Probabilistic Classifiers

Irina Rish

IBM T.J. Watson Research Center

rish@us.ibm.com, http://www.research.ibm.com/people/r/rish/

February 20, 2002

Probabilistic Classification

- Bayesian Decision Theory
  - Bayes decision rule (revisited):
    - Bayes risk, 0/1 loss, optimal classifier, discriminability
  - Probabilistic classifiers and their decision surfaces:
    - Continuous features (Gaussian distribution)
    - Discrete (binary) features (+ class-conditional independence)
- Parameter estimation
  - "Classical" statistics: maximum-likelihood (ML)
  - Bayesian statistics: maximum *a posteriori* (MAP)
- Common used classifier: naïve Bayes
  - VERY simple: class-conditional feature independence
  - VERY efficient (empirically); why and when? – still an open question

Bayesian decision theory

- **Make a decision that minimizes the overall expected cost (loss)**
  - Advantage: theoretically guaranteed optimal decisions
  - Drawback: probability distributions are assumed to be known (in practice, estimation of those distribution from data can be a hard problem)
- Classification problem as an example of a decision problem
  - Given observed properties (features) of an object, find its class. Examples:
    - Sorting fish by its type (sea bass or salmon) given observed features such as lightness and length
    - Video character recognition
    - Face recognition
    - Document classification using word counts
    - Guessing user's intentions (potential buyer or not) by his web transactions
    - Intrusion detection

Notation and definitions

- C - a **state of nature (class):** a random variable with distribution $\mathrm{P}(C)$
  - $\Omega = \{\omega_1, ..., \omega_n\}$ is a set of possible states of nature (class labels)
- $\mathrm{X} = (X_1, ..., X_d)$ - **feature vector** in feature space S
  - Continuous X_i : $S = \Re^\mathrm{d}$ and $\mathbf{X}$ has **probability density** $\mathrm{p}(\mathbf{X} \mid C)$
  - Discrete X_i : $S = D^d$, $D = \{1, ..., k\}$ and $\mathbf{X}$ has **probability** $\mathrm{P}(\mathbf{X} \mid C)$
- $\alpha(\mathrm{x}) : S \rightarrow A$ - **decision rule**
  - $A = \{\alpha_1, ..., \alpha_a\}$ is a set of **actions (decisions)**
  - Example: $\mathrm{A} = \{\omega_1, ..., \omega_n\}$ and $\alpha(\mathrm{x}) : S \rightarrow \Omega$ is a classifier
- $\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$ - **loss function** (cost of decision α_i given state ω_j)
  - Example: **0/1 loss** ($\lambda_{ij} = 1$ if $i \neq j$ and $\lambda_{ij} = 0$ if $i = j$)
- $\mathrm{R}(\alpha_i \mid \mathrm{x}) = \sum_{j=1}^{n} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid \mathrm{x})$ - **conditional risk** of action α_i given x
- $R = \int R(\alpha(\mathrm{x}) \mid \mathrm{x}) p(\mathrm{x}) d\mathrm{x}$ - **risk (total expected loss)**

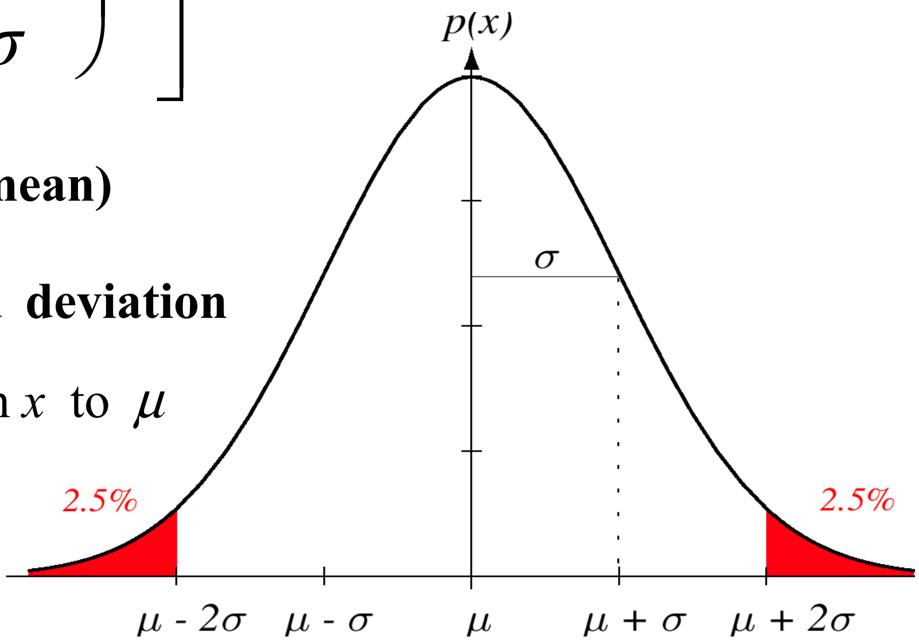
Gaussian (normal) density $N(\mu,\sigma)$

$$p(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

$\mu = \mathrm{E}[x] = \int_{-\infty}^{\infty} xp(x)dx -$ **expected value (mean)**

$\sigma^2 = \mathrm{E}[(x-\mu)^2] -$ **variance**, $\sigma -$ **standard deviation**

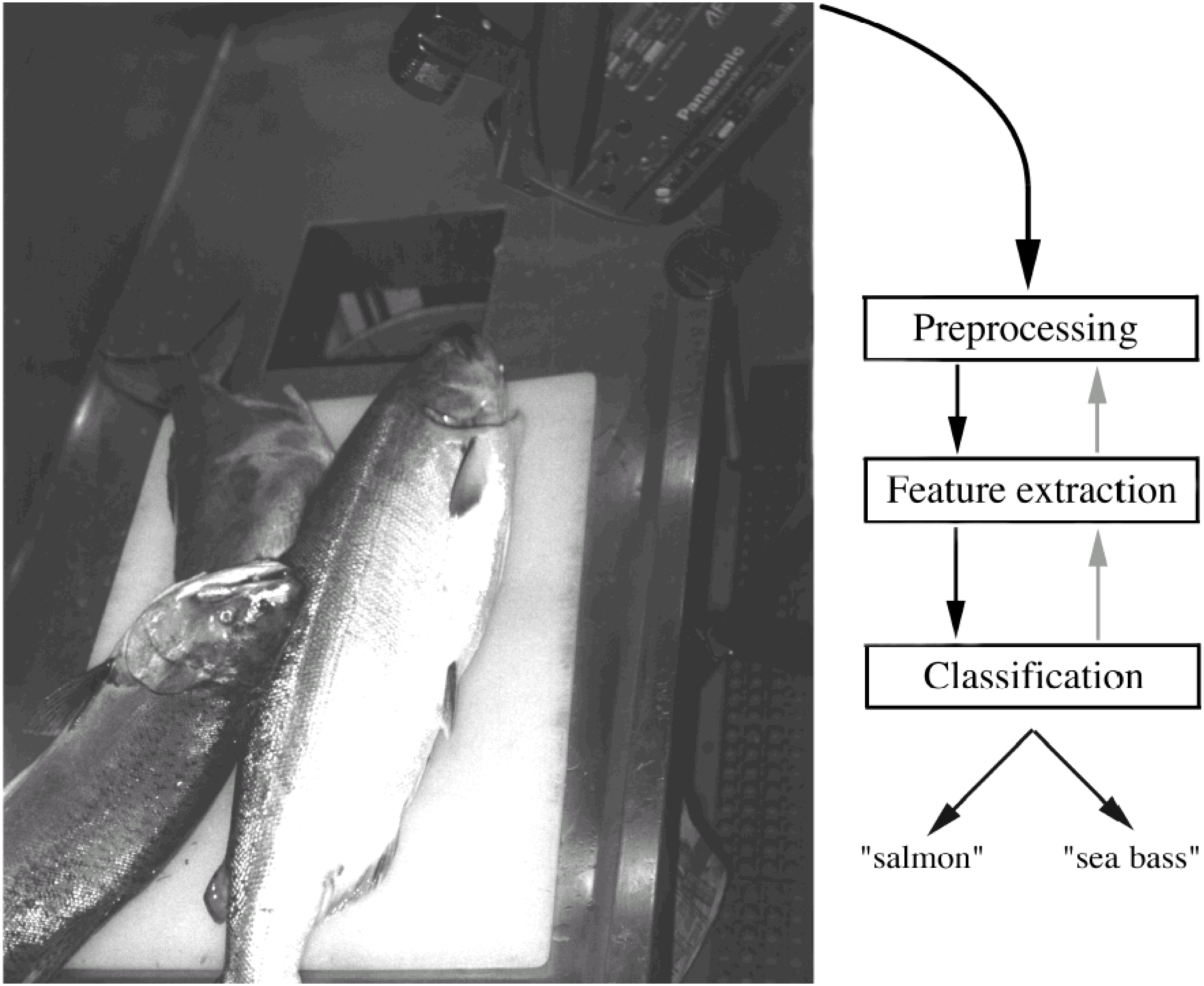
$r = \frac{|x-\mu|}{\sigma} -$ **Mahalanobis distance** from x to μ



Interesting property (homework problem:) $\mathrm{N}(\mu,\sigma)$ has **maximum entropy** $H(p(x))$ among all p(x) with given mean and variance!

$$H(p(x)) = -\int p(x)\ln p(x)dx \text{ (in nats)}$$

When learning from data, max - entropy distributi ons are most reasonable since they impose 'no additional structure' besides what is given as constaints



Preprocessing

Feature extraction

Classification

"salmon"

"sea bass"

Bayes rule for binary classification

- Given only **priors** $P(C=\omega_i)=P(\omega_i)$,

  choose $\boldsymbol{g^*(x)}=\omega_1$ *if* $P(\omega_1)>P(\omega_2)$, $\boldsymbol{g^*(x)}=\omega_2$ otherwise

- Given **evidence x,** update $P(\omega_i)$ using

  Bayes formula : $P(\omega_i \mid \mathrm{x})=\dfrac{p(\mathrm{x}\mid\omega_i)P(\omega_i)}{p(\mathrm{x})}$

  where $p(x)=\sum_{j=1}^{n} p(\mathrm{x}\mid\omega_i)P(\omega_i)$ - **evidence** probability,

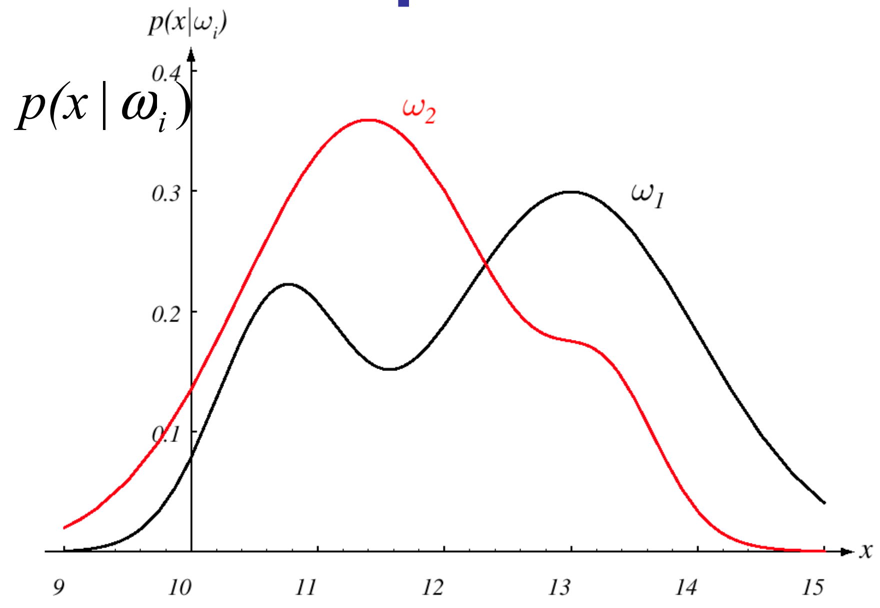
  $p(\mathrm{x}\mid\omega_i)$ - **likelihood**, $P(\omega_i)$ – **prior**

