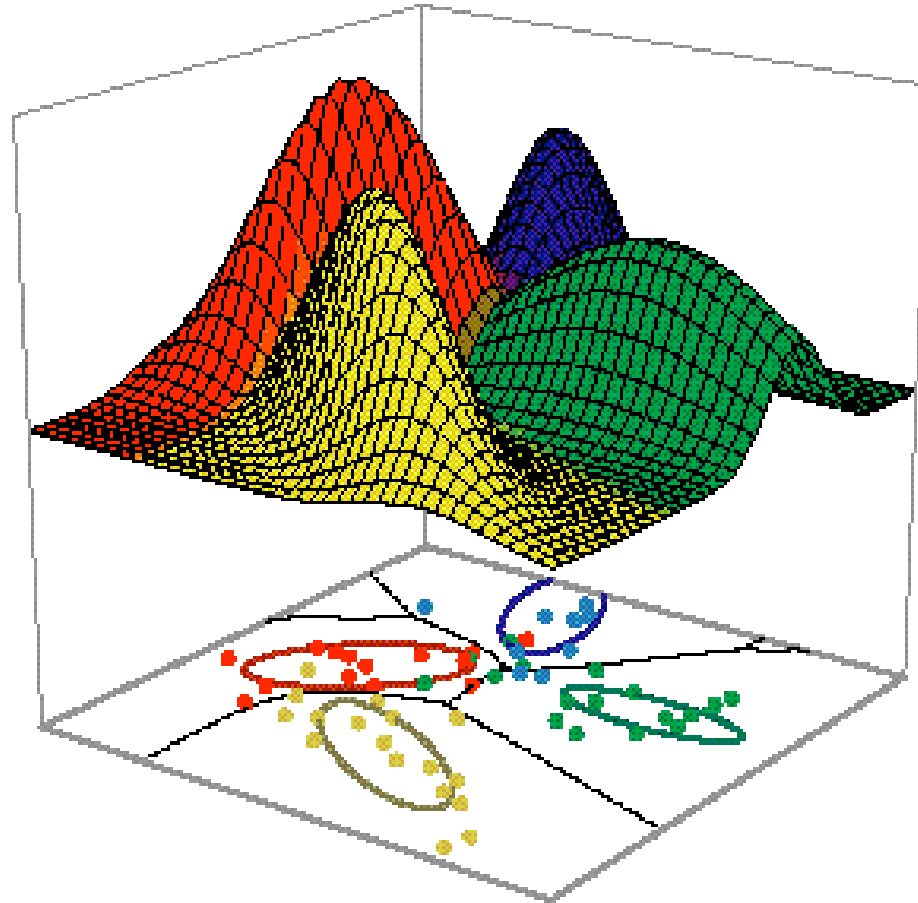


Machine Learning for Context Aware Computing

Bayesian Decision Theory



Machine Learning for Context Aware Computing

index of contents

introduction (Chapter 1 – Pattern Classification, Duda/Hart/Stork)

> machine perception

- an example
- pattern recognition systems
- the design cycle
- learning and adaptation
- conclusion

Bayesian decision theory (Chapter 2 – Pattern Classification)

Machine Learning for Context Aware Computing

machine perception

build a machine that can recognize patterns:

- speech recognition
- fingerprint identification
- DNA sequence identification
- OCR (optical character recognition)

accurate pattern recognition immensely useful

deeper understanding by solving problems

algorithm and hardware design is influenced by knowledge how these are solved in nature

Machine Learning for Context Aware Computing

index of contents

introduction to ML

- machine perception

- > an example

- pattern recognition systems

- the design cycle

- learning and adaptation

- conclusion

Bayesian decision theory

Machine Learning for Context Aware Computing

an example – fish packing plant (1/8)

aims

- wants automate process of incoming fish
- pilot project: separate sea bass from salmon
- using optical sensing

problem analysis

- take sample pictures
- extract features
 - > length, width
 - > lightness
 - > number and shape of fins
- notice noise or variations in the images
 - variation in lighting
 - position of the fish on the conveyor

Machine Learning for Context Aware Computing

an example – fish packing plant (2/8)

model

- differences between the population, different models
- hypothesize the class of models
- choose best corresponding model

preprocessing

- use a segmentation operation to isolate fishes from...
 - > one another
 - > background

feature extraction

- information from a single fish is sent to a feature extractor whose purpose is to reduce the data by measuring certain features
- the features are passed to a classifier

