Machine Learning

Lecture 8: PE & Naive

杨思蓓

SIST

Email: yangsb@shanghaitech.edu.cn

Content

- Maximum-Likelihood Estimation
- Bayesian Estimation

So Far...

Bayesian framework

We could design an optimal classifier if we knew:

$P(y_i)$: priors

$P(x \mid y_i)$: class-conditional densities

Unfortunately, we rarely have this complete information!

Design a classifier based on a set of labeled training samples

Assume priors are known (or, estimate from the data)

Need sufficient no. of training samples for estimating class-conditional densities, especially when the dimensionality of the feature space is large

Parameter Estimation

Assumption about the problem: parametric model of $P(x \mid y_i)$ is available

Normality of $P(x \mid y_i)$

$$P(x \mid y_i) \sim N(\mu_i, \Sigma_i)$$

Characterized by 2 parameters

Estimation techniques

Maximum-Likelihood (ML) and Bayesian estimation

Results of the two procedures are nearly identical, but the approaches are different

Frequentist & Bayesian

Parameters in ML estimation are fixed but unknown!

MLE: Best parameters are obtained by maximizing the probability of obtaining the samples observed

Bayesian parameter estimation procedure, by its nature, utilizes whatever prior information is available about the unknown parameter

Bayesian methods view the parameters as random variables having some known prior distribution; How do we know the priors?

In either approach, we use $P(y_i \mid x)$ for our classification rule!

Maximum-Likelihood Estimation

Has good convergence properties as the sample size increases; estimated parameter value approaches the true value as n increases

Simpler than any other alternative technique

General principle

Assume we have c classes $D_1, \cdots D_c$

The samples are drawn according to $p(x|y_j)$, iid.

$$p(x|\mathrm{y}_j) \equiv p(x|y_j, \boldsymbol{\theta}_j)$$

$$p(x|y_j) \sim N(\boldsymbol{\mu}_j, \Sigma_j)$$
$$\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j, \Sigma_j)$$

Use class y_j samples to estimate class y_j parameters

Maximum-Likelihood Estimation

Use the information in training samples to estimate $\theta = (\theta_1, \theta_2, \ldots, \theta_c)$; θ_i ($i = 1, 2, \ldots, c$) is associated with the i-th category

Suppose sample set D contains n iid samples, $x_1, x_2, \ldots, x_n$

$$p(D|\theta) = \prod_{k=1}^{n} p(x_k|\theta)$$

$p(D|\theta)$ is called the likelihood of θ w.r.t. the set of samples.

ML estimate of θ is, by definition, the value θ that maximizes $p(D \mid \theta)$

"It is the value of θ that best agrees with the actually observed training samples"

