Data Mining: Classification 2

Jakramate Bootkrajang

School of Computer Science
Chiang Mai University

Adapted from materials by Ata Kabán

Outline

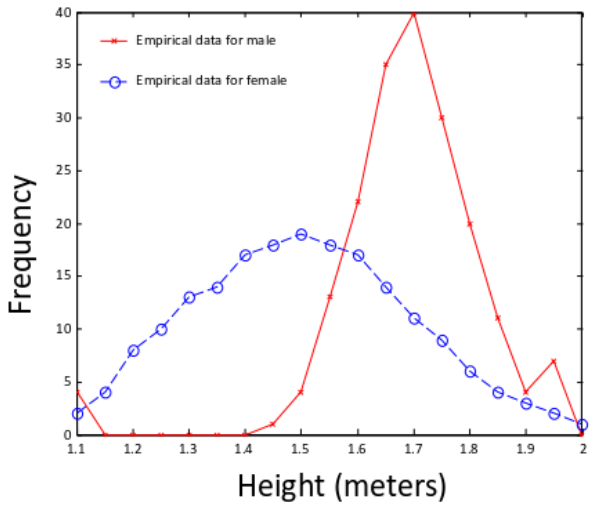
- Introduction
- Bayesian Learning
- Generative Classifier
- Discriminative Classifier
- Nearest Neighbour Classifier
- Classifier Evaluation [Maybe]

Introduction

- We have already seen how Bayes rule can be turned into a classifier
- Our examples so far only consider discrete attributes (e.g. $\{$sunny, warm$\}$, $\{+, -\}$
- Today we learn how to do this when the data attributes are continuous valued

An example problem

- Task: predict gender of individuals based on their heights
- Given
  - 100 height examples of women
  - 100 height examples of man



Class posterior

- From Bayes rule we can obtain the class posteriors of male:
- $P(h = 1|x) = \frac{P(x|h{=}1)(P(h{=}1)}{P(x|h{=}0)(P(h{=}0)+P(x|h{=}1)(P(h{=}1)}$
- The denominator is the probability of measuring the height vale x irrespective of the class.
- If we can compute this, we are done. We can use it to predict gender for height.

Class priors

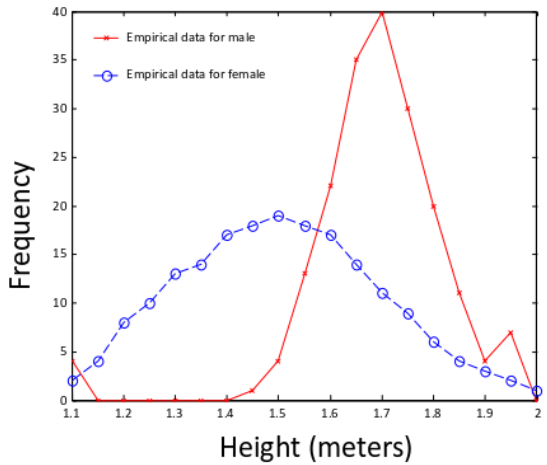
- We can encode the values of the hypothesis (class) as 1(male) and 0(female). So, $h \in \{0, 1\}$.
- Since in this example we had the same number of males and females, we have $P(h = 1) = P(h = 0) = 0.5$.
- These are priors of class membership as they can be set before measuring any data.
- In cases when class proportions are imbalanced, we can use the priors to make predictions even before seeing any data.

Class-conditional likelihood

- Our measurements are heights. This is our data, x.
- Class-conditional likelihoods
  - $P(x|h = 1)$: probability that a male has height x metres.
  - $P(x|h = 0)$: ??
- How to get this?

Discriminant function (1/3)

- When does our prediction switch from prediction $h = 0$ vs predicting $h = 1$?



- When the measured height passes a certain threshold...or...more precisely, when $P(h = 0|x) = P(h = 1|x)$.