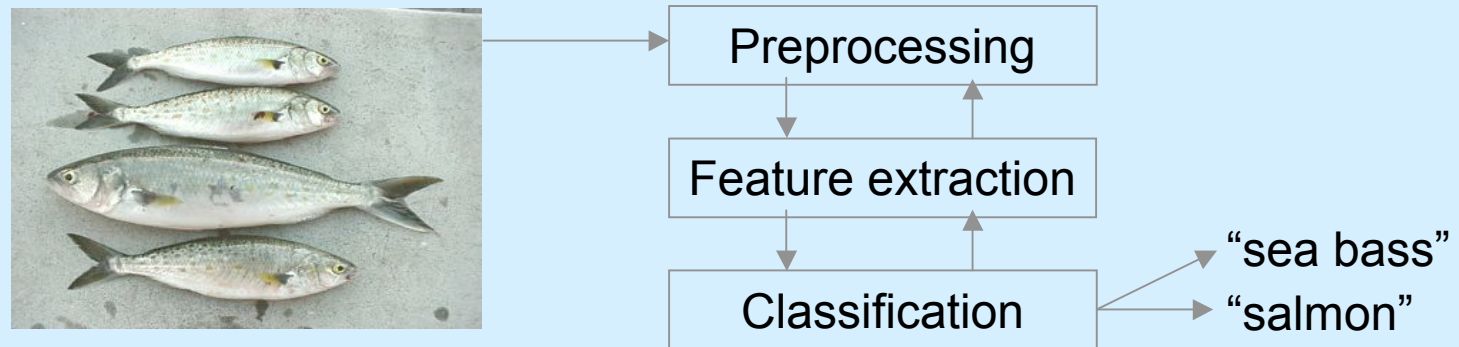
Machine Learning for Context Aware Computing

an example – fish packing plant (3/8)

classification

- evaluates evidence
- makes final decision

overview



training samples

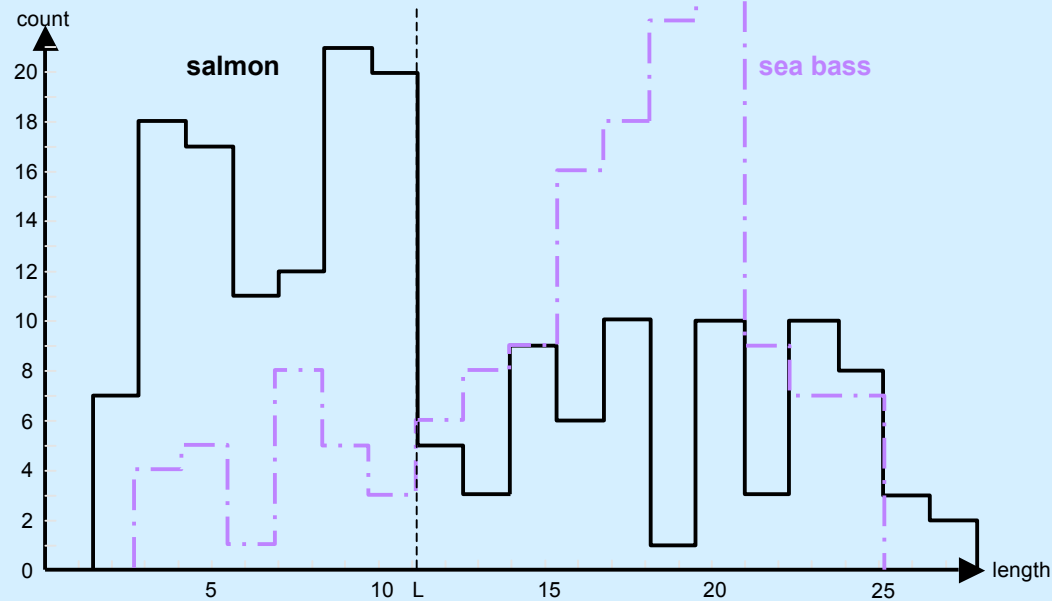
- suppose sea bass is generally longer than a salmon
- length obvious feature, try to classify by the length L
- obtain training samples by making length measurements

