CS 6501 Natural Language Processing

Text Classification

Yangfeng Ji

September 3, 2018

Department of Computer Science
University of Virginia

UNIVERSITY of VIRGINIA | ENGINEERING

Case I: Like a business?

[Pang et al., 2002]

Potential Group Project

Yelp Dataset Challenge

Discover what insights lie hidden in our data.

What is the dataset challenge?

The challenge is a chance for students to conduct research or analysis on our data and share their discoveries with us. Whether you're trying to figure out how food trends start or identify the impact of different connections from the local graph, you'll have a chance to win cash prizes for your work! See some of the past winners and hundreds of academic papers written using the dataset.

Natural Language Processing & Sentiment Analysis

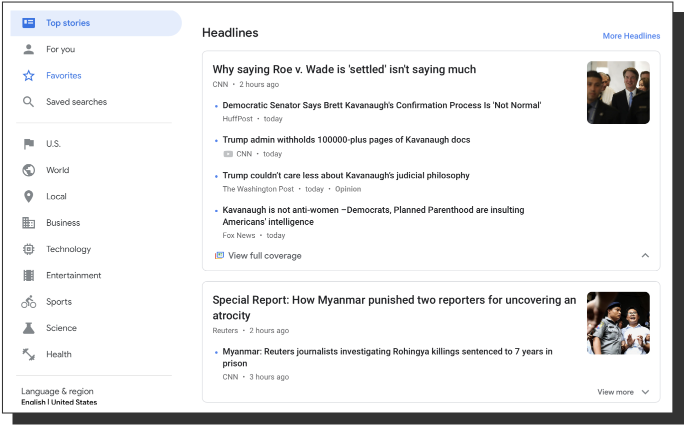
What's in a review? Is it positive or negative? Our reviews contain a lot of metadata that can be mined and used to infer meaning, business attributes, and sentiment.

Case II: Topic Classification

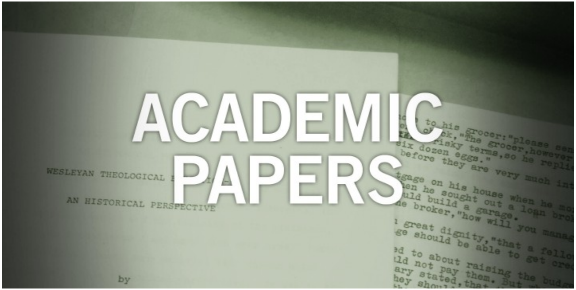
Example topics

- Business
- Arts
- Technology
- Sports
- . . .

Google News Category



Case III: Scientific Literature Analysis



- Claim
- Arguments
- Good writing?

Open Research Corpus

`https://labs.semanticscholar.org/corpus/`

Data:

- $\sim$ 39M papers
- $\sim$ 12M authors

[Ammar et al., 2018]

Potential Group Projects

`https://labs.semanticscholar.org/corpus/`

- Citation recommendation based on paper abstract
- Relation analysis on different machine learning methods
- Influence analysis of machine/deep learning on biomedical domain

Overview

1. Problem Definition
2. Case Study: Sentiment Analysis
3. Bag-of-Words Representation
4. Perceptron Algorithm
5. Classification Evaluation

Classification

- Input: a text $\boldsymbol{x}$
- Output: $y \in \mathcal{Y}$, where $\mathcal{Y}$ is the predefined category set

$$\text{Text } \boldsymbol{x} \longrightarrow \boxed{\text{Classifier}} \longrightarrow \text{Category } y$$

Mathematical Formulation

- Input: a numeric representation $\boldsymbol{x}$ of text d
- Output: scores $\Psi(\boldsymbol{x}, y; \boldsymbol{\theta}) \in \mathbb{R}, \forall y \in \mathcal{Y}$

Mathematical Formulation

- Input: a numeric representation $\boldsymbol{x}$ of text d
- Output: scores $\Psi(\boldsymbol{x}, y; \boldsymbol{\theta}) \in \mathbb{R}, \forall y \in \mathcal{Y}$

Classification

$$\hat{y} = \arg\max_{y' \in \mathcal{Y}} \Psi(\boldsymbol{x}, y'; \boldsymbol{\theta}) \tag{1}$$