Bayes decision rule :

choose $\boldsymbol{g^*(x)}=\omega_1$ if $\mathrm{P}(\omega_1\mid x)>P(\omega_2\mid x)$,

$\boldsymbol{g^*(x)}=\omega_2$ otherwise

Example: one-dimensional case



Since $p(\mathrm{x})$ does not depend on ω_i in

$$P(\omega_i\,|\,\mathrm{x}) = \frac{p(\mathrm{x}\,|\,\omega_i)P(\omega_i)}{p(\mathrm{x})},$$ we get

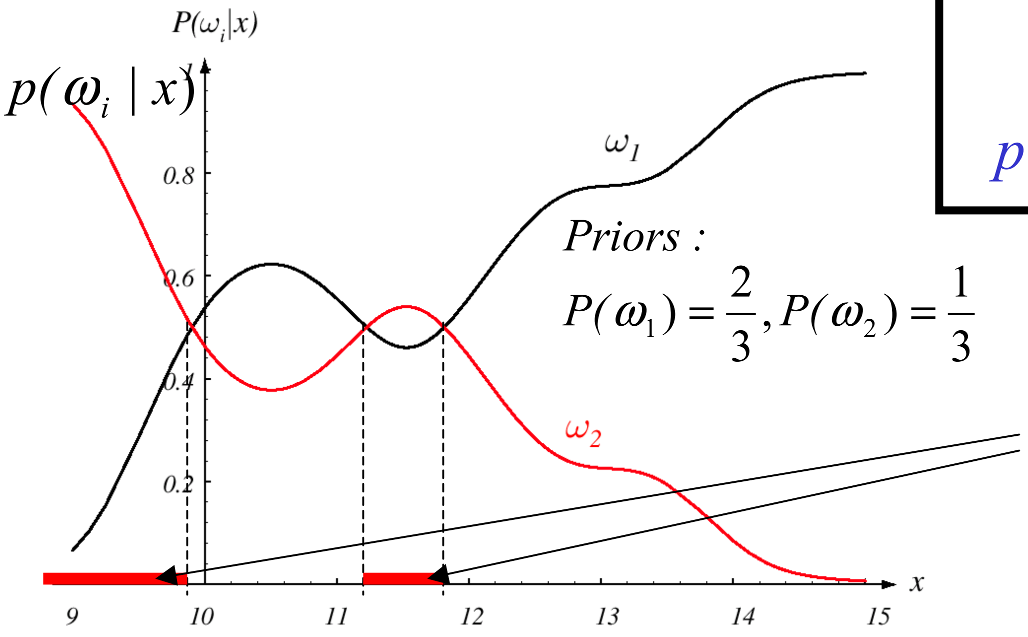
Bayes decision rule:

$$g^*(x) = \omega_1 \;\; if$$

$$P(\omega_1|x) > P(\omega_2|x)$$

♍

$$p(x\,|\,\omega_1)P(\omega_1) > p(x\,|\,\omega_2)P(\omega_2)$$



Priors :

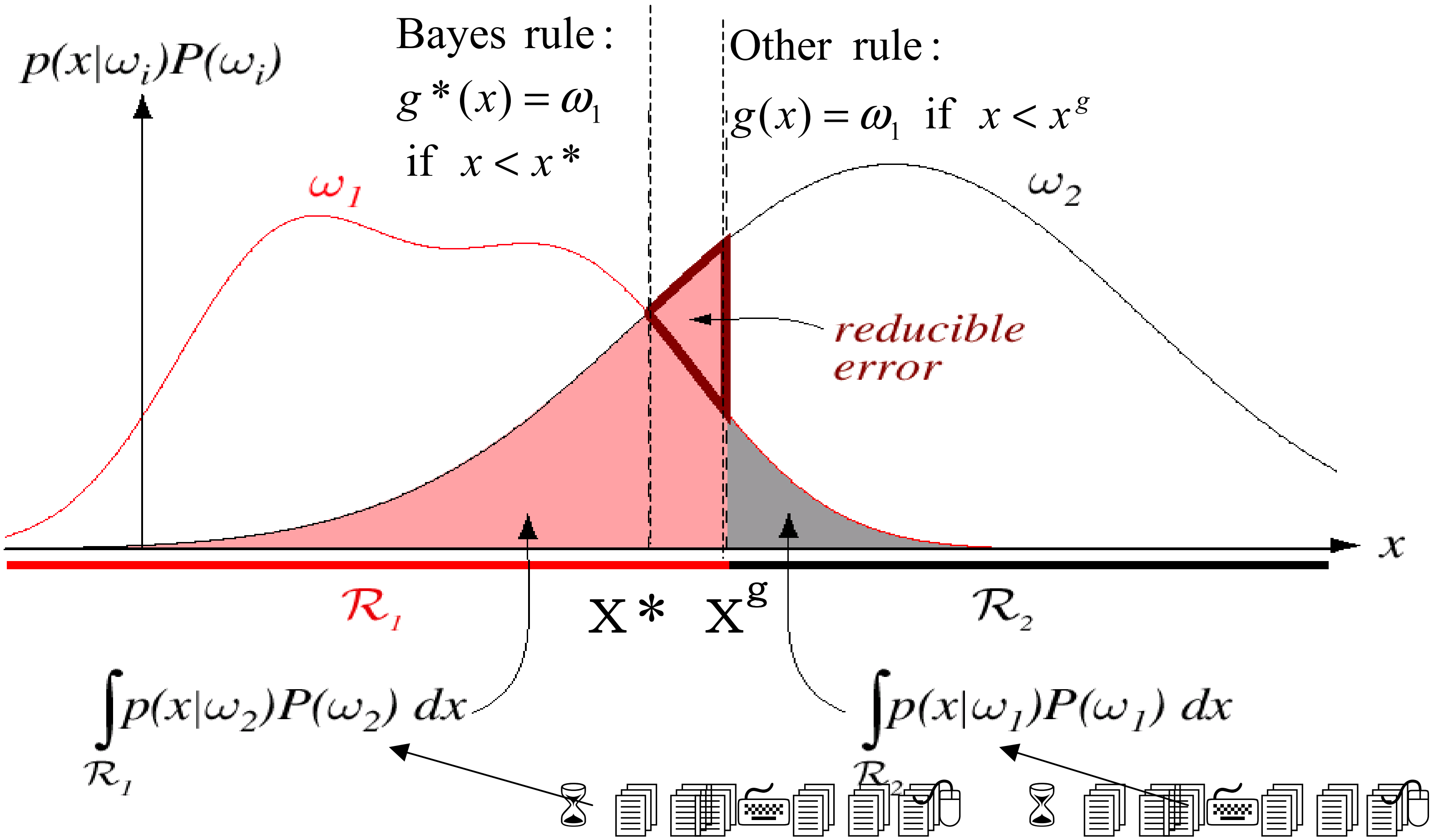
$$P(\omega_1) = \frac{2}{3}, P(\omega_2) = \frac{1}{3}$$

Decison regions :

$$R_1^* = \{\mathrm{x}\,|\;\; \mathrm{P}(\omega_1\,|\,\mathrm{x}) > P(\omega_2\,|\,\mathrm{x})\}$$

$$R_2^* = \{\mathrm{x}\,|\;\; \mathrm{P}(\omega_1\,|\,\mathrm{x}) \le P(\omega_2\,|\,\mathrm{x})\}$$

Optimality of Bayes rule: idea



$$P_g(error|x) = P(g(x)\neq\omega|x) = P(g(x)=\omega_1,\omega_2|x) + P(g(x)=\omega_2,\omega_1|x)$$

$$\boxed{P_{g*}(error|x) \le P_g(error|x) \quad \text{for any } g(x): S\to\Omega}$$

Proof – see next page…

Optimality of Bayes rule: proof

$$P_{g*}(error|\mathrm{x}) = P(g*(\mathrm{x}) \neq \omega|\mathrm{x}) = P(g*(\mathrm{x}) = \omega_1, \omega_2|\mathrm{x}) + P(g*(\mathrm{x}) = \omega_2, \omega_1|\mathrm{x}) =$$

$$= \underbrace{P(g*(\mathrm{x}) = \omega_1 | \omega_2, \mathrm{x})}_{p_1} P(\omega_2|\mathrm{x}) + \underbrace{P(g*(\mathrm{x}) = \omega_2 | \omega_1, \mathrm{x})}_{p_2} P(\omega_1|\mathrm{x})$$

1) $if\ \ P(\omega_1|x) > P(\omega_2|x),\ \text{then}\ \ g*(\mathrm{x}) = \omega_1\ \ \text{and}\ p_1 = 1,\ \ p_2 = 0 \Rightarrow$

$$P_{g*}(error|\mathrm{x}) = P(\omega_2|\mathrm{x})$$

2) $if\ \ P(\omega_1|x) \leq P(\omega_2|x),\ \text{then}\ \ g*(\mathrm{x}) = \omega_2\ \ \text{and}\ p_1 = 0,\ \ p_2 = 1 \Rightarrow$

$$P_{g*}(error|\mathrm{x}) = P(\omega_1|\mathrm{x})$$

Thus, $P_{g*}(error|\mathrm{x}) = \min\{P(\omega_1|\mathrm{x}), P(\omega_2|\mathrm{x})\}$, and, for any $\mathrm{g}(\mathbf{x}) : \mathrm{S} \rightarrow \Omega$

$$P_{g*}(error) = \int_{-\infty}^{+\infty} P_{g*}(error\,|\,x)\,p(x)dx \leq P_g(error)$$

General Bayesian Decision Theory

Given :
- set of available **actions (decisions)** $A = \{\alpha_1, \dots, \alpha_a\}$
- **loss function** (cost of decision α_i given ω_j) $\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$

Find : a **decision rule** $\alpha(\mathrm{x}) : S \rightarrow A$ minimizing

the **total expected loss (risk)** : $R = \int R(\alpha(\mathrm{x}) \mid \mathrm{x})\, p(\mathrm{x}) d\mathrm{x},$

where $\mathrm{R}(\alpha_i \mid \mathrm{x}) = \sum_{j=1}^{n} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid \mathrm{x})$ is the **conditional risk**

of action α_i given x and $P(\omega_i \mid x) = \dfrac{p(x \mid \omega_i) P(\omega_i)}{p(x)}$ (Bayes formula)

- ***Bayes decision rule :*** always minimize conditional risk

$$\boldsymbol{a}^*(\mathbf{x}) = \arg\min_{\alpha_i} \mathrm{R}(\alpha_i \mid \mathrm{x}) = \arg\min_{\alpha_i} \sum_{j=1}^{n} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid \mathrm{x})$$

- ***Bayes decision rule*** yields minimum overall risk (called **Bayes risk**) :

$$R^* = \min_{\alpha(\mathrm{x})} \int R(\alpha(\mathrm{x}) \mid \mathrm{x})\, p(\mathrm{x}) d\mathrm{x}$$

Zero-one loss classification

Let $\alpha(\mathrm{x}) = \mathrm{g}(\mathbf{x})$ (action = classification), i.e. $\alpha_i = \omega_i$, $\mathrm{i} = 1,\dots,\mathrm{n}$

$$\boxed{\textbf{Zero - one loss :}\ \lambda_{ij} = \lambda(\alpha_i \mid \omega_j) = \begin{cases} 1 \text{ if } i \neq j \\ 0 \text{ if } i = j \end{cases}}$$

Then **conditional risk = classification error**

$$R(\alpha_i \mid \mathrm{x}) = \sum_{j=1}^{n} \lambda(\alpha_j \mid \omega_j) P(\omega_j \mid \mathrm{x}) = \sum_{j \neq i} P(\omega_j \mid \mathrm{x}) = 1 - P(\omega_i \mid \mathrm{x}) = P(error \mid \mathrm{x})$$

Bayes rule:

$g^*(\mathrm{x}) = \omega_i$ if

$\mathrm{P}(\omega_i \mid \mathrm{x}) > \mathrm{P}(\omega_j \mid \mathrm{x})$ for all $\mathrm{i} \neq \mathrm{j}$

