Basic Logistic Regression Classifier

Natalie Parde
UIC CS 421

In general, supervised machine learning systems for text classification have four main components.

- **Feature representation** of the input
  - Typically, a **vector** of features $[x_1^{(j)}, x_2^{(j)}, \ldots, x_n^{(j)}]$ for a given instance $x^{(j)}$
- **Classification function** that computes the estimated class, $\hat{y}$
  - Sigmoid
  - Softmax
  - Etc.
- **Objective function** or **loss function** that computes error values on training instances
  - Cross-entropy loss function
- **Optimization function** that seeks to minimize the loss function
  - Stochastic gradient descent

To build a logistic regression classifier….

- Train **weights** *w* and a **bias** *b* using **stochastic gradient descent** and **cross-entropy loss**
- Use a **sigmoid classification function**
- Test performance by computing P(y|x) and returning the **highest-probability label**

Binary Logistic Regression

- Goal:
    - Train a classifier that can decide whether a new input observation belongs to class *a* or class *b*
- To do this, the classifier learns a **vector of weights** (one associated with each input feature) and a **bias term**
- A given **weight indicates how important its corresponding feature is** to the overall classification decision
    - Can be positive or negative
- The **bias term is a real number** that is added to the weighted inputs

Binary Logistic Regression

- To make a classification decision, the classifier:
  - Multiplies each feature for an input instance *x* by its corresponding weight (learned from the training data)
  - Sums the weighted features
  - Adds the bias term *b*
- This results in a weighted sum of evidence for the class:
  - $z = b + \sum_i w_i x_i$

Bias term

Weight for feature *i*

Feature *i* for instance *x*

⁎ Vector Notation

- Letting w be the weight vector and x be the input feature vector, we can also represent the weighted sum z using vector notation:
  - $z = \boldsymbol{w} \cdot \boldsymbol{x} + b$

Vector of all weights

Bias term

Vector of all features for *x*

Multiplying feature values by their weights means that *z* is a linear function of *x*

- What we really want is a **probability** ranging from 0 to 1
- To do this, we pass *z* through the sigmoid function, $\sigma(z)$
  - Also called the **logistic function**, hence the name **logistic regression**

Sigmoid Function

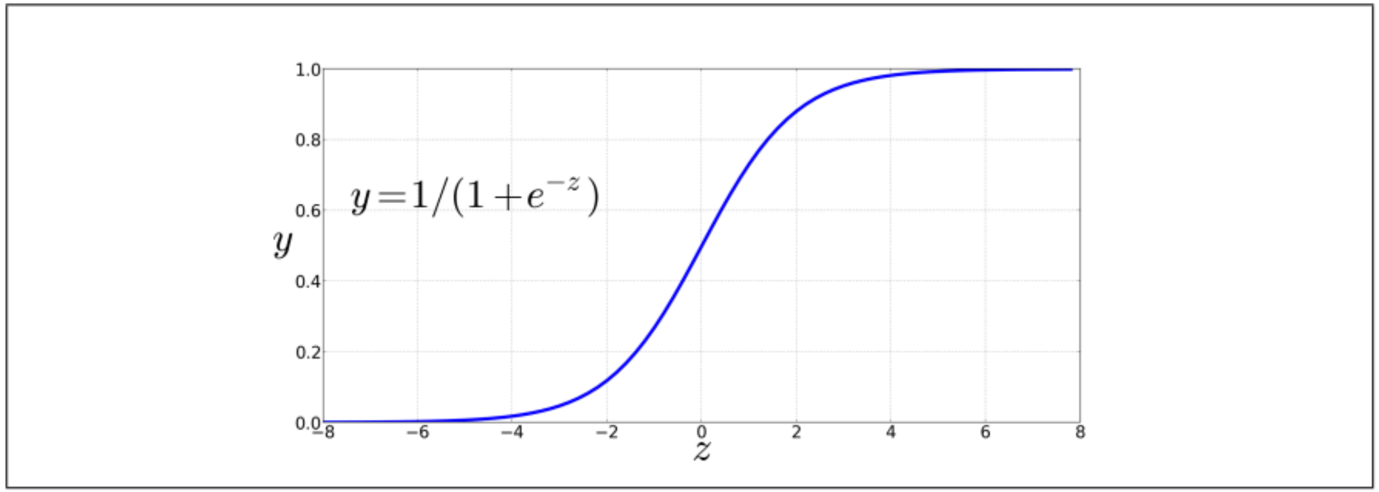

Figure 5.1 The sigmoid function $y = \frac{1}{1+e^{-z}}$ takes a real value and maps it to the range $[0, 1]$. It is nearly linear around 0 but outlier values get squashed toward 0 or 1.

Source: https://web.stanford.edu/~jurafsky/slp3/5.pdf

- Sigmoid Function:
  - $\sigma(x) = \frac{1}{1+e^{-x}}$
- Given its name because when plotted, it looks like an *s*
- Results in a value *y* ranging from 0 to 1
  - $y = \sigma(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-w\cdot x+b}}$

There are many useful properties of the sigmoid function!

- Maps a real-valued number to a 0 to 1 range
  - Just what we need for a probability….
- Squashes outlier values towards 0 or 1
- Differentiable
  - Necessary for learning….

In binary logistic regression, to make the probability for all classes sum to one….

- $P(y = 1) = \sigma(z)$
- $P(y = 0) = 1 - \sigma(z)$

How do we make a classification decision?

