



## Lecture 7 (part II): Classification DD2431

A. Maki

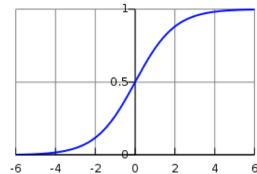
Autumn, 2015

### Logistic regression

An approach to learning functions (of the form  $f: x \rightarrow y$ ) or  $P(y | x)$  where  $y$  is **discrete-valued**, typically a boolean, and  $x$  is a vector (of discrete or continuous variables)

Sigmoid/Logistic function:

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



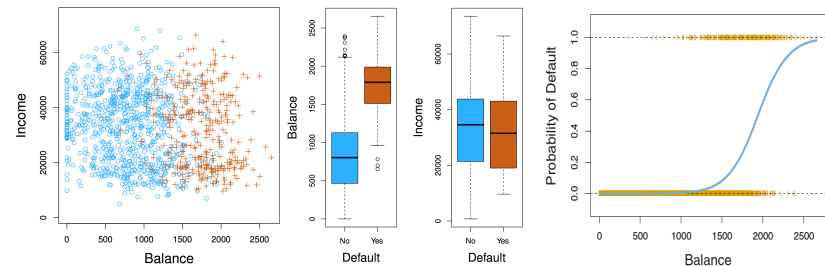
We model  $P(y | x)$  using a sigmoid function that gives **outputs between 0 and 1** (interpretable as probability) for all input values of  $x$

### We will visit

- Naïve Bayes classifier (visited in part I)
- Logistic Regression (binary classification)
- Discriminative and Generative models

Classification => a qualitative output; to assign an observation to a category (class)

### Example: Credit card default data



We are to predict customers that are likely to default

$y$  (default) is **categorical**: Yes/No

$x$  contains variables: annual income, monthly balance

Figures from An Introduction to Statistical Learning (G. James et al.)

## Model/hypothesis representation

In linear regression we had:  $f(x) = w^T x$

Here we use:  $f_w(x) = \frac{1}{1 + e^{-w^T x}}$  so that  $0 \leq f_w(x) \leq 1$

Interpretation of  $f$ : **estimated probability**

that  $y = 1$  given  $x$ , parameterized by  $w$

$$f_w(x) = P(y = 1 | x, w)$$

$$P(y = 0 | x, w) = 1 - P(y = 1 | x, w)$$

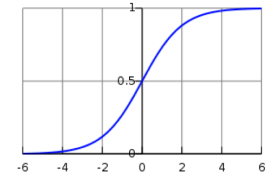
## Decision boundary

In linear regression we had:  $f(x) = w^T x$

Here we use:  $f_w(x) = \frac{1}{1 + e^{-w^T x}}$  so that  $0 \leq f_w(x) \leq 1$

Predict  $y = 1$  if  $f_w(x) \geq 0.5 \rightarrow w^T x \geq 0$

Predict  $y = 0$  if  $f_w(x) < 0.5 \rightarrow w^T x < 0$



Decision boundary

## Cost function

- Training dataset:

$$D = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$$

of  $N$  pairs of inputs  $x_n$  and targets  $y_n \in \{1, 0\}$

- Want the parameters  $w$  that minimise the error:

$$E(w) = \frac{1}{N} \sum_{n=1}^N \text{Cost}(f_w(x_n), y_n)$$

↓

$$-y \log(f_w(x)) - (1 - y) \log(1 - f_w(x))$$

## Estimating the parameters

**Gradient Decent** to find  $w$  such that  $\min_w E(w)$

$$E(w) = -\frac{1}{N} \sum_{n=1}^N [y_n \log(f_w(x_n)) + (1 - y_n) \log(1 - f_w(x_n))]$$

Repeat:  $w_i \equiv w_i - \alpha \frac{\partial}{\partial w_i} E(w)$  (Simultaneous update for all  $w_i$ )

$$= w_i - \alpha \sum_{n=1}^N (f_w(x_n) - y_n) x_{n,i}$$

For a new  $x$ , compute  $f_w(x) = \frac{1}{1 + e^{-w^T x}} = P(y = 1 | x, w)$

## Inference and decision

Three distinctive approaches to classification problem

- Discriminative function: learn a function that maps inputs directly to a class label (no access to probabilities)
- Discriminative approach
- Generative approach

Classification can be seen as inference + decision:

1. **Inference stage:** to learn a model for  $P(y | \mathbf{x})$  using training data
2. **Decision stage:** to determine optimal class membership using these posterior probabilities

## Discriminative vs Generative model

Discriminative approach:

- Directly model the posterior probabilities  $P(y | \mathbf{x})$

Generative approach:

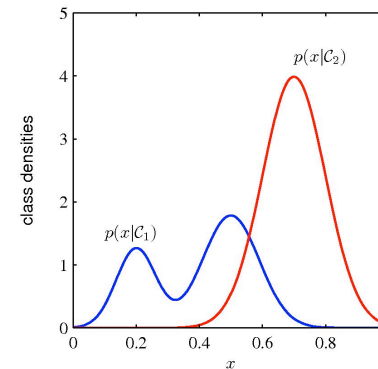
- First solve the inference of determining  $P(\mathbf{x} | y)$  for each class
- Infer the prior class probability  $P(y)$ , often just by the fraction
- Use Bayes' theorem

The difference mainly in computing  $P(\mathbf{x} | y)$

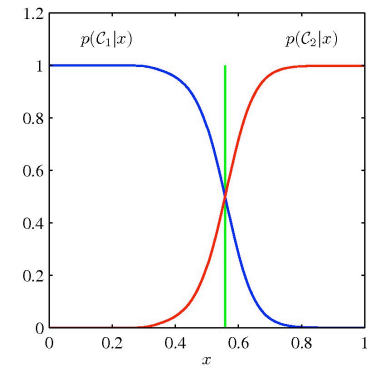
- Demanding, requiring a large training set, and computation
- + Possible to generate synthetic data points in the input space

## Example: two classes, single variable

Class-conditional densities



Posterior probabilities



Figures from Pattern Recognition and Machine Learning (C. Bishop)