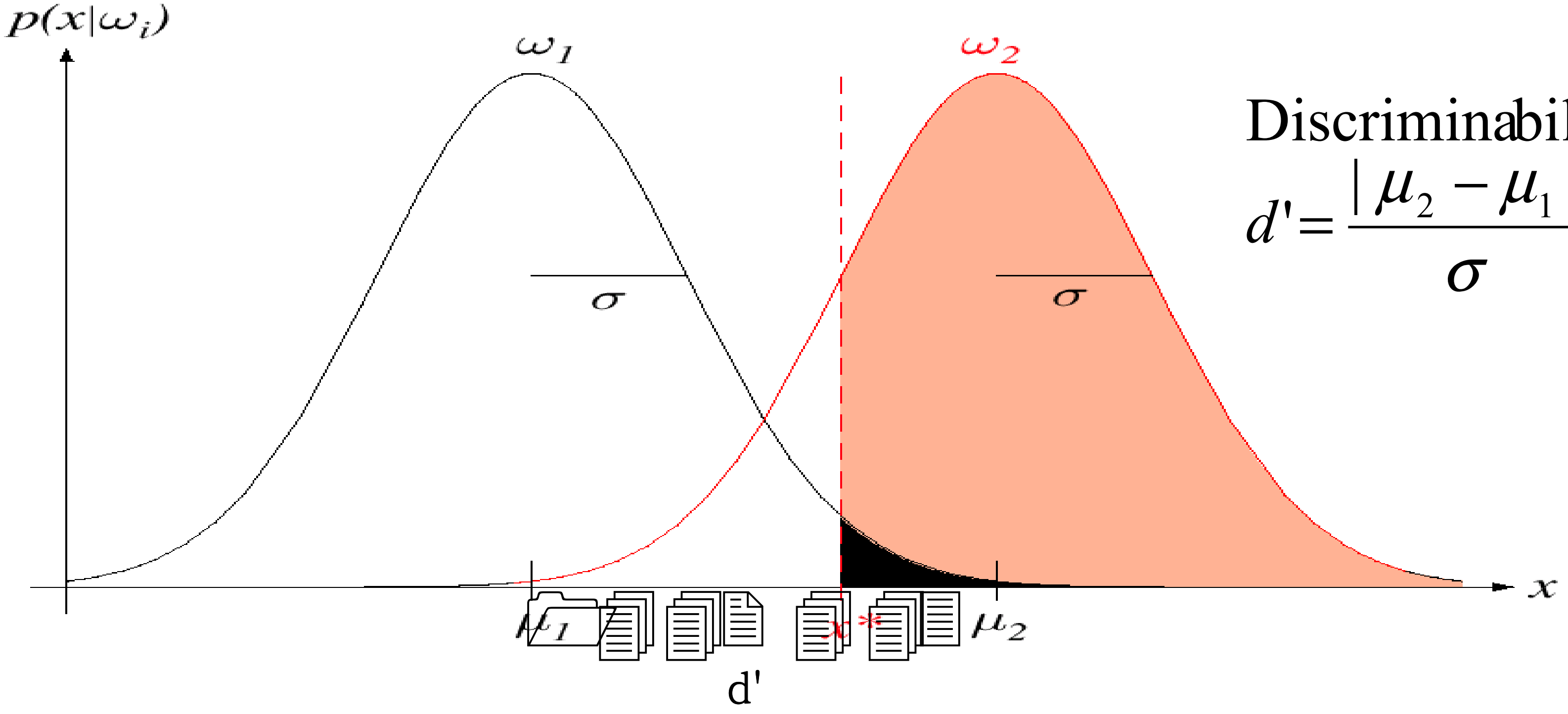
Bayes rule achieves minimum error rate $R^* = \min_{\alpha(\mathrm{x})} \int R(\alpha(\mathrm{x}) \mid \mathrm{x})\, p(\mathrm{x}) d\mathrm{x}$

Different errors, different costs

Example from signal detection theory :

ω_1 - no signal (background noise), ω_2 – signal present

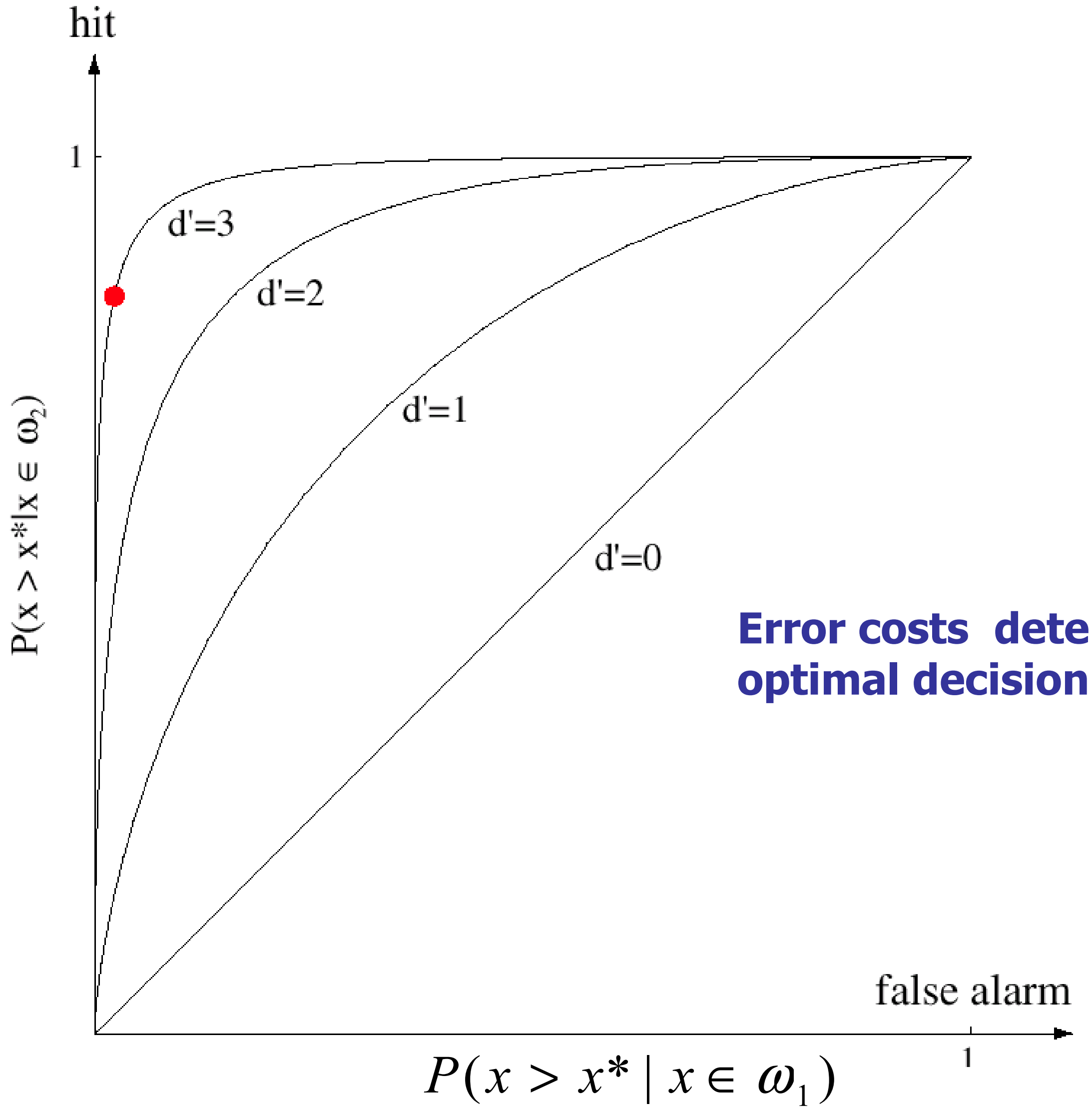
- Hit : $P(x > x^* \mid x \in \omega_2)$ (cost λ_{22})
- Miss : $P(x < x^* \mid x \in \omega_2)$ (cost λ_{12})
- False alarm : $P(x > x^* \mid x \in \omega_1)$ (cost λ_{21})
- Correct rejection : $P(x < x^* \mid x \in \omega_1)$ (cost λ_{11})



Discriminability

$$d' = \frac{|\mu_2 - \mu_1|}{\sigma}$$

Receiver operating characteristic (ROC) curve



Error costs determine optimal decision (threshold x*)

Cost-based classification

Let $\alpha_i = \omega_i, \mathrm{i} = 1,2,$ and $\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$:

$$\begin{cases} \mathrm{R}(\alpha_1 \mid \mathrm{x}) = \lambda_{11} P(\omega_1 \mid \mathrm{x}) + \lambda_{12} P(\omega_2 \mid \mathrm{x}) \\ \mathrm{R}(\alpha_2 \mid \mathrm{x}) = \lambda_{21} P(\omega_1 \mid \mathrm{x}) + \lambda_{22} P(\omega_2 \mid \mathrm{x}) \end{cases}$$

Bayes rule:

$$g^*(\mathrm{x}) = \omega_1 \text{ if}$$

$$\mathrm{R}(\alpha_1 \mid \mathrm{x}) < \mathrm{R}(\alpha_2 \mid \mathrm{x}), \text{ i.e.}$$

$$(\lambda_{21} - \lambda_{11}) P(\omega_1 \mid \mathrm{x}) > (\lambda_{12} - \lambda_{22}) P(\omega_2 \mid \mathrm{x})$$

Assuming $\lambda_{21} > \lambda_{11}$ and $\lambda_{12} > \lambda_{22}$

(errors cost more than correct decision) :

$$\frac{P(\omega_1 \mid \mathrm{x})}{P(\omega_2 \mid \mathrm{x})} > \frac{(\lambda_{12} - \lambda_{22})}{(\lambda_{21} - \lambda_{11})}, \text{ or}$$

Bayes rule :

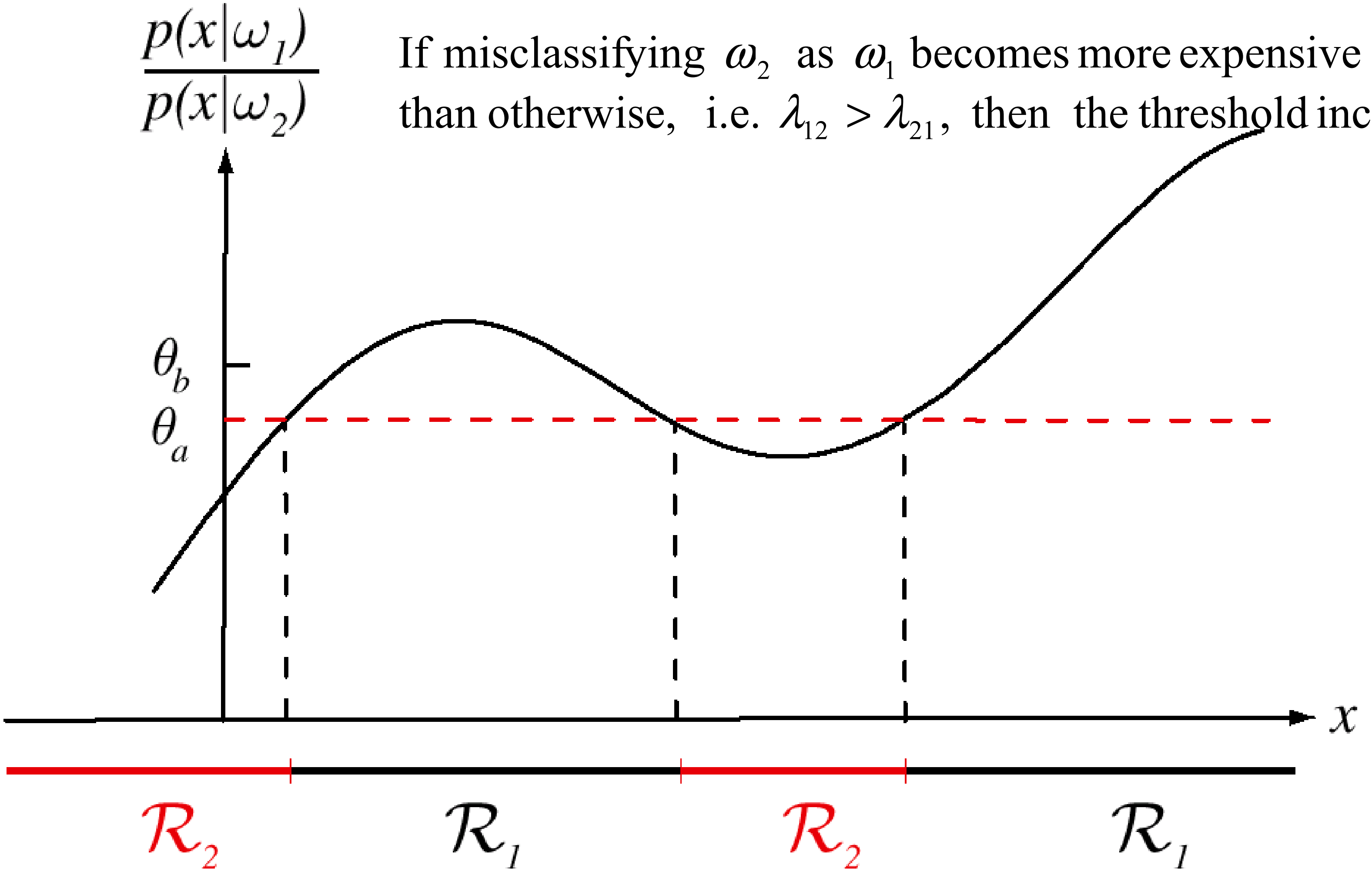
$$g^*(\mathrm{x}) = \omega_1 \text{ if}$$

