Chapter 5 Logistic Regression

9/9/20 10:53 PM

Discriminative classifier v.s. generative classifier

- Generative model:
    - Example: naïve Bayes
    - Uses likelihood term
    - Asks how to generate features of documents if we know it's in a certain class
- Discriminative model:
    - Directly compute the conditional probability given features
    - Perhaps will learn to assign weights to better distinguish classes, but will not generate example of any classes

Components of probablistic machine learning classifier:

- A feature representation of inputs
- A classification function that computes the estimated class, through conditional probability
- An object function for learning, usually minimizing error of training samples
- An algorithm for optimizing the object function, e.g. stochastic gradient descent

Sigmoid function / Logistic function

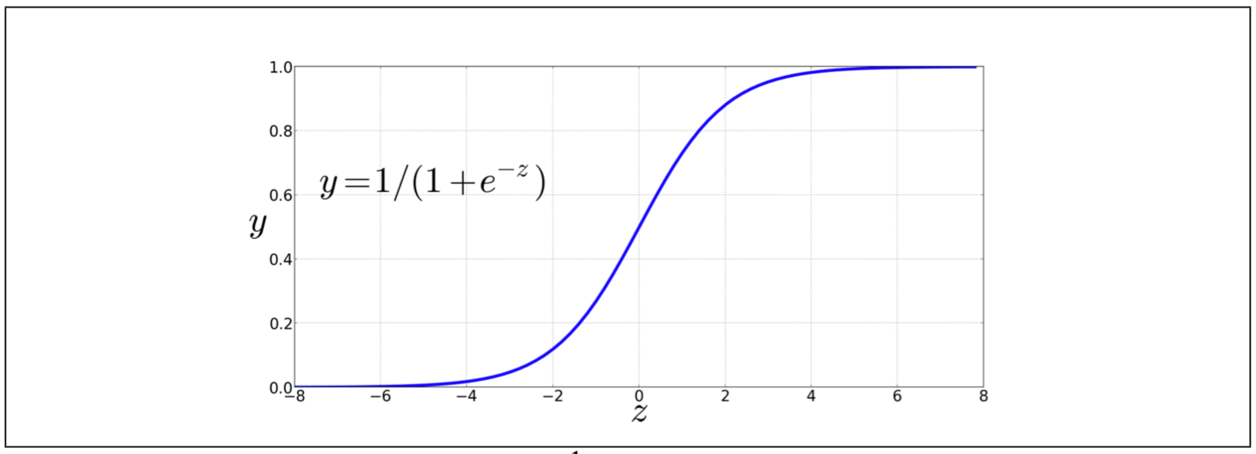

Figure 5.1 The sigmoid function $y = \frac{1}{1+e^{-z}}$ takes a real value and maps it to the range $[0,1]$. It is nearly linear around 0 but outlier values get squashed toward 0 or 1.

Representation learning

- Ways to learn features automatically in an unsupervised way from the input
- Rather than designing features by human effort every time

Logistic Regression v.s. Naïve Bayes

- Naïve Bayes requires independence assumption, which is generally not true for large corpus
- But Naïve Bayes works very well on small data sets and short documents
- Logistic Regression is much more robust on correlated features

Multinomial Logistic Regression

- Softmax function:

  The softmax of an input vector $z = [z_1, z_2, ..., z_k]$ is thus a vector itself:

$$\text{softmax}(z) = \left[\frac{e^{z_1}}{\sum_{i=1}^{k} e^{z_i}}, \frac{e^{z_2}}{\sum_{i=1}^{k} e^{z_i}}, ..., \frac{e^{z_k}}{\sum_{i=1}^{k} e^{z_i}} \right]$$

- Probability predicted:

$$p(y = c|x) = \frac{e^{w_c \cdot x + b_c}}{\sum_{j=1}^{k} e^{w_j \cdot x + b_j}}$$