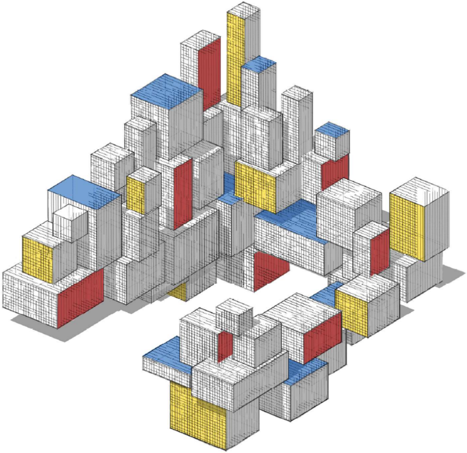
Bayes Classification



Purpose

In this lecture we discuss:

- Classification via Bayes' rule
- Naïve Bayes method
- Discriminant functions

Classification via Bayes' Rule

Recall that, for the 0-1 loss, the optimal classifier for classes $0, \ldots, c-1$ is

$$g^*(\boldsymbol{x}) = \underset{y \in \{0,\ldots,c-1\}}{\operatorname{argmax}} \ \mathbb{P}[Y = y \,|\, \boldsymbol{X} = \boldsymbol{x}]. \tag{1}$$

In a Bayesian setting, the conditional probability $\mathbb{P}[Y = y \,|\, \boldsymbol{X} = \boldsymbol{x}]$ is interpreted as a posterior probability. Using Bayesian notation, we write $f(y \,|\, \boldsymbol{x})$ for $\mathbb{P}[Y = y \,|\, \boldsymbol{X} = \boldsymbol{x}]$.

Recall from Bayesian analysis that the posterior is of the form

$$f(y \,|\, \boldsymbol{x}) \propto f(\boldsymbol{x} \,|\, y) f(y), \tag{2}$$

where $f(\boldsymbol{x} \,|\, y)$ is the likelihood of obtaining feature vector $\boldsymbol{x}$ from class y and $f(y)$ is the prior probability of class y.

Bayes Optimal Decision Rule

By making various modeling assumptions about the prior (e.g., all classes are *a priori* equally likely) and the likelihood function, one obtains the posterior pdf via Bayes' formula (2).

A class $\widehat{y}$ is then assigned to a feature vector $\boldsymbol{x}$ according to the highest posterior probability; that is, we classify according to the Bayes optimal decision rule:

$$\widehat{y} = \underset{y}{\operatorname{argmax}}\, f(y \,|\, \boldsymbol{x}), \tag{3}$$

which is exactly (1).

Naïve Bayes Method

Since the discrete density $f(y \,|\, \boldsymbol{x})$, $y = 0, \ldots, c-1$ is usually not known, the aim is to approximate it well with a function $g(y \,|\, \boldsymbol{x})$ from some class of functions $\mathcal{G}$.

Suppose a feature vector $\boldsymbol{x} = [x_1, \ldots, x_p]^\top$ of p features has to be classified into one of the classes $0, \ldots, c-1$. For example, the classes could be different people and the features could be various facial measurements, such as the width of the eyes divided by the distance between the eyes, or the ratio of the nose height and mouth width.

In the naïve Bayes method, the class of approximating functions $\mathcal{G}$ is chosen such that

$$g(\boldsymbol{x} \,|\, y) = g(x_1 \,|\, y) \cdots g(x_p \,|\, y)$$

That is, conditional on the label, all features are independent.

Naïve Bayes Method

Assuming a uniform prior for y, the posterior pdf is thus:

$$g(y \,|\, \boldsymbol{x}) \propto \prod_{j=1}^{p} g(x_j \,|\, y),$$

where the marginal pdfs $g(x_j \,|\, y)$, $j = 1, \ldots, p$ belong to a given class of approximating functions $\mathcal{G}$.

If the approximating class $\mathcal{G}$ is such that $(X_j \,|\, y) \sim \mathcal{N}(\mu_{yj}, \sigma^2)$, $y = 0, \ldots, c-1$, $j = 1, \ldots, p$, then posterior pdf is:

$$g(y \,|\, \boldsymbol{\theta}, \boldsymbol{x}) \propto \exp\left(-\frac{1}{2} \sum_{j=1}^{p} \frac{(x_j - \mu_{yj})^2}{\sigma^2}\right).$$

To classify $\boldsymbol{x}$, simply take the y that maximizes the unnormalized posterior pdf.

Naïve Bayes Method

We can write the posterior pdf as

$$g(y \mid \boldsymbol{\theta}, \boldsymbol{x}) \propto \exp\left(-\frac{1}{2}\frac{\|\boldsymbol{x} - \boldsymbol{\mu}_y\|^2}{\sigma^2}\right),$$

where $\boldsymbol{\mu}_y := [\mu_{y1}, \ldots, \mu_{yp}]^\top$ and $\boldsymbol{\theta} := \{\boldsymbol{\mu}_0, \ldots, \boldsymbol{\mu}_{c-1}, \sigma^2\}$ collects all model parameters.

The probability $g(y \mid \boldsymbol{\theta}, \boldsymbol{x})$ is maximal when $\|\boldsymbol{x} - \boldsymbol{\mu}_y\|$ is minimal.

Thus $\widehat{y} = \mathrm{argmin}_y \|\boldsymbol{x} - \boldsymbol{\mu}_y\|$ is the classifier.