$$\frac{P(\mathrm{x} \mid \omega_1)}{P(\mathrm{x} \mid \omega_2)} > \frac{(\lambda_{12} - \lambda_{22})}{(\lambda_{21} - \lambda_{11})} \frac{\mathrm{P}(\omega_2)}{\mathrm{P}(\omega_1)} = \theta_\lambda$$

Example: one-dimensional case

$$\frac{P(\mathrm{x}\,|\,\omega_1)}{P(\mathrm{x}\,|\,\omega_2)} > \frac{(\lambda_{12}-\lambda_{22})}{(\lambda_{21}-\lambda_{11})}\,\frac{\mathrm{P}(\omega_2)}{\mathrm{P}(\omega_1)} = \theta_\lambda$$

If misclassifying ω_2 as ω_1 becomes more expensive than otherwise, i.e. $\lambda_{12} > \lambda_{21}$, then the threshold increases

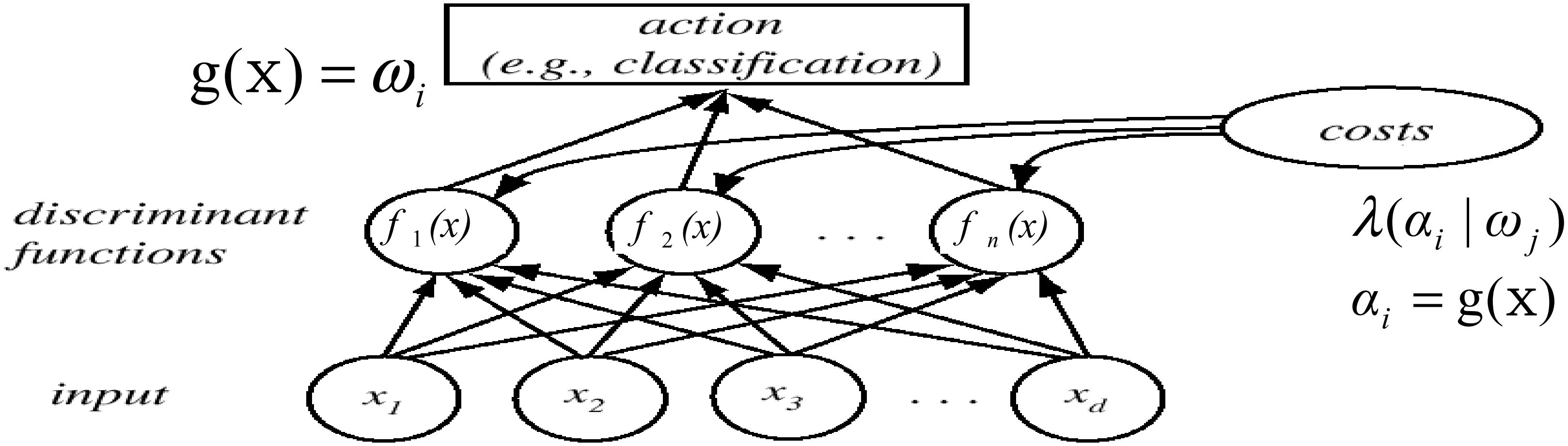


Discriminant functions and classification

- **Multi - category classification :**
  - *discriminant functions* : $f_i(\mathrm{x}),\ \ i = 1,\ldots,n$
    (e.g., $f_i^*(\mathrm{x}) = -R(\alpha_i \mid \mathrm{x})$ for Bayes classifier)
  - *classifier* : $\mathrm{g}(\mathrm{x}) = \omega_i$ if $f_i(\mathrm{x}) > f_j(\mathrm{x})$ for all $i \neq j$

- **Two - category classification :**
  - *single discriminant function* : $f(\mathrm{x}) \equiv f_1(\mathrm{x}) - f_2(\mathrm{x})$
  - *classifier* : $\mathrm{g}(\mathrm{x}) = \omega_1$ if $f(\mathrm{x}) > 0$

$\mathrm{g}(\mathrm{x}) = \omega_i$



$\lambda(\alpha_i \mid \omega_j)$

$\alpha_i = \mathrm{g}(\mathrm{x})$

Bayes Discriminant Functions

Bayes discriminant functions for zero - one loss classification :

$$f_i^*(\mathrm{x}) = -R(\alpha_i \mid \mathrm{x}) \;= \mathrm{P}(\omega_\mathrm{i} \mid \mathrm{x}) - 1$$

Every $f_i(\mathrm{x})$ can be replaced by $\mathrm{h}(f_i(\mathrm{x}))$ where $\mathrm{h}(\cdot)$ is monotonically increasing !

Multi-category:

$$f_i^*(\mathrm{x}) = P(\omega_i \mid \mathrm{x}) = \frac{p(\mathrm{x} \mid \omega_i)P(\omega_i)}{p(\mathrm{x})}$$

$$f_i^*(\mathrm{x}) = p(\mathrm{x} \mid \omega_i)P(\omega_i)$$

$$f_i^*(\mathrm{x}) = \log p(\mathrm{x} \mid \omega_i) + \log P(\omega_i)$$

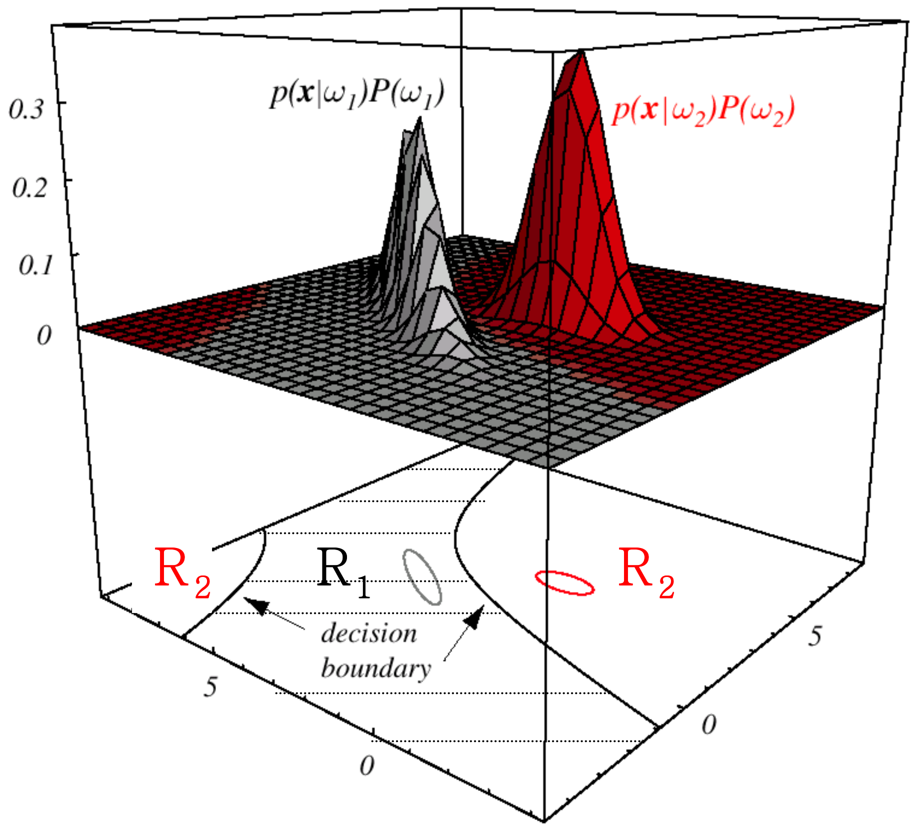
Two-category:

$$f^*(\mathrm{x}) \equiv f_1^*(\mathrm{x}) - f_1^*(\mathrm{x})$$

$$f^*(\mathrm{x}) = P(\omega_1 \mid \mathrm{x}) - P(\omega_2 \mid \mathrm{x})$$

$$f^*(\mathrm{x}) = \log \frac{p(\mathrm{x} \mid \omega_1)}{p(\mathrm{x} \mid \omega_2)} + \log \frac{P(\omega_1)}{P(\omega_2)}$$

Decision regions and surfaces:



Discriminant functions: examples

	Features	Discriminant functions
Discrete	**Binary, conditionally independent** $P(x_i, x_j \mid C) = P(x_i \mid C)P(x_j \mid C)$	linear (hyperplanes)
Continuous	Multivariate Gaussian $p(\mathrm{x} \mid \omega_i) =$ $\frac{1}{(2\pi)^{d/2} \mid \Sigma_\mathrm{i} \mid^{1/2}} \exp\left[-\frac{1}{2}(\mathrm{x} - \mu_\mathrm{i})^t \Sigma_\mathrm{i}^{-1} (\mathrm{x} - \mu_\mathrm{i}) \right]$	
	• **Same covariance matrix :** $\Sigma_\mathrm{i} = \Sigma$ for all classes ω_i	linear (hyperplanes)
	• **General case :** arbitrary Σ_i	hyperquadrics (hyperellipsoids, hyperparabaloids, hyperhyperboloids)

Conditionally independent binary features ⇔ linear classifier (‘naïve Bayes’-see later)

$$x=(x_1,\ldots,x_n),\, x_i \in \{0,1\} - \mathit{features},\quad c \in \{\omega_1,\omega_2\} - \mathit{class}$$

$$\mathit{Let}\ p_i = P(x_i = 1 \mid \omega_1),\quad q_i = P(x_i = 1 \mid \omega_2).\quad \mathit{Then}$$

$$P(x \mid \omega_1) = \prod_{i=1}^{n} p_i^{x_i}(1-p_i)^{1-x_i},\qquad P(x \mid \omega_2) = \prod_{i=1}^{n} q_i^{x_i}(1-q_i)^{1-x_i}$$

Discriminant function f(x) (decision rule: if f(x) > 0, choose ω_1):

$$f(x) = \log\frac{P(\omega_1 \mid x)}{P(\omega_2 \mid x)} = \log\frac{P(x \mid \omega_1)P(\omega_1)}{P(x \mid \omega_2)P(\omega_2)} = \log\prod_{i=1}^{n}\left(\frac{p_i}{q_i}\right)^{x_i}\left(\frac{1-p_i}{1-q_i}\right)^{1-x_i}\left(\frac{P(\omega_1)}{P(\omega_2)}\right) =$$

$$= \sum_{i=1}^{n}\left(x_i\log\frac{p_i}{q_i} + (1-x_i)\log\frac{1-p_i}{1-q_i}\right) + \log\left(\frac{P(\omega_1)}{P(\omega_2)}\right) =$$

$$f(x) = \sum_{i=1}^{n} w_i x_i + w_0,\ \text{where}\ w_i = \log\frac{p_i(1-q_i)}{q_i(1-p_i)},\ w_0 = \sum_{i=1}^{n}\log\frac{1-p_i}{1-q_i} + \log\frac{P(\omega_1)}{P(\omega_2)}$$