Key Questions

$$\hat{y} = \arg\max_{y' \in \mathcal{Y}} \Psi(\boldsymbol{x}, y'; \boldsymbol{\theta}) \tag{2}$$

1. How to represent a text as $\boldsymbol{x}$?
2. How to formulate score function $\Psi(\boldsymbol{x}, y; \boldsymbol{\theta})$

Overview

1. Problem Definition

2. Case Study: Sentiment Analysis

3. Bag-of-Words Representation

4. Perceptron Algorithm

5. Classification Evaluation

Sentiment Analysis

Task: predicting user sentiment polarity (POSITIVE or NEGATIVE) based on a review

Twitter Sentiment Analysis

Donald J. Trump
@realDonaldTrump

Follow

Happy Labor Day! Our country is doing better than ever before with unemployment setting record lows. The U.S. has tremendous upside potential as we go about fixing some of the worst Trade Deals ever made by any country in the world. Big progress being made!

4:28 AM - 3 Sep 2018

A Simple Predictor

Example I: POSITIVE

Super quick and really friendly staff. I like starting off my mornings at this store!!

- SentiWordnet: a publicly available word sentiment polarity dictionary.

Another Example

Example II: POSITIVE

Din Tai Fung, every time I go eat at anyone of the locations around the King County area,
I keep being reminded on why I have to keep coming back to this restaurant.
…

- No signal word

Counting Words Can Be Risky

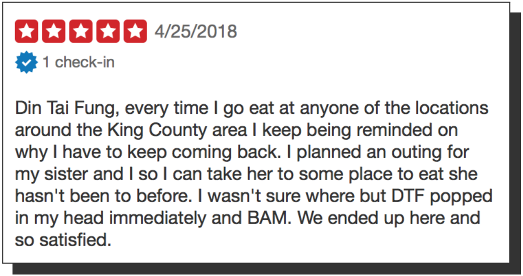
Donald J. Trump
@realDonaldTrump

Follow

Happy Labor Day! Our country is doing better than ever before with unemployment setting record lows. The U.S. has tremendous upside potential as we go about fixing some of the worst Trade Deals ever made by any country in the world. Big progress being made!

4:28 AM - 3 Sep 2018

Data Driven Approach



- Discover the relationship between words and sentiment polarity from data
- Need a collection of texts and their category labels $\{(x^{(i)}, y^{(i)})\}$

Basic Idea of Statistical Machine Learning

Given $\{(\boldsymbol{x}^{(i)}, y^{(i)})\}$

- Principle: Discover the statistical relationship between patterns and categories from training data.
- Goal: To make better decisions for unseen data points in a *test* set (generalization power)

Standard Setup of Statistical Machine Learning

- Training set $\mathcal{T} = \{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$

- Test set $\mathcal{U} = \{(\boldsymbol{x}^{(l)}, y^{(l)})\}_{l=1}^{L}$

Standard Setup of Statistical Machine Learning

- Training set $\mathcal{T} = \{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$
- Development set $\mathcal{D} = \{(\boldsymbol{x}^{(j)}, y^{(j)})\}_{j=1}^{M}$
- Test set $\mathcal{U} = \{(\boldsymbol{x}^{(l)}, y^{(l)})\}_{l=1}^{L}$

Simple Framework

$$\hat{y} = \arg\max_{y' \in \mathcal{Y}} \Psi(\boldsymbol{x}, y'; \boldsymbol{\theta}) \tag{3}$$

Score function

$$\Psi(\boldsymbol{x}, y; \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, y) \tag{4}$$

- $\boldsymbol{f}(\boldsymbol{x}, y)$: feature function
- $\boldsymbol{\theta}$: classification weights

Questions

Score function

$$\Psi(\boldsymbol{x}, y; \boldsymbol{\theta}) = \boldsymbol{\theta}^{\top} \boldsymbol{f}(\boldsymbol{x}, y) \tag{5}$$

- How to design a feature function $\boldsymbol{f}(\boldsymbol{x}, y)$?
  - Bag-of-words representation
- How to learn $\boldsymbol{\theta}$?
  - Perceptron algorithm

Overview

1. Problem Definition
2. Case Study: Sentiment Analysis
3. Bag-of-Words Representation
4. Perceptron Algorithm
5. Classification Evaluation

Bag-of-Words Representation

From texts to bag of words:

