwich})\big)\\ &= \begin{bmatrix}0.1 & 0.0\end{bmatrix}^T.\end{aligned}$$

2 Binary Logistic Regression

1. For binary logistic regression, we have the following dataset:

$$\mathcal{D} = \left\{ \left(\mathbf{x}^{(1)}, y^{(1)} \right), \ldots, \left(\mathbf{x}^{(N)}, y^{(N)} \right) \right\} \text{ where } \mathbf{x}^{(i)} \in \mathbb{R}^M, y^{(i)} \in \{0, 1\}$$

A couple of reminders from lecture

1. $$\sigma(\boldsymbol{\theta}^T \mathbf{x}^{(i)}) = \frac{1}{1 + \exp(-\boldsymbol{\theta}^T \mathbf{x}^{(i)})} = \frac{\exp(\boldsymbol{\theta}^T \mathbf{x}^{(i)})}{1 + \exp(\boldsymbol{\theta}^T \mathbf{x}^{(i)})}$$

2. $$p\left(y^{(i)} \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}\right) = \begin{cases} \sigma(\boldsymbol{\theta}^T \mathbf{x}^{(i)}) & y^{(i)} = 1 \\ 1 - \sigma(\boldsymbol{\theta}^T \mathbf{x}^{(i)}) & y^{(i)} = 0 \end{cases}$$
$$= \sigma(\boldsymbol{\theta}^T \mathbf{x}^{(i)})^{y^{(i)}} (1 - \sigma(\boldsymbol{\theta}^T \mathbf{x}^{(i)}))^{(1 - y^{(i)})}$$

3. $$\boldsymbol{\phi}^{(i)} = \sigma(\boldsymbol{\theta}^T \mathbf{x}^{(i)})$$

4. $$\frac{\partial \sigma(z)}{\partial z} = \sigma(z)(1 - \sigma(z))$$

5. if $z = f(\boldsymbol{\theta})$ then
$$\frac{\partial \sigma(f(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}_j} = \sigma(f(\boldsymbol{\theta}))(1 - \sigma(f(\boldsymbol{\theta}))) \frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}$$
In binary logistic regression, this is
$$\frac{\partial \boldsymbol{\phi}^{(i)}}{\partial \boldsymbol{\theta}_j} = \boldsymbol{\phi}^{(i)} * (1 - \boldsymbol{\phi}^{(i)}) * \frac{\partial \boldsymbol{\theta}^T \mathbf{x}^{(i)}}{\partial \theta_j}$$

6. remember that
$$\frac{\partial \log(f(z))}{\partial z} = \frac{1}{f(z)} \frac{\partial f(z)}{\partial z}$$

2. (a) Write down our objective function, $J(\boldsymbol{\theta})$, which is $\frac{1}{N}$ times the negative conditional log-likelihood of data, in terms of N and $p\left(y^{(i)} \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}\right)$ where $\boldsymbol{\theta} \in \mathbb{R}^M$. As usual, assume $y^{(i)}$ are independent and identically distributed.

$$J(\boldsymbol{\theta}) = -\frac{1}{N}\log(\prod_{i=1}^{N} p\left(y^{(i)} \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}\right))$$

$$J(\boldsymbol{\theta}) = -\frac{1}{N}\sum_{i=1}^{N} \log(p\left(y^{(i)} \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}\right))$$

(b) Write $J(\boldsymbol{\theta})$ in terms of $\sigma(\boldsymbol{\theta}^T\mathbf{x}^{(i)})$. simplify as much as possible. Then write in terms of $\boldsymbol{\phi}^{(i)}$

$$\begin{aligned} J(\boldsymbol{\theta}) &= -\frac{1}{N}\sum_{i=1}^{N} \log\left(\sigma(\boldsymbol{\theta}^T\mathbf{x}^{(i)})^{y^{(i)}}(1-\sigma(\boldsymbol{\theta}^T\mathbf{x}^{(i)}))^{(1-y^{(i)})}\right) \\ &= -\frac{1}{N}\sum_{i=1}^{N} (y^{(i)}\log\left(\sigma(\boldsymbol{\theta}^T\mathbf{x}^{(i)})\right) + (1-y^{(i)})\log\left(1-\sigma(\boldsymbol{\theta}^T\mathbf{x}^{(i)})\right)) \\ &= -\frac{1}{N}\sum_{i=1}^{N} (y^{(i)}\log\left(\boldsymbol{\phi}^{(i)}\right) + (1-y^{(i)})\log\left(1-\boldsymbol{\phi}^{(i)}\right)) \end{aligned}$$

(c) In stochastic gradient descent, we use only a single $\mathbf{x}^{(i)}$. Given $\boldsymbol{\phi}^{(i)} = \sigma(\boldsymbol{\theta}^T\mathbf{x}^{(i)})$ and

$$J^{(i)}(\boldsymbol{\theta}) = -y^{(i)}\log(\boldsymbol{\phi}^{(i)}) - (1-y^{(i)})\log(1-\boldsymbol{\phi}^{(i)})$$