- Choose a **decision boundary**
  - For binary classification, often 0.5
- For a test instance *x*, assign a label *c* if $P(y = c|x)$ is greater than the decision boundary
  - If performing binary classification, assign the other label if $P(y = c|x)$ is lower than or equal to the decision boundary

Example: Sigmoid Classification

I'm just thrilled that I have five final exams on the same day. 🙄

Sarcastic or not sarcastic?

Example: Sigmoid Classification

I’m just thrilled that I have five final exams on the same day. 🙄

Sarcastic or not sarcastic?

Feature
Contains 🙄
Contains 😊
Contains “I’m”

Example: Sigmoid Classification

I'm just thrilled that I have five final exams on the same day. 🙄

Sarcastic or not sarcastic?

Feature	Weight
Contains 🙄	2.5
Contains 😊	-3.0
Contains "I'm"	0.5

Example: Sigmoid Classification

I'm just thrilled that I have five final exams on the same day. 🙄

Sarcastic or not sarcastic?

Feature	Weight
Contains 🙄	2.5
Contains 😊	-3.0
Contains "I'm"	0.5

Positively associated with sarcasm

Negatively associated with sarcasm

Example: Sigmoid Classification

I'm just thrilled that I have five final exams on the same day. 🙄

Sarcastic or not sarcastic?

Feature	Weight	Value
Contains 🙄	2.5	1
Contains 😊	-3.0	0
Contains "I'm"	0.5	1

Example: Sigmoid Classification

I'm just thrilled that I have five final exams on the same day. 🙄

Sarcastic or not sarcastic?

Feature	Weight	Value
Contains 🙄	2.5	1
Contains 😊	-3.0	0
Contains "I'm"	0.5	1

Bias = 0.1

Example: Sigmoid Classification

I'm just thrilled that I have five final exams on the same day. 🙄

Sarcastic or not sarcastic?

Feature	Weight	Value
Contains 🙄	2.5	1
Contains 😊	-3.0	0
Contains "I'm"	0.5	1

Bias = 0.1

$$z = b + \sum_i w_i x_i$$

$$y = \sigma(z) = \frac{1}{1+e^{-z}}$$

Example: Sigmoid Classification

I'm just thrilled that I have five final exams on the same day. 🙄

Sarcastic or not sarcastic?

Feature	Weight	Value
Contains 🙄	2.5	1
Contains 😊	-3.0	0
Contains "I'm"	0.5	1

Bias = 0.1

$$z = b + \sum_i w_i x_i$$

$$y = \sigma(z) = \frac{1}{1+e^{-z}}$$

$$P(\text{sarcasm}|x) = \sigma\big(0.1 + (2.5 * 1 + (-3.0) * 0 + 0.5 * 1)\big) = \sigma(0.1 + 3.0) = \sigma(3.1) = \frac{1}{1 + e^{-3.1}} = 0.96$$

Example: Sigmoid Classification

I'm just thrilled that I have five final exams on the same day. 🙄

Sarcastic or not sarcastic?

Feature	Weight	Value
Contains 🙄	2.5	1
Contains 😊	-3.0	0
Contains "I'm"	0.5	1

Bias = 0.1

$$z = b + \sum_i w_i x_i$$

$$y = \sigma(z) = \frac{1}{1+e^{-z}}$$

$$P(\text{sarcasm}|x) = \sigma\big(0.1 + (2.5 * 1 + (-3.0) * 0 + 0.5 * 1)\big) = \sigma(0.1 + 3.0) = \sigma(3.1) = \frac{1}{1 + e^{-3.1}} = 0.96$$

$$P(\text{not sarcasm}|x) = 1 - \sigma\big(0.1 + (2.5 * 1 + (-3.0) * 0 + 0.5 * 1)\big) = 1 - \sigma(0.1 + 3.0) = 1 - \sigma(3.1) = 1 - \frac{1}{1 + e^{-3.1}} = 1 - 0.96 = 0.04$$

Example: Sigmoid Classification

I'm just thrilled that I have five final exams on the same day. 🙄

Sarcastic or not sarcastic?

Feature	Weight	Value
Contains 🙄	2.5	1
Contains 😊	-3.0	0
Contains "I'm"	0.5	1

Bias = 0.1

$$z = b + \sum_i w_i x_i$$

$$y = \sigma(z) = \frac{1}{1+e^{-z}}$$

$$P(\text{sarcasm}|x) = \sigma\big(0.1 + (2.5 * 1 + (-3.0) * 0 + 0.5 * 1)\big) = \sigma(0.1 + 3.0) = \sigma(3.1) = \frac{1}{1 + e^{-3.1}} = 0.96$$

$$P(\text{not sarcasm}|x) = 1 - \sigma\big(0.1 + (2.5 * 1 + (-3.0) * 0 + 0.5 * 1)\big) = 1 - \sigma(0.1 + 3.0) = 1 - \sigma(3.1) = 1 - \frac{1}{1 + e^{-3.1}} = 1 - 0.96 = 0.04$$

A little bit about features….

- Anything can be a feature!
  - Specific words or n-grams
  - Information from external lexicons
  - Grammatical elements
  - Part-of-speech tags
- In neural classification models, the feature vector often includes word embeddings
  - More about these soon!

Learning in Logistic Regression

- How are the parameters of a logistic regression model, *w* and *b*, learned?
  - **Loss function**
  - **Optimization function**
- Goal: Learn parameters that make $\hat{y}$ for each training observation as close as possible to the true y