Case1: independent Gaussian features with same variance for all classes: $\Sigma_i = \sigma^2 I$

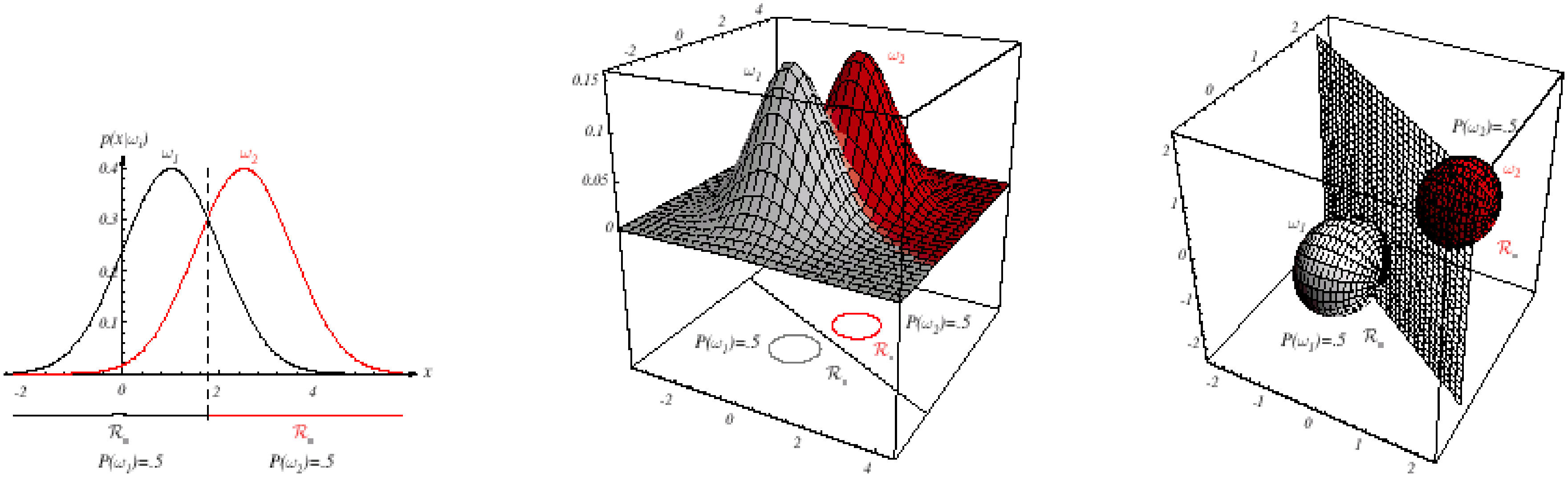


FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d - 1$ dimensions, perpendicular to the line separating the means. In these one-, two-, and three-dimensional examples, we indicate $p(\mathbf{x}|\omega_i)$ and the boundaries for the case $P(\omega_1) = P(\omega_2)$. In the three-dimensional case, the grid plane separates $\mathcal{R}_1$ from $\mathcal{R}_2$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Note: linear separating surfaces!

Case 2 : generalization to dependent features having same covariances for all classes : $\Sigma_i = \Sigma$

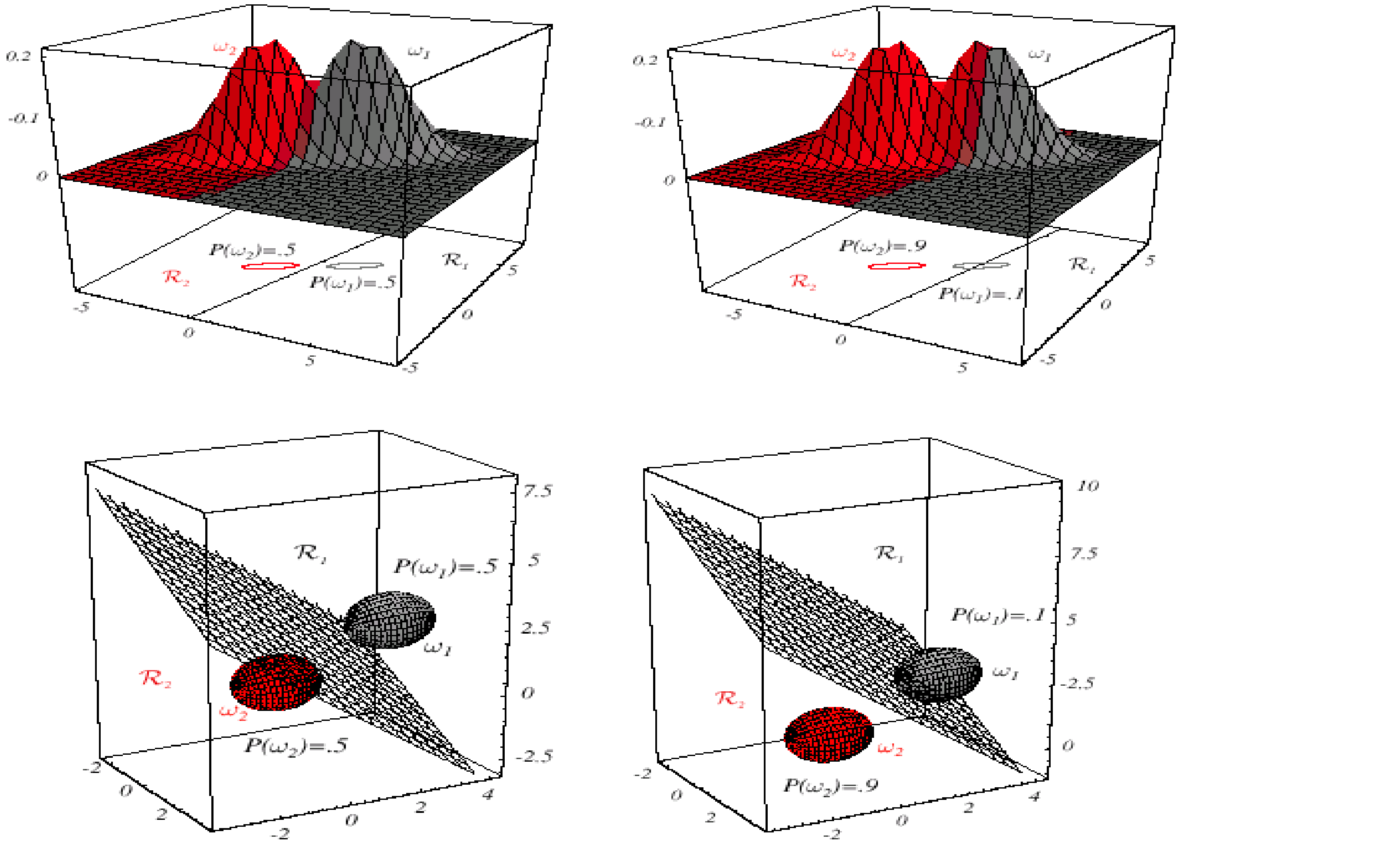


FIGURE 2.12. Probability densities (indicated by the surfaces in two dimensions and ellipsoidal surfaces in three dimensions) and decision regions for equal but asymmetric Gaussian distributions. The decision hyperplanes need not be perpendicular to the line connecting the means. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Case 3: unequal covariance matrices
One dimension: multiply-connected decision regions

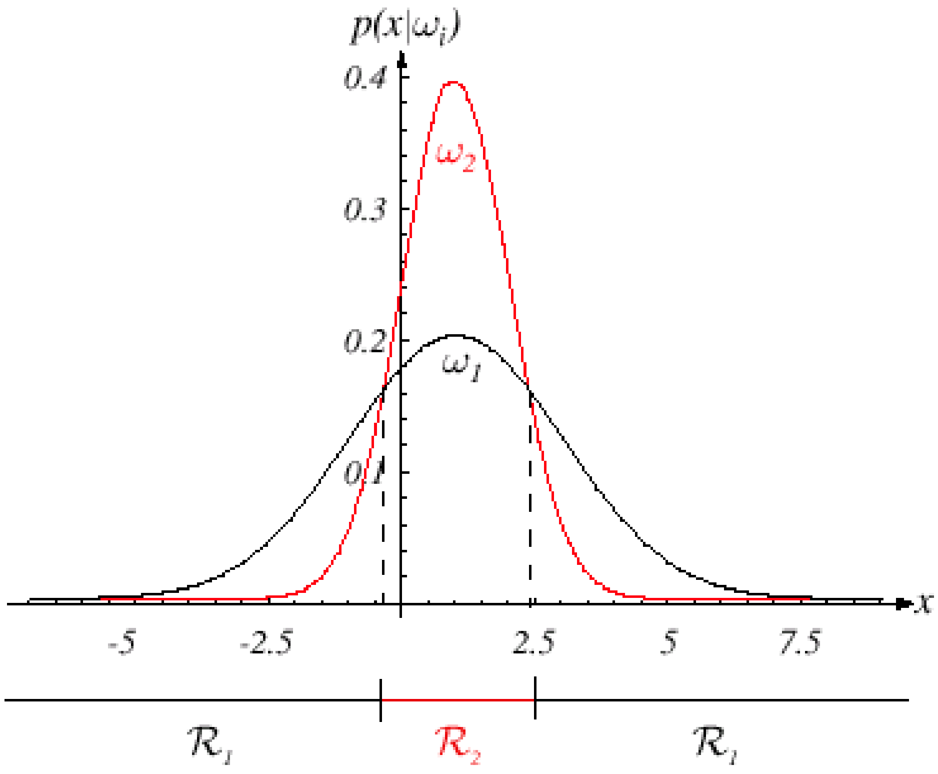


FIGURE 2.13. Non-simply connected decision regions can arise in one dimensions for Gaussians having unequal variance. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Case 3, many dimensions: hyperquadric surfaces

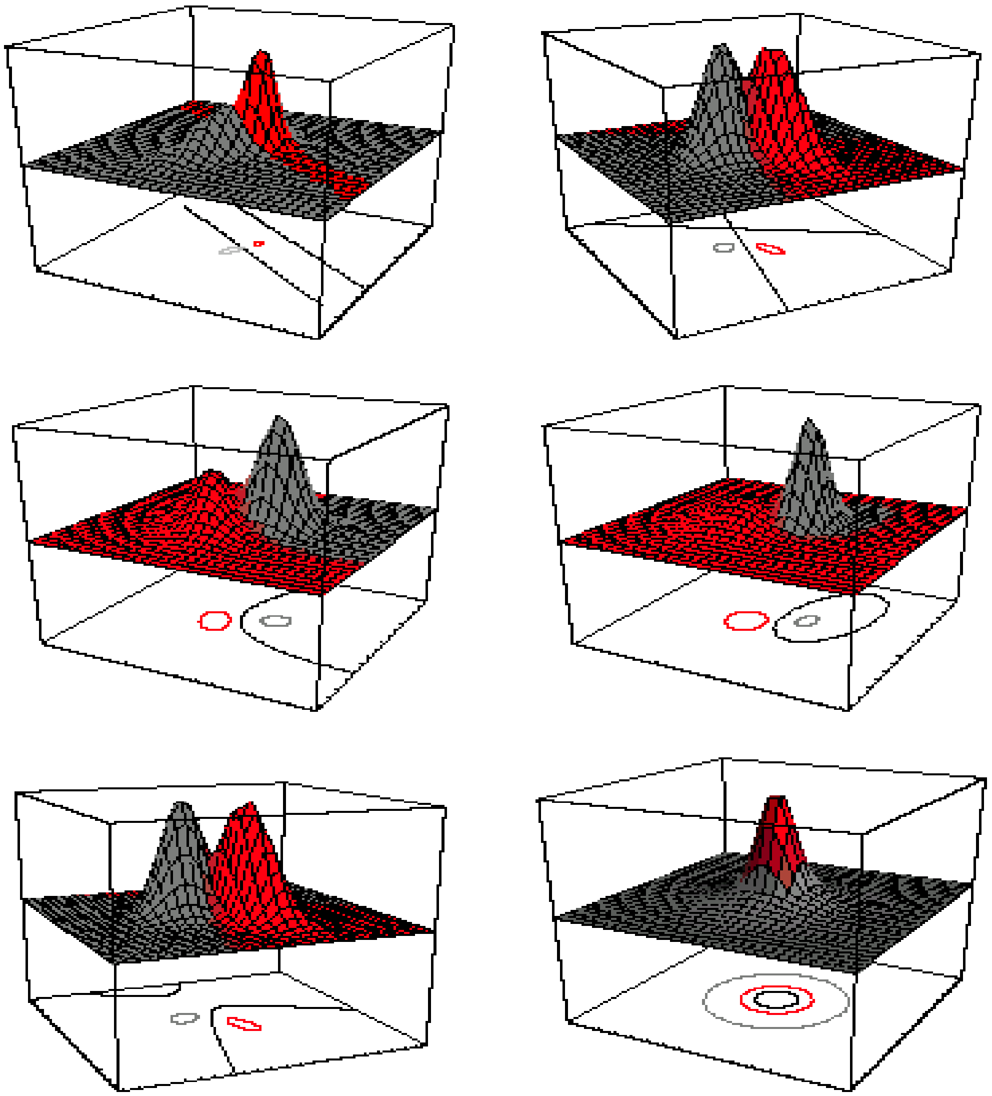


FIGURE 2.14. Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadric, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric. These variances are indicated by the contours of constant probability density. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Summary