Super quick and really friendly staff. I like starting off my mornings at this store!! $\Rightarrow$ SUPER,QUICK,AND, REALLY, FRIENDLY, $\cdots$, STORE

NLTK function

```
nltk.tokenize.wordpunct_tokenize
```

Vocab

Given the texts from training set, build a vocab first

$$\left\{ \begin{array}{c} \text{SUPER} \\ \cdots \\ \text{QUICK} \\ \text{FOOD} \\ \text{FRIENDLY} \\ \text{EAT} \\ \cdots \\ \text{STAFF} \end{array} \right\} \tag{6}$$

Feature Function

Example I: POSITIVE

Super quick and really friendly staff. I like starting off my mornings at this store!!

$$\left\{\begin{array}{c} \langle \textsc{super},\text{-}1 \rangle \\ \langle \textsc{super},1 \rangle \\ \cdots \\ \langle \textsc{friendly},\text{-}1 \rangle \\ \langle \textsc{friendly},1 \rangle \\ \langle \textsc{eat},\text{-}1 \rangle \\ \langle \textsc{eat},1 \rangle \\ \cdots \\ \langle \textsc{delicious},\text{-}1 \rangle \\ \langle \textsc{delicious},1 \rangle \end{array}\right\}$$

Feature Function

Example I: POSITIVE

Super quick and really friendly staff. I like starting off my mornings at this store!!

$$\left\{\begin{array}{c} \langle \text{SUPER},\text{-1} \rangle \\ \langle \text{SUPER},1 \rangle \\ \cdots \\ \langle \text{FRIENDLY},\text{-1} \rangle \\ \langle \text{FRIENDLY},1 \rangle \\ \langle \text{EAT},\text{-1} \rangle \\ \langle \text{EAT},1 \rangle \\ \cdots \\ \langle \text{DELICIOUS},\text{-1} \rangle \\ \langle \text{DELICIOUS},1 \rangle \end{array}\right\} \qquad \begin{bmatrix} 0 \\ 1 \\ \cdots \\ 0 \\ 1 \\ 0 \\ 0 \\ \cdots \\ 0 \\ 0 \end{bmatrix} = f(x, y)$$

Preprocessing for Building Vocab

1. convert all characters to lowercase

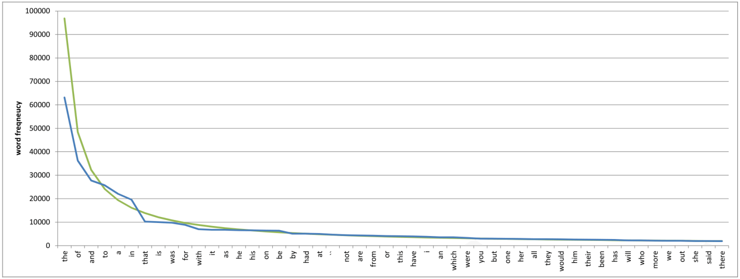
$$\text{UVa}, \text{UVA} \rightarrow \textsc{uva}$$

Preprocessing for Building Vocab

1. convert all characters to lowercase

$$\text{UVa}, \text{UVA} \longrightarrow \textsc{uva}$$

2. map low frequency words to a special token UNK



Zipf's law: $f(w_t) \propto 1/r_t$

Information Embedded in BoW Representation

- Lose:
  - word order
  - sentence boundary
  - paragraph boundary
  - …
- Keep: words in texts

Alternative Formulation

Example I: POSITIVE

Super quick and really friendly staff. I like starting off my mornings at this store!!

Vocab $\quad\quad\quad\quad \boldsymbol{f}(\boldsymbol{x}) \quad\quad\quad\quad \boldsymbol{\theta}_y$

$$\begin{Bmatrix} \text{SUPER} \\ \cdots \\ \text{QUICK} \\ \text{FOOD} \\ \text{FRIENDLY} \\ \text{EAT} \\ \cdots \\ \text{STAFF} \end{Bmatrix}$$

Alternative Formulation (Cont.)

$$\left\{\begin{array}{c} \langle\textsc{super},\text{-}1\rangle \\ \langle\textsc{super},1\rangle \\ \dots \\ \langle\textsc{friendly},\text{-}1\rangle \\ \langle\textsc{friendly},1\rangle \\ \langle\textsc{eat},\text{-}1\rangle \\ \langle\textsc{eat},1\rangle \\ \dots \\ \langle\textsc{delicious},\text{-}1\rangle \\ \langle\textsc{delicious},1\rangle \end{array}\right\}$$

Overview

1. Problem Definition
2. Case Study: Sentiment Analysis
3. Bag-of-Words Representation
4. Perceptron Algorithm
5. Classification Evaluation

Question

How to learn $\boldsymbol{\theta}$ from training examples?