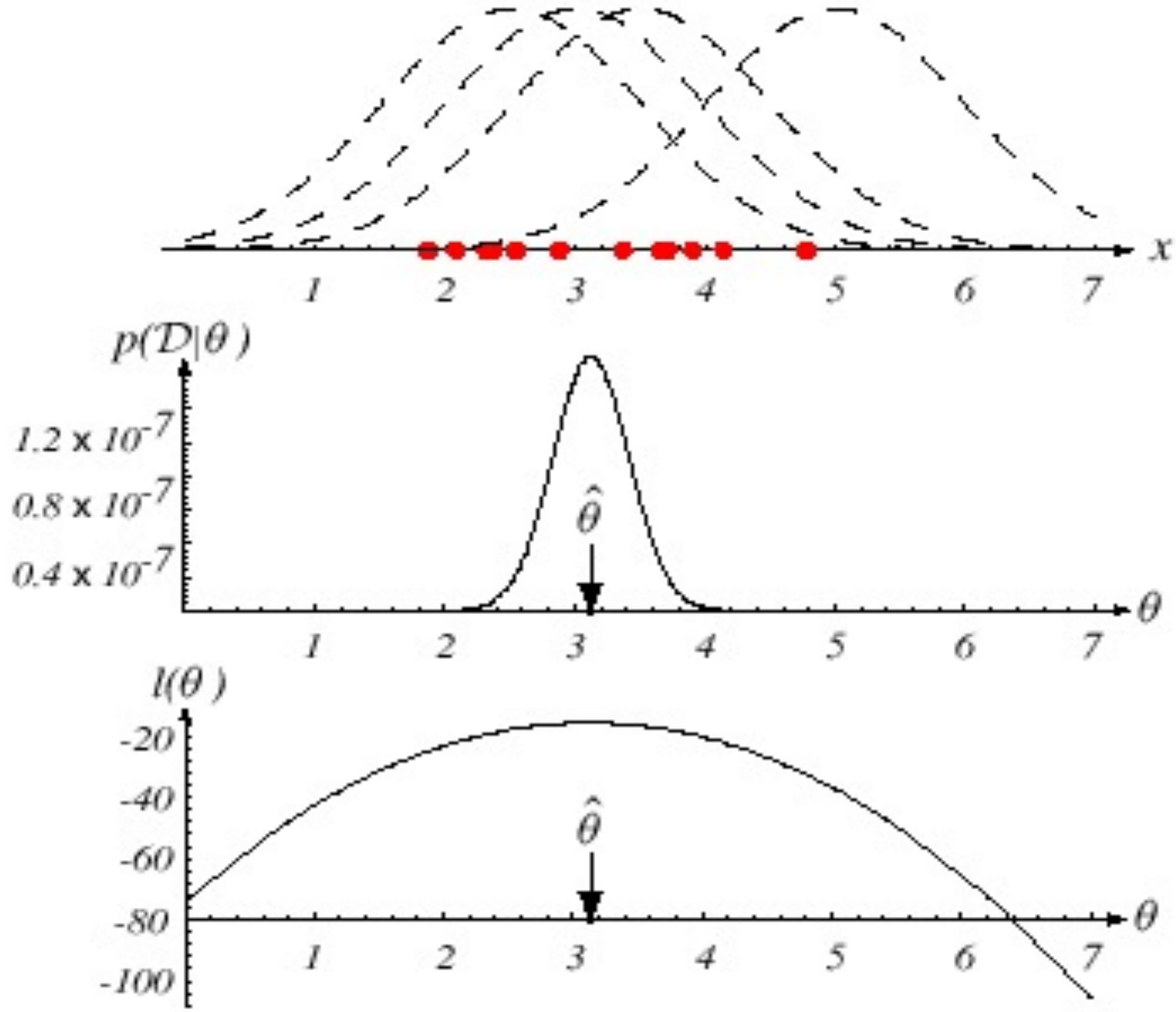

FIGURE 3.1. The top graph shows several training points in one dimension, known or assumed to be drawn from a Gaussian of a particular variance, but unknown mean. Four of the infinite number of candidate source distributions are shown in dashed lines. The middle figure shows the likelihood $p(\mathcal{D}|\theta)$ as a function of the mean. If we had a very large number of training points, this likelihood would be very narrow. The value that maximizes the likelihood is marked $\hat{\theta}$; it also maximizes the logarithm of the likelihood—that is, the log-likelihood $l(\theta)$, shown at the bottom. Note that even though they look similar, the likelihood $p(\mathcal{D}|\theta)$ is shown as a function of θ whereas the conditional density $p(x|\theta)$ is shown as a function of x. Furthermore, as a function of θ, the likelihood $p(\mathcal{D}|\theta)$ is not a probability density function and its area has no significance. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.

Optimal Estimation

We define $l(\theta)$ as the log-likelihood function

$$l(\theta) = \ln P(D \mid \theta)$$

New problem statement:

determine θ that maximizes the log-likelihood

$$\theta^* = arg\max_{\theta} l(\theta)$$

Let $\theta = (\theta_1, \theta_2, \ldots, \theta_p)^t$ and ∇_θ be the gradient operator

$$\nabla_\theta = \left[\frac{\partial}{\partial\theta_1}, \frac{\partial}{\partial\theta_2}, \cdots, \frac{\partial}{\partial\theta_p}\right]^T$$

Set of necessary conditions for an optimum is:

$$\nabla_\theta l = 0$$

$$\nabla_\theta l = \sum_{k=1}^{n} \nabla_\theta \ln P(x_k|\theta)$$

Example: Gaussian with unknown μ

P(x | μ) ~ N(μ, Σ)
(Samples are drawn from a multivariate normal population)

$$\ln P(x_k|\mu) = -\frac{1}{2}\ln[(2\pi)^d|\Sigma|] - \frac{1}{2}(\boldsymbol{x}_k - \boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}_k - \boldsymbol{\mu})$$

$$\nabla_\mu \ln P(x_k|\mu) = \Sigma^{-1}(\boldsymbol{x}_k - \boldsymbol{\mu})$$

therefore the ML estimate for μ must satisfy:

$$\sum_{k=1}^{n} \Sigma^{-1}(\boldsymbol{x}_k - \boldsymbol{\mu}) = 0$$

Example: Gaussian with unknown μ

Multiplying by Σ and rearranging, we obtain:

$$\boldsymbol{\mu}^* = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{x}_k$$

which is the arithmetic average or the mean of the samples of the training samples!