- Bayesian decision theory:
  - In theory: tells you how to make optimal (minimum-risk) decisions
  - In practice: where do you get those probabilities from?
    - Expert knowledge + learning from data (see next; also, Chapter 3)
- Note: some typos in Chapter 2
  - Page 25, second line after the equation 12: must be $\mathrm{R}(\alpha_i(\mathrm{x}) \mid \mathrm{x}), \text{not } \mathrm{R}(\alpha_i(\mathrm{x}))$
  - Page 27, third line before section 2.3.1:

    $$\text{switch } \omega_1 \text{ and } \omega_2, \text{ and reverse inequality in } \lambda_{21} > \lambda_{12}$$
  - Page 50 and 51, in Fig. 2.20 and Fig 2.21, replace x-axis label by

    $$P(x > x^* \mid x \in \omega_1)$$
  - Same replacement in problem 9, page 75
  - Section 2.11 (Bayesian belief networks): contains several mistakes; ignore for now. Bayesian networks will be covered later.

Parameter Estimation

- In general, given **training data** $\mathbf{D} = \{\mathbf{y}^1,\ldots,\mathbf{y}^N\}$, where $\mathbf{y}^j = (\mathbf{x}^j, \omega^j)$,

  we wish to find (estimate) $P(C = \omega_i)$ and $P(\mathbf{x} \mid \omega_i)$ - e.g., density estimation problem
- Usually, estimating $P(\mathbf{x} \mid \omega_i)$ is hard, especially in high - dimensional feature spaces
- Solution : simplifying assumptions (e.g., parametric form of $P(\mathbf{x} \mid \omega_i)$), or feature independence, etc.)
- We consider first a fixed parametric distribution approach (e.g., Gaussian, multinomial, etc.)
- Then learning = parameter estimation from data
- Example: assume $p(x \mid \omega_i)$, is Gaussian $N(\mu_i, \sigma_i)$, estimate μ_i, σ_i
- Two major approaches : classical statistical (ML) and Bayesian (MAP)
- Philosophical difference : is parameter a 'physical' constant or a random variable?

Maximum likelihood (ML) and Maximum a posteriory (MAP) estimates

- Assume independent and identically distributed (i.i.d.) samples

$$\mathbf{D} = \{\mathbf{y}^1, ..., \mathbf{y}^N\}, \text{ where } \mathbf{y}^j = (\mathbf{x}^j, \omega^j)$$

- Assumea parametricdistribution $p(\mathbf{x} \mid \omega_i, \Theta_i)$, where Θ_i is a parametervector

Example: $\Theta_i = (\mu_i, \sigma_i)$ for Gaussian $p(\mathbf{x} \mid \omega_i, \Theta_i)$

- Wealsoassumethat Θ_i for different classes are independent

(can be estimated separately in same way)

- Then $P(D \mid \Theta) = \prod_{j=1}^{N} p(\mathbf{x}^j \mid \Theta)$

- Maximum-likelihoodestimate (Θ is an unknownconstant):

$$\hat{\Theta} = \arg\max_{\Theta} l(\Theta) = P(D \mid \Theta)$$

- Maximum a posterioryestimate (Θ is an unknown randomvariable,with prior $P(\Theta)$)

$$\hat{\Theta} = \arg\max_{\Theta} P(\Theta \mid D) = \arg\max_{\Theta} P(D \mid \Theta) P(\Theta)$$

- Note that ML = MAP with uniformprior

ML estimate: Gaussian distribution

- known σ, estimate μ :

$$\hat{\mu} = \frac{1}{N}\sum_{j=1}^{N} x^{j}$$

- both μ and σ are unknown :

$$\hat{\mu} = \frac{1}{N}\sum_{j=1}^{N} x^{j}, \quad \hat{\sigma} = \frac{1}{N}\sum_{j=1}^{N} (x^{j} - \hat{\mu})^{2}$$

- Note that $\hat{\sigma}$ is biased, i.e. $\mathrm{E}\left[\frac{1}{N}\sum_{j=1}^{N} (x^{j} - \hat{\mu})^{2}\right] = \frac{N-1}{N}\sigma^{2} \neq \sigma^{2}$

- Unbiased estimate would be $\hat{\sigma} = \frac{1}{N-1}\sum_{j=1}^{N} (x^{j} - \hat{\mu})^{2}$

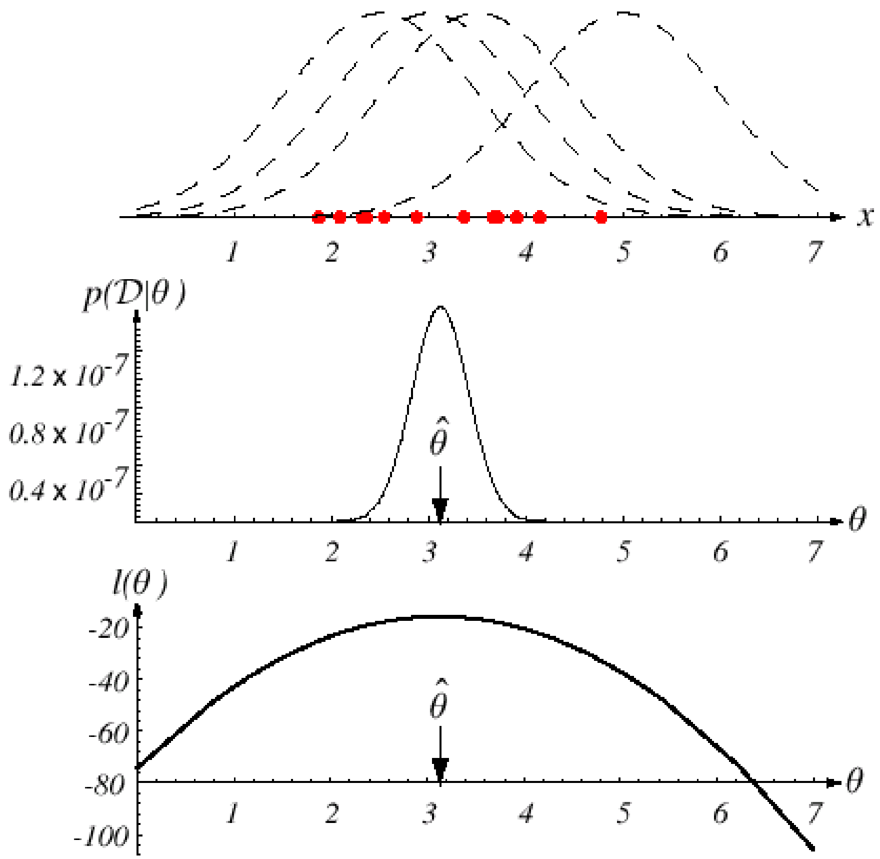


FIGURE 3.1. The top graph shows several training points in one dimension, known or assumed to be drawn from a Gaussian of a particular variance, but unknown mean. Four of the infinite number of candidate source distributions are shown in dashed lines. The middle figure shows the likelihood $p(\mathcal{D}|\theta)$ 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*. Copyright © 2001 by John Wiley & Sons, Inc.

Bayesian (MAP) estimate with increasing sample size

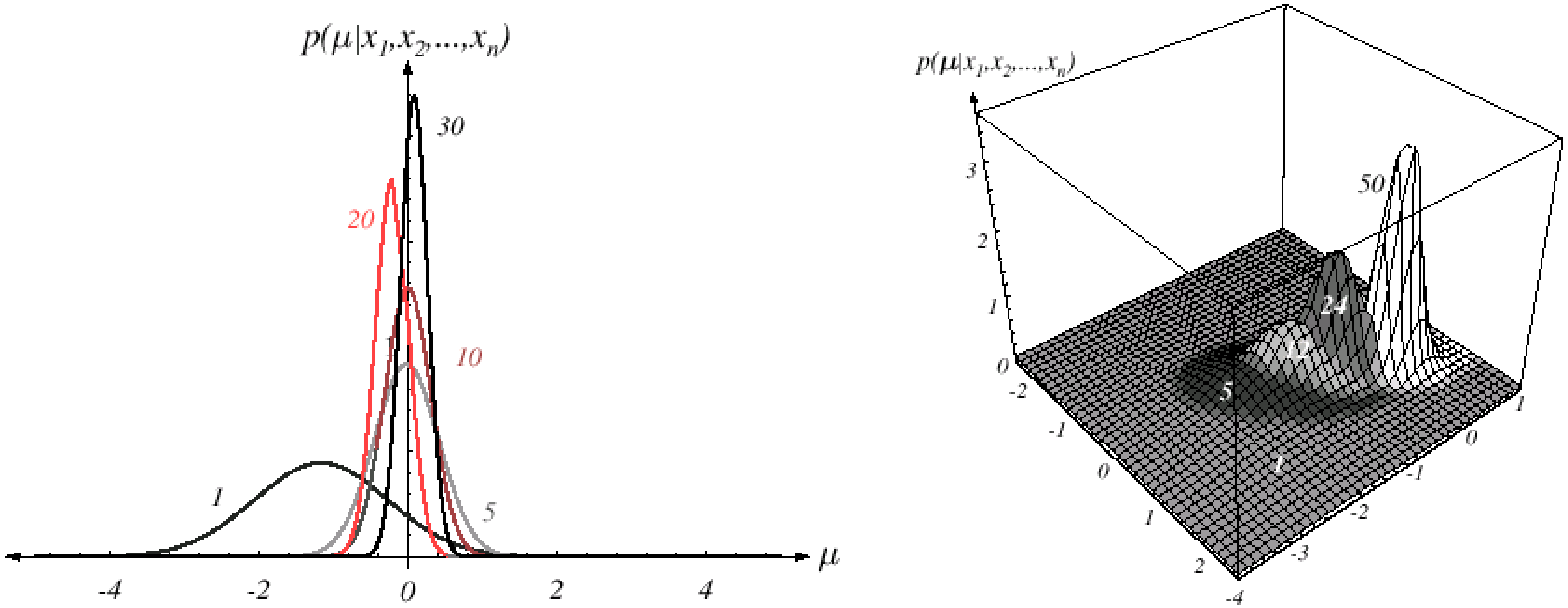
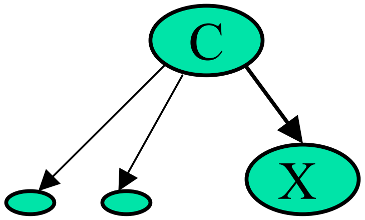