Discriminant function (2/3)

- If we make a measurement, say we get $x = 1.7m$.
- If we compute the posteriors and find $P(h = 1|x = 1.7) > P(h = 0|x = 1.7)$
- We then decide to predict $h = 1$.
- If we measured $x = 1.2m$, we will get $P(h = 1|x = 1.2) < P(h = 0|x = 1.2)$

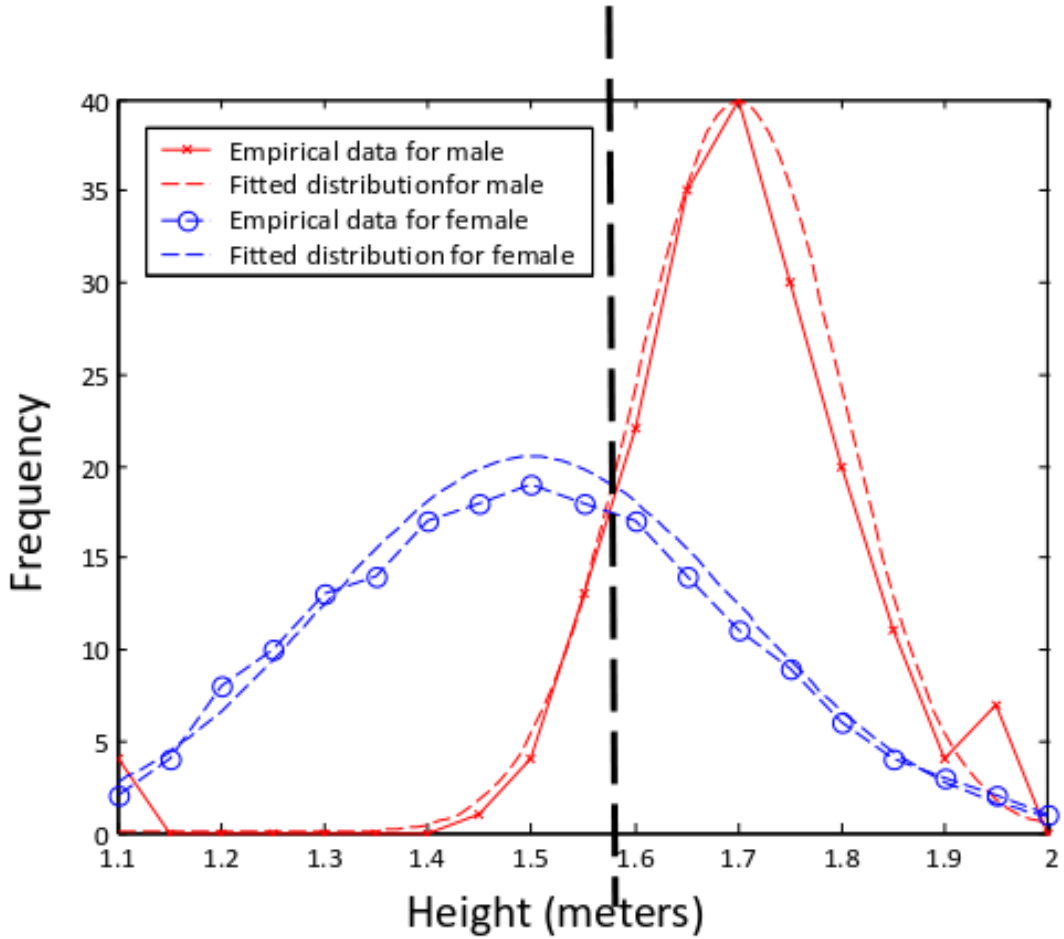
Discriminant function (3/3)

- We can define a discriminant function as: $f_1(x) = \frac{P(h{=}1|x)}{P(h{=}0|x)}$ and compare the value to 1.
- It's more convenient to have the switching at 0 $f_2(x) = \log \frac{P(h{=}1|x)}{P(h{=}0|x)}$.
- Then the sign of this function defines the prediction $f_2(x) > 0 = \text{male}$, $f_2(x) < 0 = \text{female}$

How do we compute it?

- Using Bayes rule
  $f_2(x) = \log \frac{P(h{=}1|x)}{P(h{=}0|x)} = \log \frac{P(x|h{=}1)P(h{=}1)}{P(x|h{=}0)P(h{=}0)}$.
- Now, we need the class conditional likelihood terms, $P(x|h=1), P(x|h=0)$
- We will model each class by a Gaussian distribution. (Other distribution is possible)

Illustration - our 1D example



Univariate Gaussian (Normal Distribution)