Example: Gaussian with unknown μ and Σ

Consider first the univariate case: $\theta = (\theta_1, \theta_2) = (\mu, \sigma^2)$

$$\ln p(x_k|\boldsymbol{\theta}) = -\frac{1}{2}\ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

$$\nabla_{\boldsymbol{\theta}} l = \nabla_{\boldsymbol{\theta}} \ln p(x_k|\boldsymbol{\theta}) = \begin{bmatrix} \frac{1}{\theta_2}(x_k - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x_k-\theta_1)^2}{2\theta_2^2} \end{bmatrix}$$

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})^2$$

Example: Gaussian with unknown μ and Σ

Multivariate case is basically very similar

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k$$

Sample covariance matrix

$$\widehat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\mu})(\mathbf{x}_k - \hat{\mu})^t$$

Bayesian Estimation

Bayesian learning approach for classification problems

In MLE, θ was supposed to have a fixed value

In BE, θ is a random variable

The computation of posterior probabilities $P(y_i \mid x)$ lies at the heart of Bayesian classification

To emphasize the training data: compute $P(y_i \mid x, D)$

Given the training sample set D, Bayes formula can be written

$$\boxed{P(\omega_i|\mathbf{x}, \mathcal{D}) = \frac{p(\mathbf{x}|\omega_i, \mathcal{D})P(\omega_i|\mathcal{D})}{\sum_{j=1}^{c} p(\mathbf{x}|\omega_j, \mathcal{D})P(\omega_j|\mathcal{D})}.}$$

We assume that the true values of the a priori probabilities are known or obtainable from a trivial calculation:

We substitute $P(\omega_i) = P(\omega_i|D)$

Furthermore, we can separate the training samples by class into c subsets $D_1, D_2, \cdots, D_c$, with the samples in D_i belonging to y_i

$$P(\omega_i|\mathbf{x}, \mathcal{D}) = \frac{p(\mathbf{x}|\omega_i, \mathcal{D}_i)P(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{x}|\omega_j, \mathcal{D}_j)P(\omega_j)}.$$

In essence, we have c separate problems of the following form: use a set D of samples drawn independently according to the fixed but unknown probability distribution $p(\boldsymbol{x})$ to determine

$$P(x|D)$$

Bayesian Parameter Estimation: Gaussian Case

Goal: Estimate θ using the a-posteriori density $P(\theta \mid D)$

The univariate Gaussian case: $P(\mu \mid D)$

μ is the only unknown parameter

$$P(x \mid \mu) \sim N(\mu, \sigma^2)$$

$$P(\mu) \sim N(\mu_0, \sigma_0^2)$$

μ_0 and σ_0 are known!

$$P(\mu \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \mu).P(\mu)}{\int P(\mathcal{D} \mid \mu).P(\mu)d\mu} \quad (1)$$

$$= \alpha \prod_{k=1}^{k=n} P(x_k \mid \mu).P(\mu)$$

Reproducing density

$$P(\mu \mid \mathcal{D}) \sim N(\mu_n, \sigma_n^2) \quad (2)$$

The updated parameters of the prior:

$$\mu_n = \left(\frac{n\sigma_0^2}{n_0\sigma_0^2 + \sigma^2}\right)\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}.\mu_0$$