FIGURE 3.2. Bayesian learning of the mean of normal distributions in one and two dimensions. The posterior distribution estimates are labeled by the number of training samples used in the estimation. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001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

Parameter estimation: discrete features



Multinomial P(x|C)

$$\theta_k^i = P(x = k \mid C = \omega_i)$$

- ML-estimate: $\hat{\Theta} = \arg\max_{\Theta} \log P(D|\Theta)$

counts

$$\boxed{\mathrm{ML}(\theta_k^i) = \frac{N_{x=k,c=i}}{\sum_k N_{x=k,c=i}}}$$

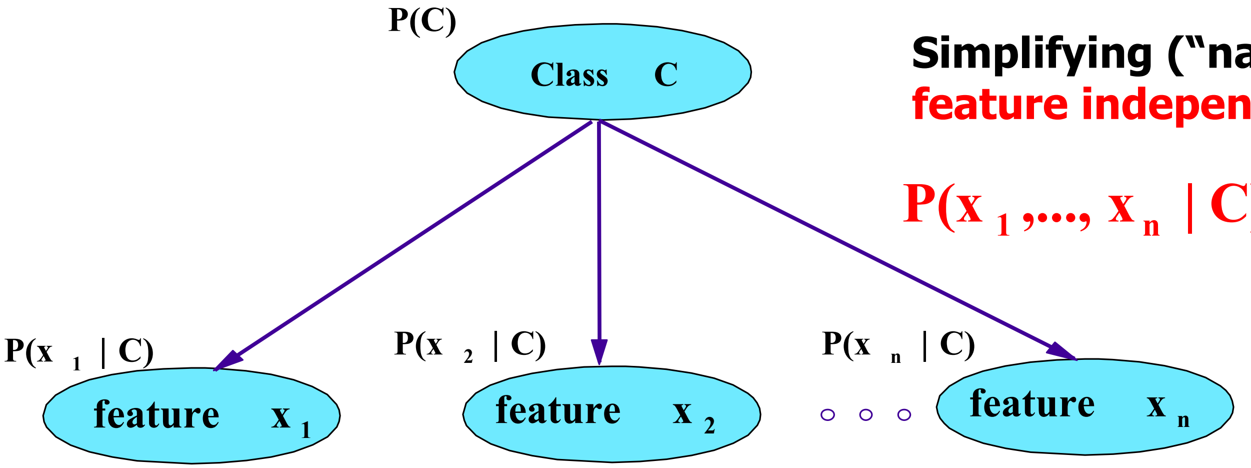
- MAP-estimate $\max_{\Theta} \log \underbrace{P(D \mid \Theta) P(\Theta)}$

Conjugate priors - **Dirichlet** $Dir(\theta_{\mathbf{pa}_X} \mid \alpha_{1,\mathbf{pa}_X}, \ldots, \alpha_{m,\mathbf{pa}_X})$

$$\boxed{\mathrm{MAP}(\theta_{x,\mathbf{pa}_X}) = \frac{N_{x,\mathbf{pa}_X} + \alpha_{x,\mathbf{pa}_X}}{\sum_x N_{x,\mathbf{pa}_X} + \underbrace{\sum_x \alpha_{x,\mathbf{pa}_X}}}}$$

Equivalent sample size (prior knowledge)

An example: naïve Bayes classifier



Simplifying ("naïve") assumption:
feature independence given class

$$P(x_1,\dots,x_n \mid C) = \prod_{i=1}^{n} P(x_i \mid C)$$

1. Bayes(-optimal) classifier:
   given an (unlabeled) instance $\bar{x} = (x_1,\dots,x_n)$, choose most likely class:

$$BO(x) = \arg\max_i P(C = i \mid \bar{x})$$

2. Naïve Bayes classifier:

By Bayes rule $P(C = i \mid \bar{x}) = \frac{P(\bar{x} \mid C = i)P(C = i)}{P(\bar{x})}$, and by independence assumption

$$NB(x) = \arg\max_i \prod_{j=1}^{n} P(C = i)P(x_j \mid C = i)$$

State-of-the-art

- Optimality results
  - Linear decision surface for binary features (Minsky 61, Duda&Hart 73)
  - (polynomial for general nominal features - Duda&Hart 1973, Peot 96)
  - Optimality for OR and AND concepts (Domingos&Pazzani 97)
  - **No XOR-containing concepts on nominal features** (Zhang&Ling 01)

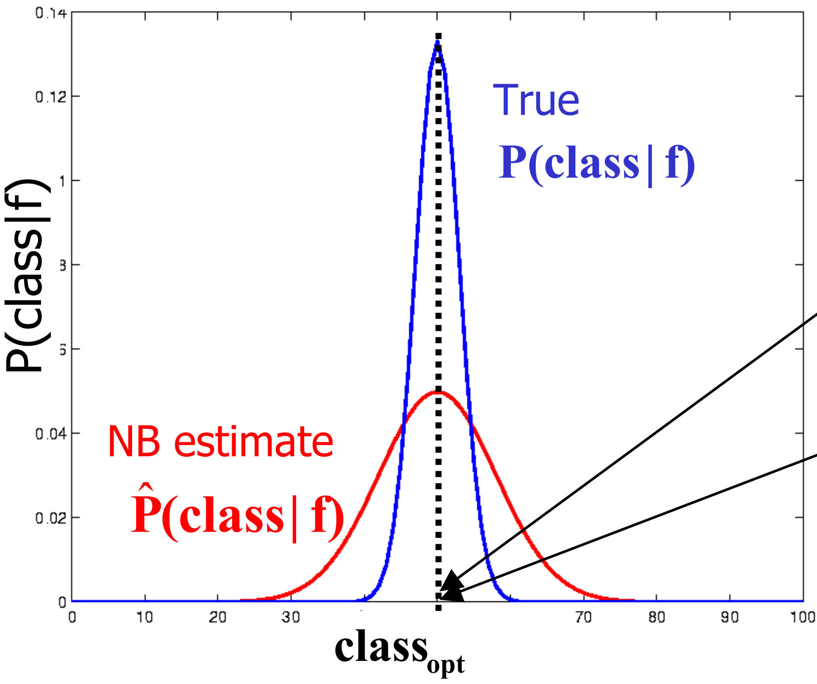
- Algorithmic improvements
  - Boosted NB (Elkan 97) is equivalent to multilayer perceptron
  - Augmented NB (TAN, Bayes Nets – e.g., Friedman et al 97)
  - Other improvments (combining with Decision Trees (Kohavi), w/ error-correcting output coding (ECOC) (Ghani, ICML 2000), etc.

- Still, open problems remain:
  - NB error estimate/bounds based on domain properties

Why Naïve Bayes often works well (despite independence assumption)?

Wrong P(C|x) estimates do not imply wrong classification!

Domingos&Pazzani, 97, J. Friedman 97, etc.
"Statistical diagnosis based on conditional independence does not require it", J. Hilden 84



Bayes-optimal: $\text{class}_{\text{opt}} = \arg\max_{i} P(\text{class}_i \mid f)$

Naïve Bayes: $\text{class}_{\text{NB}} = \arg\max_{i} \hat{P}(\text{class}_i \mid f)$

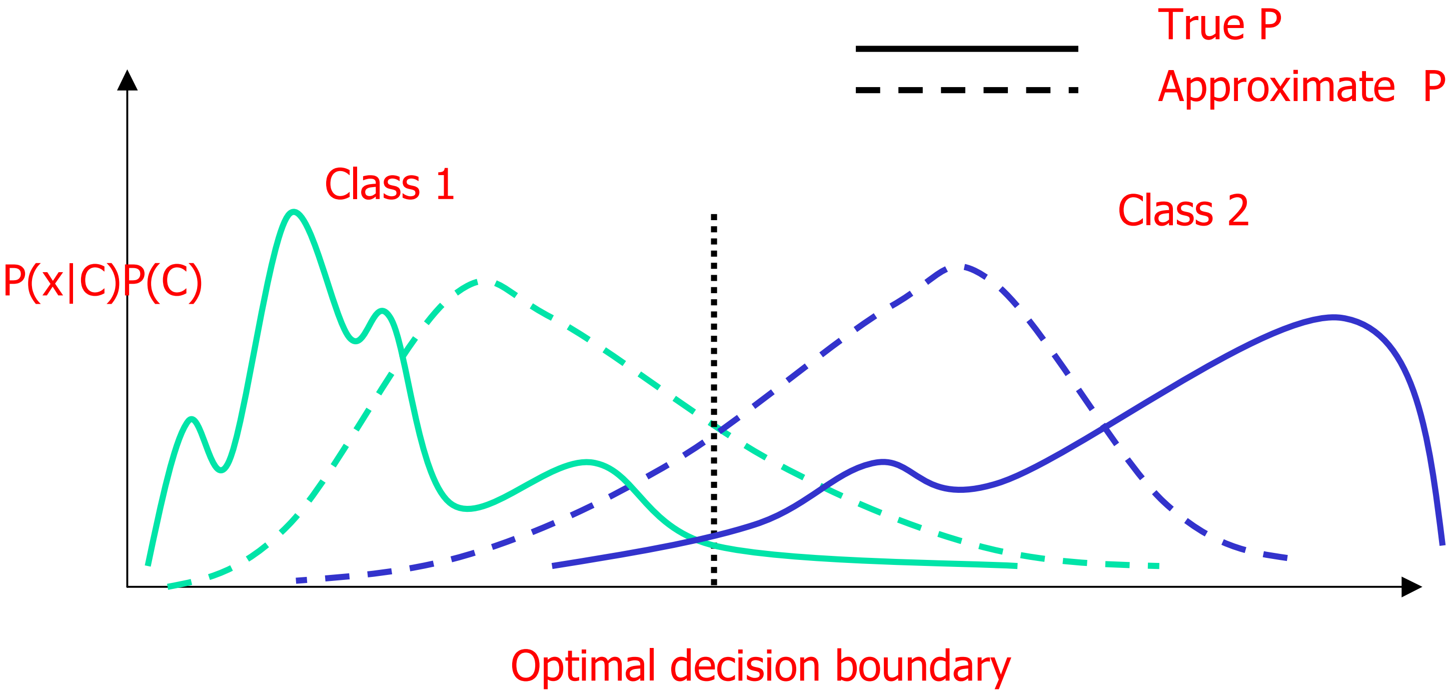
Major questions remain:

Which P(c,x) are 'good' for NB?

What domain properties "predict" NB accuracy?

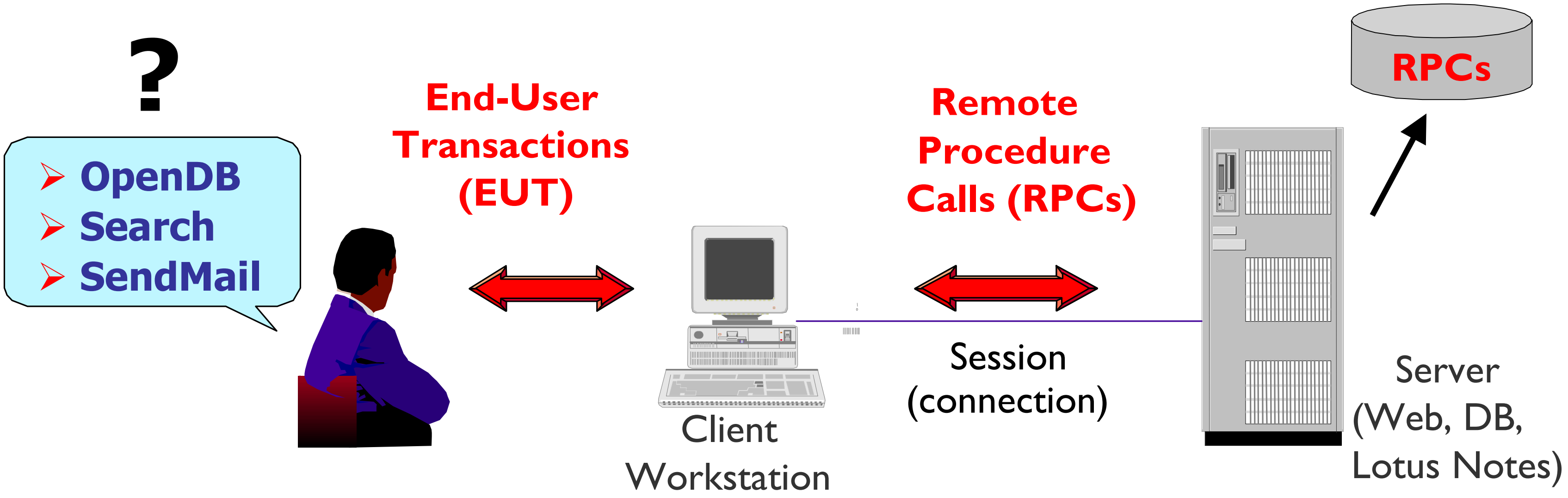
General question:

characterizing distributions P(X,C) and their approximations Q (X,C) that can be 'far' from P(X,C), but yield low classification error



Note: one measure of 'distance' between distributions can be relative entropy, or KL-divergence (see hw problem11, chap.3) $D(P\,||\,Q) = \int P(z)\log\frac{P(z)}{Q(z)}dz$

Case study: using Naïve Bayes for Transaction Recognition Problem



?