(Online) Supervised Learning

Given a training example $(\boldsymbol{x}, y)$

- Predict $\hat{y}$ as

$$\hat{y} = \arg\max_{y'} \boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, y')$$

- If $y \neq \hat{y}$, update $\boldsymbol{\theta}$

Perceptron Algorithm: Updating rule

If $\hat{y} \neq y$:

$$\begin{aligned} \boldsymbol{\theta}^{(\text{new})} \leftarrow \boldsymbol{\theta}^{(\text{old})} &+ \boldsymbol{f}(\boldsymbol{x}, y) && \rightsquigarrow && \text{ground truth} \\ &- \boldsymbol{f}(\boldsymbol{x}, \hat{y}) && \rightsquigarrow && \text{predicted label} \end{aligned} \tag{7}$$

[Eisenstein, 2018, Sec. 2.2.1]

Justification

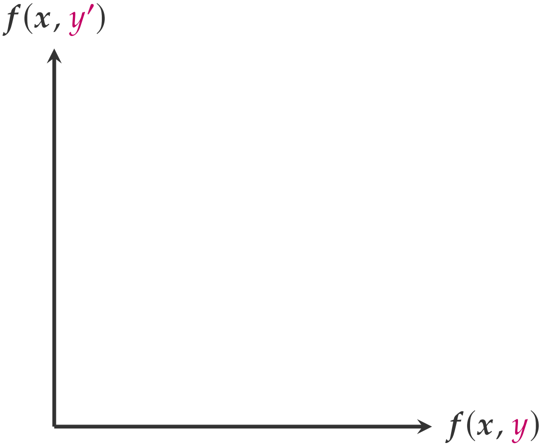
$$\boldsymbol{\theta}^{(\text{new})} \leftarrow \boldsymbol{\theta}^{(\text{old})} + \boldsymbol{f}(\boldsymbol{x}, y) - \boldsymbol{f}(\boldsymbol{x}, \hat{y}) \tag{8}$$

$\forall y' \in \mathcal{Y}$

$$\begin{aligned}
(\boldsymbol{\theta}^{(\text{new})})^\top \boldsymbol{f}(\boldsymbol{x}, y') &= (\boldsymbol{\theta}^{(\text{old})})^\top \boldsymbol{f}(\boldsymbol{x}, y') \\
&+ (\boldsymbol{f}(\boldsymbol{x}, y))^\top \boldsymbol{f}(\boldsymbol{x}, y') \\
&- (\boldsymbol{f}(\boldsymbol{x}, \hat{y}))^\top \boldsymbol{f}(\boldsymbol{x}, y')
\end{aligned}$$

Geometric Interpretation

$$\theta^{(\text{new})} \leftarrow \theta^{(\text{old})} + f(x, y) - f(x, \hat{y}) \tag{9}$$



