2.4 Logistic Regression

Edwin Simpson

Department of Computer Science,

University of Bristol, UK.

Limitations of Naïve Bayes

Reality: the conditional independence assumption is often violated!

$$P(\boldsymbol{x}|y) = \prod_{i=1}^{N} P(x_i|y)$$

- Closed form computations
- We can easily add new conditionally independent features

- if the document contains "*Bayes*", "*naïve*" is likely to appear regardless of the class
- if the text contains the bigram "*not good*", then the unigram "good" is certain to occur

Logistic Regression

Using Bayes' rule (a generative classifier):

$$P(y|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|y)P(y)}{P(\boldsymbol{x})}$$

Binary logistic regression (a discriminative classifier):

$$P(y|\boldsymbol{x}) = \frac{1}{1 + e^{-\sum_{i=1}^{N} \theta_i \cdot x_i}}$$

Logistic Regression

Apply weights to each feature:	$\theta_i \cdot x_i$

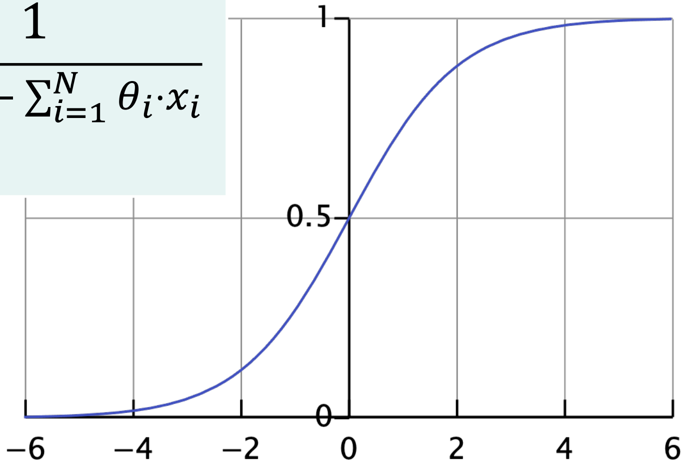
Features may be continuous, not just discrete occurrences.

Logistic Regression

Apply weights to each feature:	$\theta_i \cdot x_i$
Combine their individual contributions:	$\sum_{i=1}^{N} \theta_i \cdot x_i = \boldsymbol{\theta} \cdot \boldsymbol{x}$

Logistic Regression

Apply weights to each feature:	$\theta_i \cdot x_i$
Combine their individual contributions:	$\sum_{i=1}^{N} \theta_i \cdot x_i = \boldsymbol{\theta} \cdot \boldsymbol{x}$
Use the logistic sigmoid to map unbounded real numbers to values between 0 and 1:	$\sigma(\theta . x) = \frac{1}{1 + e^{-\sum_{i=1}^{N} \theta_i \cdot x_i}}$



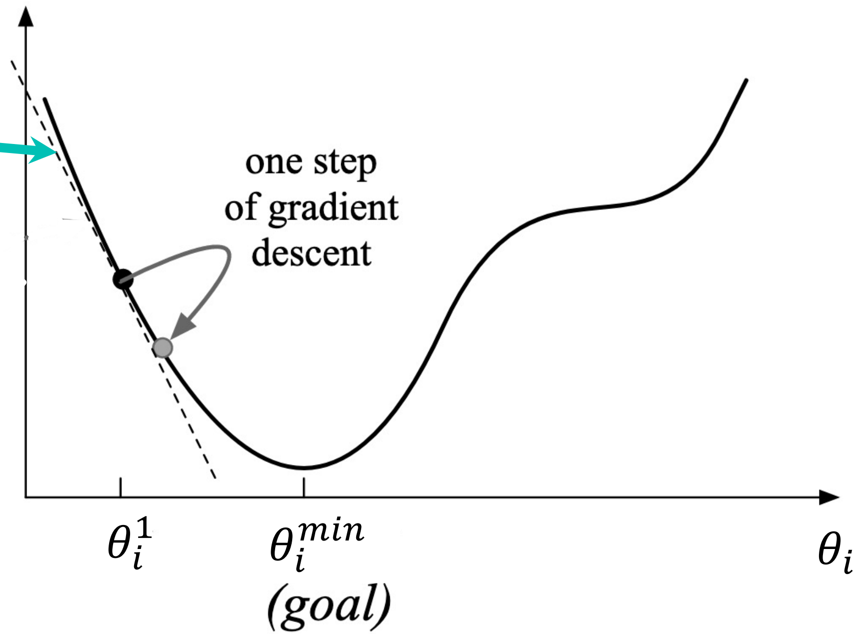
Learning Objective

- Goal: choose $\boldsymbol{\theta}$ to make training set predictions as close as possible to the training labels
- Measure how far we are from this objective using a **loss** or **cost** function:
$$L(\boldsymbol{\theta}; y) = -[y \log p(y|\boldsymbol{x}, \boldsymbol{\theta}) + (1-y) \log(1 - p(y|\boldsymbol{x}, \boldsymbol{\theta}))]$$
- Find the values of $\boldsymbol{\theta}$ that minimise the total loss on the training set using gradient descent

Gradient Descent

[Chapter 5, Speech and Language Processing (3rd edition draft), Jurafsky & Martin (2021).]

- Initialise all weights θ_i to random values
- Compute gradient of $L(\boldsymbol{\theta}; y)$ with respect to θ_i
- Increase or decrease θ_i in the opposite direction to the gradient
- Thereby climb down the hill toward the minimum of the loss



Summary

- Combinations of features often violate the naïve Bayes conditional independence assumption, which could decrease performance.
- Logistic regression is a discriminative classifier that relaxes this assumption and allows continuous-valued features.
- The parameters (weights) are learned by minimising a loss function using gradient descent.

Chapter 5, Speech and Language Processing, Jurafsky & Martin (2021).
Chapter 2, Introduction to Natural Language Processing, Eisenstein (2019).