- OpenDB
- Search
- SendMail

End-User Transactions (EUT)

Client Workstation

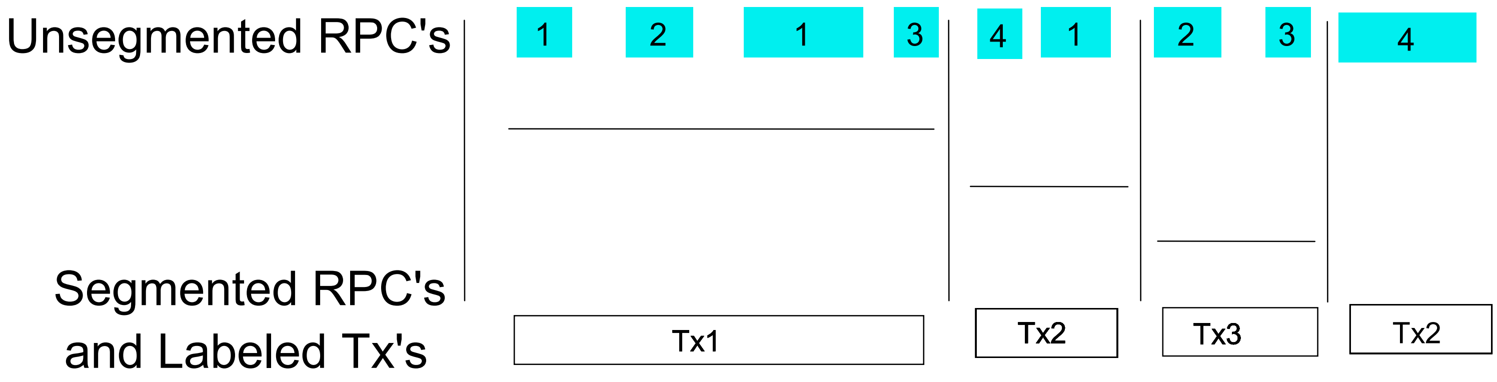
Remote Procedure Calls (RPCs)

Session (connection)

RPCs

Server (Web, DB, Lotus Notes)

Two problems: segmentation and labeling



Unsegmented RPC's

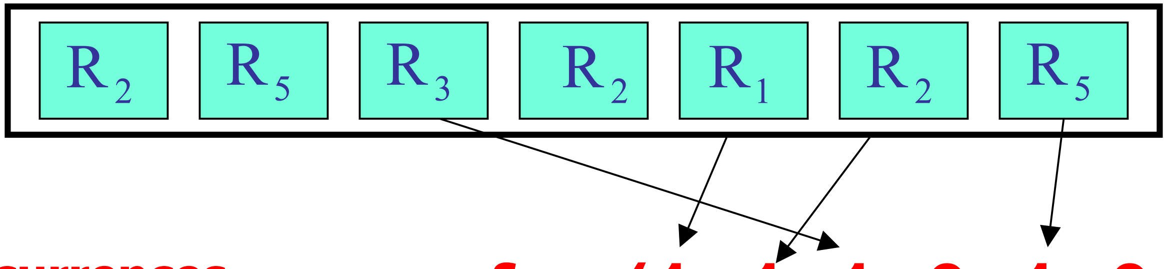
1 2 1 3 4 1 2 3 4

Segmented RPC's and Labeled Tx's

Tx1 Tx2 Tx3 Tx2

Representing transactions as feature vectors

Transaction of type *i*



- **RPC occurrences**

$$\text{Bernoulli}: \ P(R_j = 1 \mid T_i) = p_{ij}$$

$$f = (1, 1, 1, 0, 1, 0, \ldots)$$

- **RPC counts**

$$f = (1, 3, 1, 0, 2, 0, \ldots)$$

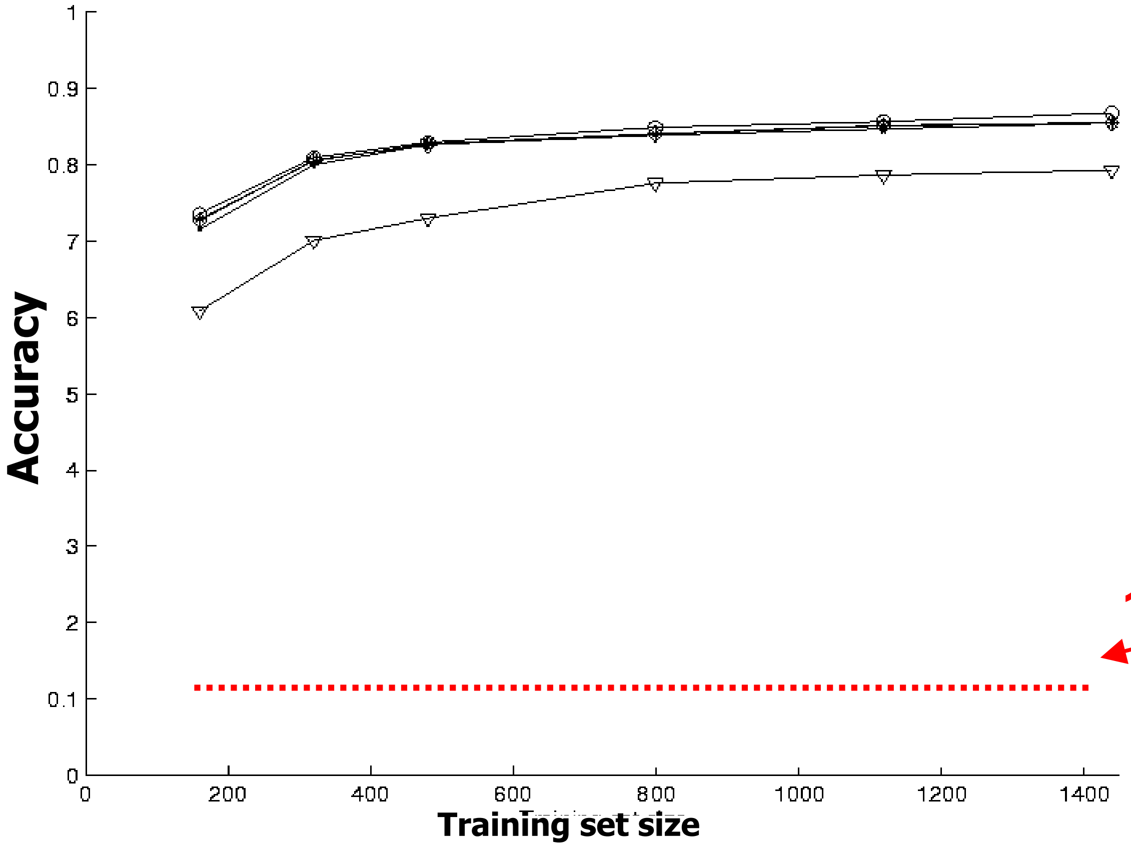
$$\text{Multinomial}: \ P(n_{i1}, \ldots, n_{iM} \mid T_i) = \frac{n!}{\prod_{j=1}^{M} n_{ij}!} \prod_{j=1}^{M} p_{ij}^{\,n_{ij}}$$

$$\text{Geometric}: \ P(n_{ij} \mid T_i) = p_{ij}^{\,n_{ij}} (1 - p_{ij})$$

$$\text{Shifted Geometric}: \ P(n_{ij} \mid T_i) = p_{ij}^{\,n_{ij} - s_{ij}} (1 - p_{ij})$$

Best fit to data (χ^2): shifted geometric

Empirical results



87 ± 2% **NB + Bernoulli, mult. or geom.**

79 ± 3% **NB + shifted geom.**

10 ± 1% **Baseline classifier: Always selects most-frequent transaction**

- **Significant improvement over baseline classifier (75%)**
- **NB is simple, efficient, and comparable to the state-of-the-art classifiers:**
  - **SVM – 85-87%, Decision Tree – 90-92%**
- **Best-fit distribution (shift. geom) - not necessarily best classifier! (?)**

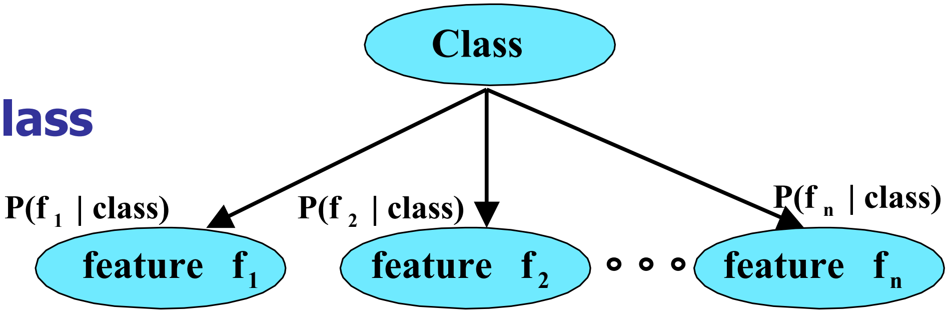
Next lecture on Bayesian topics

April 17, 2002 - lecture on recent ‘hot stuff’:
Bayesian networks, HMMs, EM algorithm

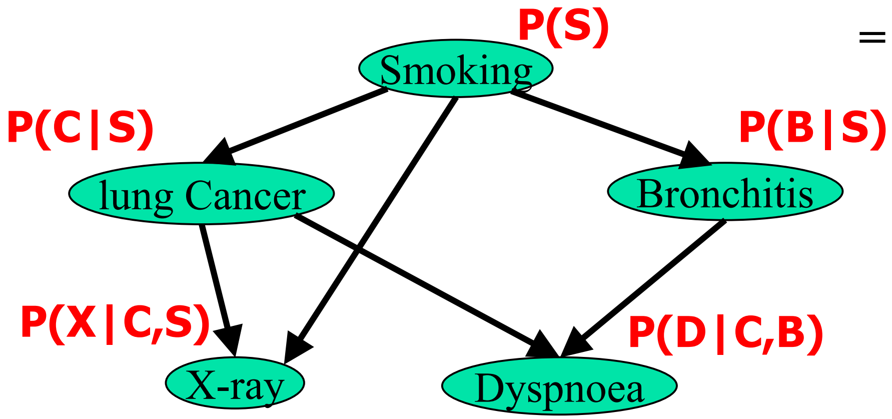
Short Preview:

From Naïve Bayes to Bayesian Networks

Naïve Bayes model:
independent features given class



Bayesian network (BN) model:
Any joint probability distributions



$$P(S, C, B, X, D)=$$

$$= P(S)\ P(C|S)\ P(B|S)\ P(X|C,S)\ P(D|C,B)$$

CPD:

C	B	D=0	D=1
0	0	0.1	0.9
0	1	0.7	0.3
1	0	0.8	0.2
1	1	0.9	0.1

Query: P (lung cancer=yes | smoking=no, dyspnoea=yes) = ?