That is, classify $\boldsymbol{x}$ as y when $\boldsymbol{\mu}_y$ is closest to $\boldsymbol{x}$ in Euclidean distance.

Usually, the parameters (here, the $\{\boldsymbol{\mu}_y\}$ and σ^2) are unknown and have to be estimated from the training data.

Example: Naïve Bayes Classification

We extend the previous classifier to the case where also the variance σ^2 depends on the class y and feature j.

The table lists the means μ and standard deviations σ of $p = 3$ normally distributed features, for $c = 4$ different classes.

How should a feature vector $\boldsymbol{x} = [1.67, 2.00, 4.23]^\top$ be classified?

	Feature 1		Feature 2		Feature 3	
Class	μ	σ	μ	σ	μ	σ
0	1.6	0.1	2.4	0.5	4.3	0.2
1	1.5	0.2	2.9	0.6	6.1	0.9
2	1.8	0.3	2.5	0.3	4.2	0.3
3	1.1	0.2	3.1	0.7	5.6	0.3

Example: Naïve Bayes Classification

The posterior pdf is

$$g(y \,|\, \boldsymbol{\theta}, \boldsymbol{x}) \propto (\sigma_{y1}\sigma_{y2}\sigma_{y3})^{-1} \exp\left(-\frac{1}{2}\sum_{j=1}^{3}\frac{(x_j - \mu_{yj})^2}{\sigma_{yj}^2}\right),$$

where $\boldsymbol{\theta} := \{\boldsymbol{\sigma}_j, \boldsymbol{\mu}_j\}_{j=0}^{c-1}$ collects all model parameters.

The (unscaled) values for $g(y \,|\, \boldsymbol{\theta}, \boldsymbol{x})$, $y = 0, 1, 2, 3$ are 53.5, 0.24, 8.37, and 3.5×10^{-6}, respectively.

Hence, the feature vector should be classified as 0. The code follows.

naiveBayes.py

```python
import numpy as np
x = np.array([1.67,2,4.23]).reshape(1,3)
mu = np.array([1.6, 2.4, 4.3,
               1.5, 2.9, 6.1,
               1.8, 2.5, 4.2,
               1.1, 3.1, 5.6]).reshape(4,3)
sig = np.array([0.1, 0.5, 0.2,
                0.2, 0.6, 0.9,
                0.3, 0.3, 0.3,
                0.2, 0.7, 0.3]).reshape(4,3)
g = lambda y: 1/np.prod(sig[y,:]) * np.exp(
      -0.5*np.sum((x-mu[y,:])**2/sig[y,:]**2));
for y in range(0,4):
    print('{:3.2e}'.format(g(y)))
```

```
5.35e+01
2.42e-01
8.37e+00
3.53e-06
```

Discriminant Functions

The Bayesian viewpoint for classification leads in a natural way to the well-established technique of discriminant analysis.

We discuss the binary classification case first, with classes 0 and 1.

We consider a class of approximating functions $\mathcal{G}$ such that, conditional on the class $y \in \{0,1\}$, the feature vector $\boldsymbol{X} = [X_1, \ldots, X_p]^\top$ has a $\mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$ distribution, i.e.,

$$g(\boldsymbol{x} \,|\, \boldsymbol{\theta}, y) = \frac{1}{\sqrt{(2\pi)^p \, |\boldsymbol{\Sigma}_y|}} \mathrm{e}^{-\frac{1}{2}\, (\boldsymbol{x}-\boldsymbol{\mu}_y)^\top \boldsymbol{\Sigma}_y^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_y)}, \quad \boldsymbol{x} \in \mathbb{R}^p, \quad y \in \{0,1\}, \tag{4}$$

where $\boldsymbol{\theta} = \{\alpha_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}_{j=0}^{c-1}$ collects all model parameters, including the probability vector $\boldsymbol{\alpha}$ (that is, $\sum_i \alpha_i = 1$ and $\alpha_i \geqslant 0$) which helps define the prior density: $g(y \,|\, \boldsymbol{\theta}) = \alpha_y,\ y \in \{0,1\}$.

Discriminant Functions

Then, the posterior density is

$$g(y \,|\, \boldsymbol{\theta}, \boldsymbol{x}) \propto \alpha_y \times g(\boldsymbol{x} \,|\, \boldsymbol{\theta}, y),$$

and, according to the Bayes optimal decision rule (3), we classify $\boldsymbol{x}$ to come from class 0 if $\alpha_0 g(\boldsymbol{x} \,|\, \boldsymbol{\theta}, 0) > \alpha_1 g(\boldsymbol{x} \,|\, \boldsymbol{\theta}, 1)$.

Or, equivalently (by taking logarithms) if,

$$\ln \alpha_0 - \frac{1}{2} \ln |\boldsymbol{\Sigma}_0| - \frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_0) > \ln \alpha_1 - \frac{1}{2} \ln |\boldsymbol{\Sigma}_1| - \frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_1).$$

The function

$$\delta_y(\boldsymbol{x}) = \ln \alpha_y - \frac{1}{2} \ln |\boldsymbol{\Sigma}_y| - \frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_y)^\top \boldsymbol{\Sigma}_y^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_y), \quad \boldsymbol{x} \in \mathbb{R}^p$$

is called the quadratic discriminant function for class $y = 0, 1$.

A point $\boldsymbol{x}$ is classified to class y for which $\delta_y(\boldsymbol{x})$ is largest.