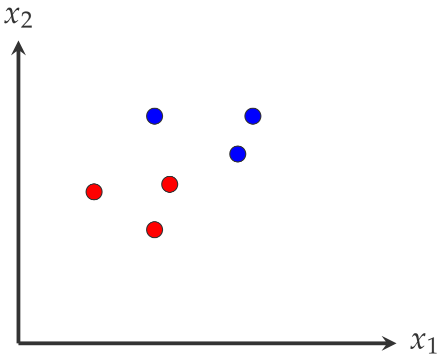
Algorithm Overview

Algorithm 3 Perceptron learning algorithm

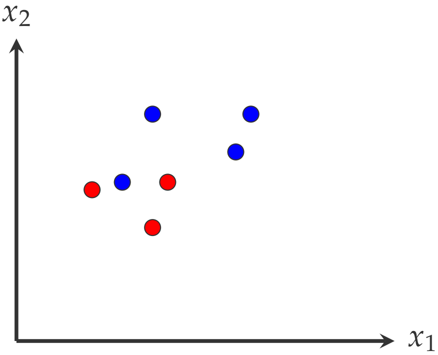
procedure PERCEPTRON($\boldsymbol{x}^{(1:N)}, y^{(1:N)}$)
    $t \leftarrow 0$
    $\boldsymbol{\theta}^{(0)} \leftarrow \mathbf{0}$
    repeat
        $t \leftarrow t + 1$
        Select an instance i
        $\hat{y} \leftarrow \text{argmax}_y \, \boldsymbol{\theta}^{(t-1)} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y)$
        if $\hat{y} \neq y^{(i)}$ **then**
            $\boldsymbol{\theta}^{(t)} \leftarrow \boldsymbol{\theta}^{(t-1)} + \boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}) - \boldsymbol{f}(\boldsymbol{x}^{(i)}, \hat{y})$
        else
            $\boldsymbol{\theta}^{(t)} \leftarrow \boldsymbol{\theta}^{(t-1)}$
    until tired
    return $\boldsymbol{\theta}^{(t)}$

Convergence: Separable Case

The algorithm will converge, if the training examples are well separated w.r.t. labels



Non-Separable Case



Averaged Perceptron

In practice, average all the classification weights $\boldsymbol{\theta}_t$ across all the iterations

$$\overline{\boldsymbol{\theta}} = \frac{1}{T} \sum_t \boldsymbol{\theta}^{(t)} \tag{10}$$

Then,

$$\Psi(\boldsymbol{x}, y) = \overline{\boldsymbol{\theta}}^\top \boldsymbol{f}(\boldsymbol{x}, y) \tag{11}$$

[Eisenstein, 2018, Sec. 2.2.2]

Overview

1. Problem Definition
2. Case Study: Sentiment Analysis
3. Bag-of-Words Representation
4. Perceptron Algorithm
5. Classification Evaluation

Standard Setup of Statistical Machine Learning

- Training set $\mathcal{T} = \{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$
- Development set $\mathcal{D} = \{(\boldsymbol{x}^{(j)}, y^{(j)})\}_{j=1}^{M}$
- Test set $\mathcal{U} = \{(\boldsymbol{x}^{(l)}, y^{(l)})\}_{l=1}^{L}$

Evaluation Measurements

- Accuracy
- Precision, recall, and F-measure

[Eisenstein, 2018, Sec 4.4]

Accuracy

$$\textsc{acc}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{N}\delta(y^{(i)}, \hat{y}^{(i)}) \tag{12}$$

δ function

$$\delta(y, \hat{y}) = \begin{cases} 1 & y = \hat{y} \\ 0 & y \neq \hat{y} \end{cases} \tag{13}$$

The loss function of perceptron algorithm

True/False Positive/Negative

For a particular category/class, e.g., user reviews with rating 1 on Yelp

		Ground truth	
		POSITIVE	NEGATIVE
Prediction	POSITIVE	True Positive (TP)	False Positive (FP)
	NEGATIVE	False Negative (FN)	True Negative (TN)

Recall, Precision and F Measure

		Ground truth	
		POSITIVE	NEGATIVE
Prediction	POSITIVE	True Positive (TP)	False Positive (FP)
	NEGATIVE	False Negative (FN)	True Negative (TN)

Recall:

$$r(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{14}$$

Precision:

$$p(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{15}$$

F measure:

$$F(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{2 \cdot p \cdot r}{p + r} \tag{16}$$

Summary

- Bag-of-words representations
- Perceptron updating rule

$$\theta^{(\text{new})} \leftarrow \theta^{(\text{old})} + f(x, y) - f(x, \hat{y}) \tag{17}$$

- Evaluation measurements
  - Accuracy
  - Recall, Precision and F-Measure

Reference

Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., Downey, D., Dunkelberger, J., Elgohary, A., Feldman, S., Ha, V., Kinney, R., Kohlmeier, S., Lo, K., Murray, T., Ooi, H.-H., Peters, M. E., Power, J., Skjonsberg, S., Wang, L. L., Wilhelm, C., Yuan, Z., van Zuylen, M., and Etzioni, O. (2018).
Construction of the literature graph in semantic scholar.
In *NAACL-HTL*.

Eisenstein, J. (2018).
Natural Language Processing.
MIT Press.

Pang, B., Lee, L., and Vaithyanathan, S. (2002).
Thumbs up?: sentiment classification using machine learning techniques.
In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.