Show that the partial derivative of $J^{(i)}(\boldsymbol{\theta})$ with respect to the jth parameter θ_j is as follows:

$$\frac{\partial J^{(i)}(\boldsymbol{\theta})}{\partial \theta_j} = (\sigma(\boldsymbol{\theta}^T\mathbf{x}^{(i)}) - y^i)x_j^{(i)}$$

Remember,

$$\frac{\partial \boldsymbol{\phi}^{(i)}}{\partial \boldsymbol{\theta}_j} = \boldsymbol{\phi}^{(i)} * (1-\boldsymbol{\phi}^{(i)}) * \frac{\partial \boldsymbol{\theta}^T\mathbf{x}^{(i)}}{\partial \theta_j}$$

note

$$\frac{\partial \boldsymbol{\theta}^T \mathbf{x}^{(i)}}{\partial \theta_j} = \mathbf{x}_j^{(i)}$$

$$\begin{aligned}
\frac{\partial J^{(i)}(\boldsymbol{\theta})}{\partial \theta_j} &= -\frac{y^{(i)}}{\boldsymbol{\phi}^{(i)}}\frac{\partial \boldsymbol{\phi}^{(i)}}{\partial \boldsymbol{\theta}_j} - \frac{(1-y^{(i)})}{1-\boldsymbol{\phi}^{(i)}}\frac{\partial (1-\boldsymbol{\phi}^{(i)})}{\partial \boldsymbol{\theta}_j} \\
&= -\frac{y^{(i)}}{\boldsymbol{\phi}^{(i)}}\frac{\partial \boldsymbol{\phi}^{(i)}}{\partial \boldsymbol{\theta}_j} + \frac{(1-y^{(i)})}{1-\boldsymbol{\phi}^{(i)}}\frac{\partial \boldsymbol{\phi}^{(i)}}{\partial \boldsymbol{\theta}_j} \\
&= -\frac{y^{(i)}}{\boldsymbol{\phi}^{(i)}}\boldsymbol{\phi}^{(i)} * (1-\boldsymbol{\phi}^{(i)}) * \frac{\partial \boldsymbol{\theta}^T \mathbf{x}^{(i)}}{\partial \theta_j} + \frac{(1-y^{(i)})}{1-\boldsymbol{\phi}^{(i)}}\boldsymbol{\phi}^{(i)} * (1-\boldsymbol{\phi}^{(i)}) * \frac{\partial \boldsymbol{\theta}^T \mathbf{x}^{(i)}}{\partial \theta_j} \\
&= (-y^{(i)}(1-\boldsymbol{\phi}^{(i)}) + (1-y^{(i)})\boldsymbol{\phi}^{(i)}))\mathbf{x}_j^{(i)} \\
&= (-y^{(i)} + y^{(i)}\boldsymbol{\phi}^{(i)} + \boldsymbol{\phi}^{(i)} - y^{(i)}\boldsymbol{\phi}^{(i)})\mathbf{x}_j^{(i)} \\
&= (\boldsymbol{\phi}^{(i)} - y^{(i)})\mathbf{x}_j^{(i)} \\
&= (\sigma(\boldsymbol{\theta}^T \mathbf{x}^{(i)}) - y^{i})\mathbf{x}_j^{(i)}
\end{aligned}$$

3. Let's go through a toy problem.

Y	X_1	X_2	X_3
1	1	2	1
1	1	1	-1
0	1	-2	1

(a) What is $J(\boldsymbol{\theta})$ of above data given initial $\boldsymbol{\theta} = \begin{bmatrix} -2 \\ 2 \\ 1 \end{bmatrix}$?

$J(\boldsymbol{\theta}) = -\frac{1}{3}[\log(\sigma(3)) + \log(\sigma(-1)) + \log(1-\sigma(-5))] \approx 0.46$

(b) Calculate $\frac{\partial J^{(1)}(\boldsymbol{\theta})}{\partial \theta_1}$, $\frac{\partial J^{(1)}(\boldsymbol{\theta})}{\partial \theta_2}$ and $\frac{\partial J^{(1)}(\boldsymbol{\theta})}{\partial \theta_3}$ for first training example. Note that $\sigma(3) \approx 0.95$.

$$\frac{\partial J^{(1)}(\boldsymbol{\theta})}{\partial \theta_1} = (\sigma(3) - 1)1 = -0.05$$
$$\frac{\partial J^{(1)}(\boldsymbol{\theta})}{\partial \theta_2} = (\sigma(3) - 1)2 = -0.10$$
$$\frac{\partial J^{(1)}(\boldsymbol{\theta})}{\partial \theta_3} = (\sigma(3) - 1)1 = -0.05$$