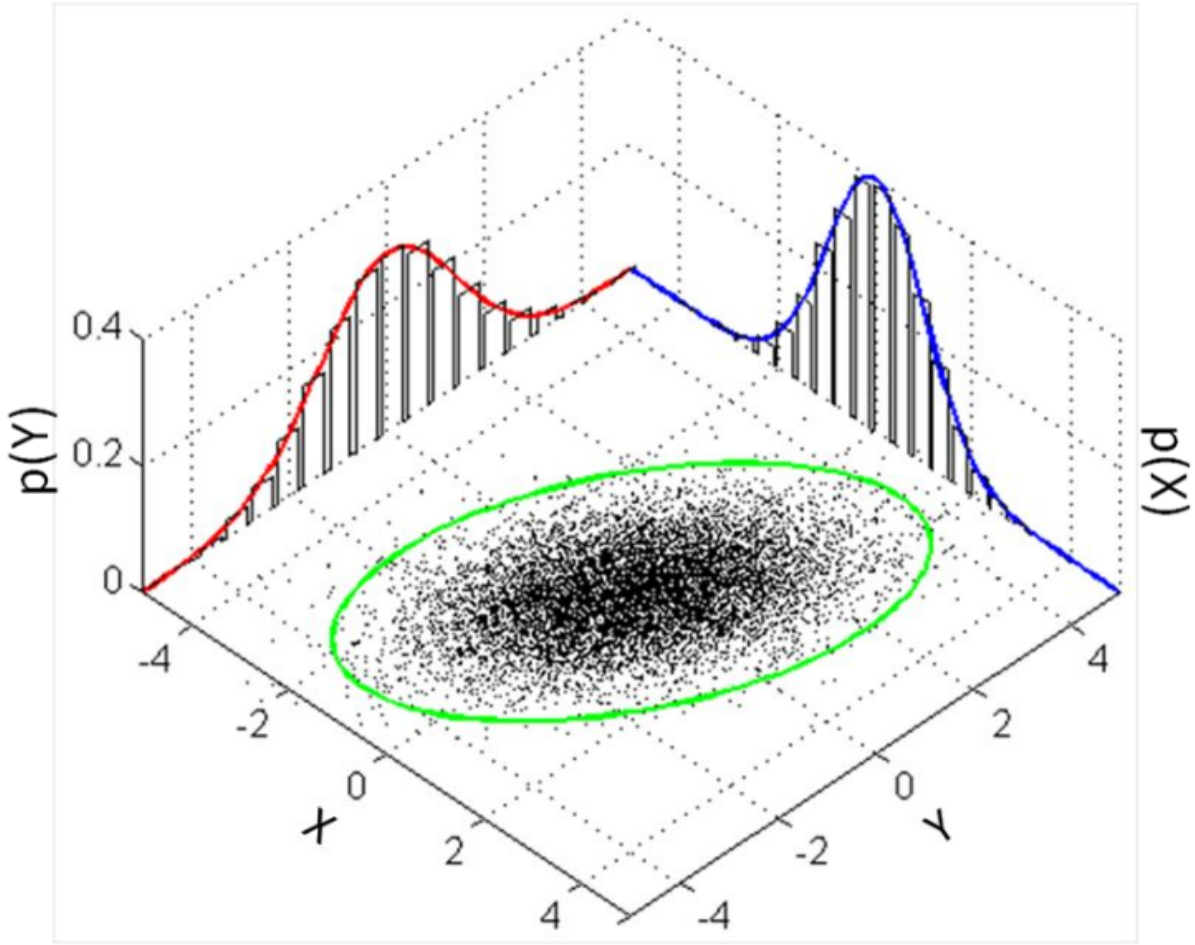
$$P(x) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left\{-\frac{(x-m_k)^2}{2\sigma_k^2}\right\}$$

- where m_k is the mean(centre), and σ_k^2 is the variance (spread). These are parameters that describe the distributions.
- We will have separate Gaussian for each class. So, the female class will have m_0 as its mean, and σ_0^2 as its variance. So as male class with m_1 and σ_1^2.
- We will estimate these parameters from the data.