$$\text{and } \sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}$$

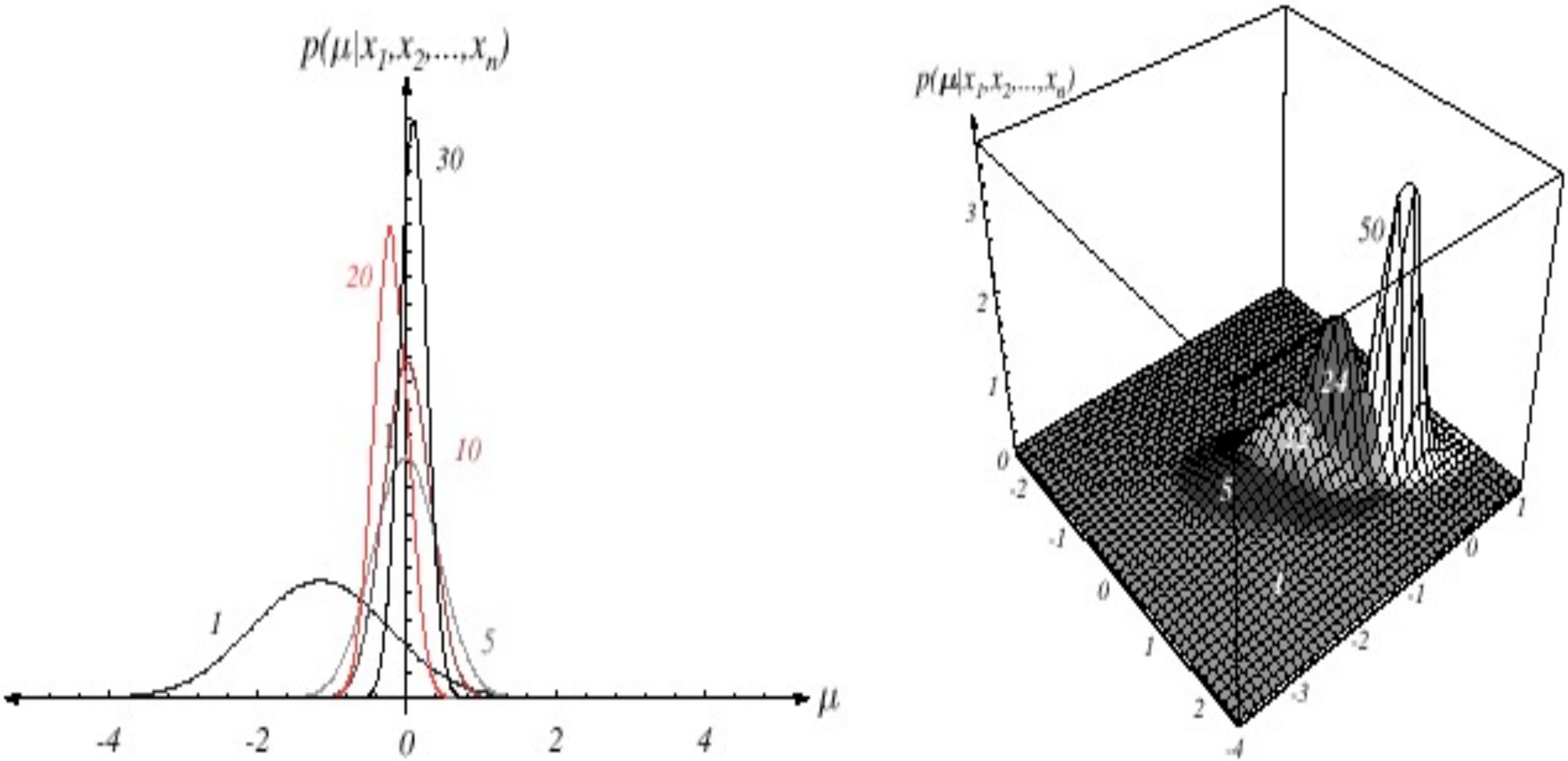

FIGURE 3.2. Bayesian learning of the mean of normal distributions in one and two dimensions. The posterior distribution estimates are labeled by the number of training samples used in the estimation. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

The univariate case $P(x \mid D)$

$P(\mu \mid D)$ has been computed

$P(x \mid D)$ remains to be computed!

$$\mathbf{P(x \mid D) = \int P(x \mid \mu).P(\mu \mid D)d\mu \ \text{ is Gaussian}}$$

It provides:

$$\mathbf{P(x \mid D) \sim N(\mu_n, \sigma^2 + \sigma_n^2)}$$

Desired class-conditional density $P(x \mid D_j, \omega_j)$

$P(x \mid D_j, \omega_j)$ together with $P(\omega_j)$ and using Bayes formula, we obtain the Bayesian classification rule:

$$\underset{\omega_j}{Max}\left[P(\omega_j \mid x, D)\right] \equiv \underset{\omega_j}{Max}\left[P(x \mid \omega_j, D).P(\omega_j)\right]$$

Bayesian Parameter Estimation: General Theory

$P(x \mid D)$ computation can be applied to any situation in which the unknown density can be parametrized: the basic assumptions are:

- The form of $P(x \mid \theta)$ is assumed known, but the value of θ is not known exactly
- Our knowledge about θ is assumed to be contained in a known prior density $P(\theta)$
- The rest of our knowledge about θ is contained in a set D of n random variables $x_1, x_2, \ldots, x_n$ that follows $P(x)$