(c) Calculate $\frac{\partial J^{(2)}(\boldsymbol{\theta})}{\partial \theta_1}$, $\frac{\partial J^{(2)}(\boldsymbol{\theta})}{\partial \theta_2}$ and $\frac{\partial J^{(2)}(\boldsymbol{\theta})}{\partial \theta_3}$ for second training example. Note that $\sigma(-1) \approx 0.25$.

$$\frac{\partial J^{(2)}(\boldsymbol{\theta})}{\partial \theta_1} = (\sigma(-1) - 1)1 = -0.75$$
$$\frac{\partial J^{(2)}(\boldsymbol{\theta})}{\partial \theta_2} = (\sigma(-1) - 1)1 = -0.75$$
$$\frac{\partial J^{(2)}(\boldsymbol{\theta})}{\partial \theta_3} = (\sigma(-1) - 1) - 1 = 0.75$$

(d) Assuming we are doing stochastic gradient descent with a learning rate of 1.0, what are the updated parameters $\boldsymbol{\theta}$ if we update $\boldsymbol{\theta}$ using the second training example?

$$\begin{bmatrix} -2 \\ 2 \\ 1 \end{bmatrix} - 1 \begin{bmatrix} -0.75 \\ -0.75 \\ 0.75 \end{bmatrix} = \begin{bmatrix} -1.25 \\ 2.75 \\ 0.25 \end{bmatrix}$$

(e) What is the new $J(\boldsymbol{\theta})$ after doing the above update? Should it decrease or increase?
$J(\boldsymbol{\theta}) = 0.09$

It should decrease for logistic classifier to learn.

(f) Given a test example where $(X_1 = 1, X_2 = 3, X_3 = 4)$, what will the classifier output following this update?
$\sigma(\boldsymbol{\theta}^T X) > 0.5 \implies Y = 1$

3 Multinomial Logistic Regression (Optional Learning)

1. Definition

   Multinomial logistic regression, also known as softmax regression or multiclass logistic regression, is a generalization of binary logistic regression.

   $\mathcal{D} = \left\{\left(\mathbf{x}^{(1)}, y^{(1)}\right), \ldots, \left(\mathbf{x}^{(N)}, y^{(N)}\right)\right\}$ where $\mathbf{x}^{(i)} \in \mathbb{R}^M, y^{(i)} \in \{1, \ldots, K\}$ for $i = 1, \ldots, N$

   Here N is the number of training examples, M is the number of features, and K is the number of possible classes, which is usually greater than two to be interesting.

$$p\left(Y^{(i)} = y^{(i)} \mid \mathbf{x}^{(i)}, \boldsymbol{\Theta}\right) = \frac{\exp\left(\boldsymbol{\Theta}_{y^{(i)}} \mathbf{x}^{(i)}\right)}{\sum_{j=1}^{K} \exp\left(\boldsymbol{\Theta}_j \mathbf{x}^{(i)}\right)} = \text{softmax}(\boldsymbol{\Theta}\mathbf{x}^{(i)})_{y^{(i)}} \tag{2}$$

   where $\boldsymbol{\Theta}$ is the parameter matrix of size $K \times (M+1)$, and $\boldsymbol{\Theta}_{y^{(i)}}$ denotes the $y^{(i)}$th **row** of $\boldsymbol{\Theta}$, which is the parameter vector for the $y^{(i)}$th class.

2. Suppose $K = 4$ and $N = 10$, $M = 3$. What could $\boldsymbol{\Theta}$ look like?

   $\boldsymbol{\Theta}$ will have K rows because there are K distinct labels. $\boldsymbol{\Theta}$ will have M+1 columns because there are M features plus a bias term. So any K by (M+1) matrix is a possible candidate for $\boldsymbol{\Theta}$.

$$\begin{bmatrix} 0.5 & -2 & 5 & 7 \\ 0 & 0.22 & 6 & 1 \\ 9 & 2 & 0.1 & 6 \\ 7 & -0.5 & 0 & 1 \end{bmatrix}$$

3. A *one-hot encoding* is a vector representation of a one dimensional integer defined as such: a vector $\mathbf{c}$ of length K is a *one-hot encoding* of integer $n \iff |\mathbf{c}| = K$ and for all $j \neq n$, $\mathbf{c}_j = 0$ and $\mathbf{c}_n = 1$. Give some examples of one-hot encodings where K = 5.

   Let $n = 1, \implies \mathbf{c} = [1, 0, 0, 0, 0]^T$
   Let $n = 3, \implies \mathbf{c} = [0, 0, 1, 0, 0]^T$
   Let $n = 4, \implies \mathbf{c} = [0, 0, 0, 1, 0]^T$

4. In multinomial logistic regression, we form the matrix $\mathbf{T}$ where the ith row of $\mathbf{T}$ is the one-hot encoding of label $y^{(i)}$. Draw $\mathbf{T}$ if $\mathbf{y} = [1, 3, 1, 4, 4]^T$ and K = 4.

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$