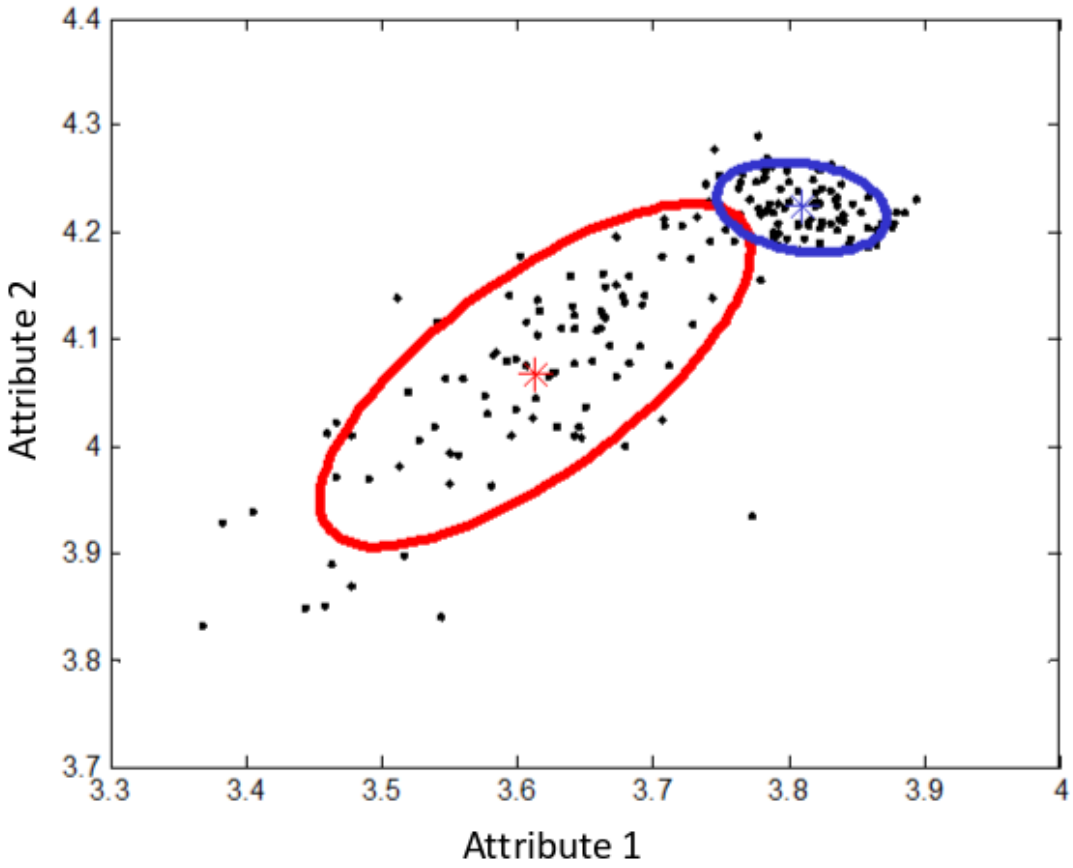
Multivariate Gaussian

- Let $\mathbf{x} = \{x_1, x_2, x_3, \ldots, x_d\}$ Let $k \in \{0, 1\}$
  $P(\mathbf{x}) =$
  $\frac{1}{\sqrt{2\pi|\Sigma_k|^2}} \exp\{-\frac{1}{2}(\mathbf{x} - \mathbf{m}_k)^T \Sigma_k^{-1}(\mathbf{x} - \mathbf{m}_k)$
- where $\mathbf{m}_k$ is the mean vector, and Σ_k is the covariance matrix.
- These parameters get estimated from the data.

Illustration - 2D example



Example with 2 classes



Naive Bayes

- The full covariance are $d \times d$
- In many situation there is not enough data to estimate full covariance.
- The Naive Bayes is again useful and tends to work well in practice.
- Using Naive assumption the covariance becomes diagonal.

What's next

- How to estimate the parameters? $\mathbf{m}$ and Σ.
- If we use naive assumption we can compute the estimates in each class separately for each feature(dimension).
- If d is small full covariance (not using Naive assumption) is expected to work better.
- In MATLAB there are built-in functions that you can use `mean()`, `cov()`, `var()`
- The sample mean and sample covariance obtain is a Maximum Likelihood estimates of the population mean and population covariance.

Multi-class classification

- We may have more than two classes. 'Healthy', 'Disease 1', 'Disease 2'.
- Our Gaussian classifier is easy to use in multi-class problem.
- We compute posterior probability for each of the classes.
- We predict class with highest posterior.

Summary

- This type of classifier is call *Generative* because it makes an assumption that the points in each class are generated by some distribution i.e., Gaussian distribution in our example.
- One can model the discriminant function directly. That is called *Discriminative* classifier. (which we will learn shortly).