The basic problem is:

"Compute the posterior density P(θ | D)"

then "Derive P(x | D)"

Using Bayes formula, we have:

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})\ d\boldsymbol{\theta}}$$

And by independence assumption:

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k|\boldsymbol{\theta})$$

Mammals vs. Non-mammals

Mammals

Non-mammals

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

Mammals vs. Non-mammals

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Naïve Bayes Classifier

Given $\boldsymbol{x} = \left(x_1, \cdots x_p\right)^T$

Goal is to predict class ω

Specifically, we want to find the value of ω that maximizes

$P(\omega|\boldsymbol{x}) = P\left(\omega|x_1, \cdots x_p\right)$

$$P\left(\omega|x_1, \cdots x_p\right) \propto P\left(x_1, \cdots x_p|\omega\right)P(\omega)$$

Independence assumption among features

$$P\left(x_1, \cdots x_p|\omega\right) = P(x_1|\omega) \cdots P\left(x_p|\omega\right)$$

How to Estimate Probabilities from Data?

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Class: $P(\omega_k) = \frac{N_{\omega_k}}{N}$

e.g., P(No) = 7/10,
P(Yes) = 3/10

For discrete attributes:

$$P(x_i|\omega_k) = \frac{|x_{ik}|}{N_{\omega_k}}$$

where $|x_{ik}|$ is number of instances having attribute x_i and belongs to class ω_k

Examples:

P(Status=Married|No) = 4/7

P(Refund=Yes|Yes)=0

How to Estimate Probabilities from Data?

For continuous attributes:

- Discretize the range into bins
  - one ordinal attribute per bin
  - violates independence assumption
- Two-way split: $(x < v)$ or $(x > v)$
  - choose only one of the two splits as new attribute
- Probability density estimation:
  - Assume attribute follows a normal distribution
  - Use data to estimate parameters of distribution (e.g., mean and standard deviation)
  - Once probability distribution is known, can use it to estimate the conditional probability $P(x_1|\omega)$

How to Estimate Probabilities from Data?

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Evade**
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Normal distribution:

$$P(x_i \mid \omega_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} \exp\left(-\frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2} \right)$$

One for each (x_i, ω_i) pair

For (Income, Class=No):

If Class=No

sample mean = 110

sample variance = 2975

$$P(Income = 120 \mid No) = \frac{1}{\sqrt{2\pi}(54.54)} \exp\left(-\frac{(120-110)^2}{2(2975)} \right) = 0.0072$$

Example of Naïve Bayes Classifier

Given a Test Record:

$$X = (\text{Refund} = \text{No}, \text{Married}, \text{Income} = 120\text{K})$$

naive Bayes Classifier:

P(Refund=Yes|No) = 3/7
P(Refund=No|No) = 4/7
P(Refund=Yes|Yes) = 0
P(Refund=No|Yes) = 1
P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No)=1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/7
P(Marital Status=Divorced|Yes)=1/7
P(Marital Status=Married|Yes) = 0

For taxable income:

If class=No:	sample mean=110
	sample variance=2975
If class=Yes:	sample mean=90
	sample variance=25

- P(X|Class=No) = P(Refund=No|Class=No)
  $\times$ P(Married| Class=No)
  $\times$ P(Income=120K| Class=No)
  = 4/7 $\times$ 4/7 $\times$ 0.0072 = 0.0024

- P(X|Class=Yes) = P(Refund=No| Class=Yes)
  $\times$ P(Married| Class=Yes)
  $\times$ P(Income=120K| Class=Yes)
  = 1 $\times$ 0 $\times$ 1.2 $\times$ 10^{-9} = 0

Since P(X|No)P(No) > P(X|Yes)P(Yes)

Therefore P(No|X) > P(Yes|X)

=> Class = No

Example of Naïve Bayes Classifier

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

A: attributes

M: mammals

N: non-mammals

$$P(A \mid M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A \mid N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A \mid M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A \mid N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

P(A|M)P(M) > P(A|N)P(N)

=> Mammals

Generalization

Bayesian Decision Theory

Naive Bayes Classifier

Naïve Bayes (Summary)

Robust to isolated noise points

Handle missing values by ignoring the instance during probability estimate calculations

Robust to irrelevant attributes

Independence assumption may not hold for some attributes

Smoothing

$$P(x_i|\omega_k) = \frac{|x_{ik}| + 1}{N_{\omega_k} + K}$$