Machine Learning for Context Aware Computing

an example – fish packing plant (4/8)

feature: length

- length alone is a poor feature



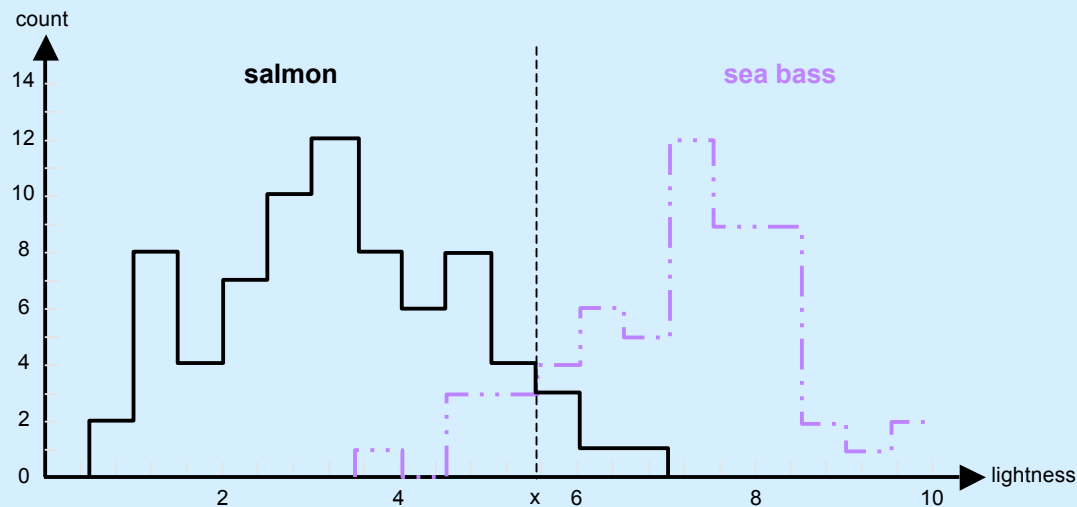
- select the lightness as a possible feature

Machine Learning for Context Aware Computing

an example – fish packing plant (5/8)

feature: lightness

- careful elimination of variations in illumination



- classes much better separated

decision boundary and cost relationship

- Move our decision boundary toward smaller values of lightness in order to minimize the cost (reduce the number of sea bass that are classified salmon!)
- task of decisions theory

Machine Learning for Context Aware Computing

an example – fish packing plant (6/8)

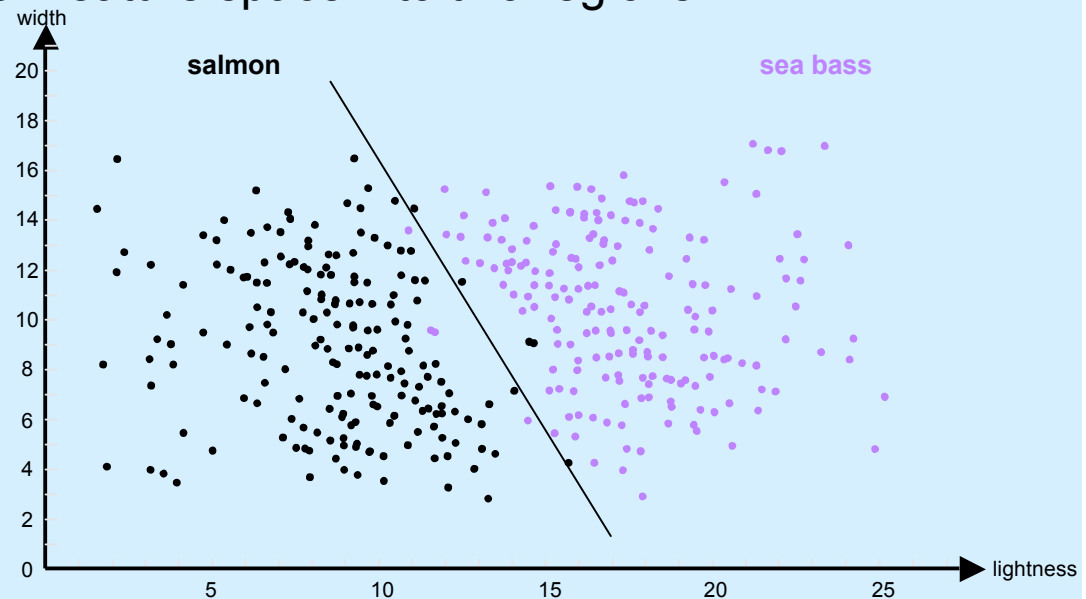
decision theory

- make decision rules, such as to minimize cost
- width as new feature to classify
- add other features that are not correlated with the ones we already have
- a precaution should be taken not to reduce the performance by adding such “noisy features”
- problem to partition feature space into two regions

$$\rightarrow x^T = [x_1, x_2]$$

$x_1 = \text{lightness}$

$x_2 = \text{width}$

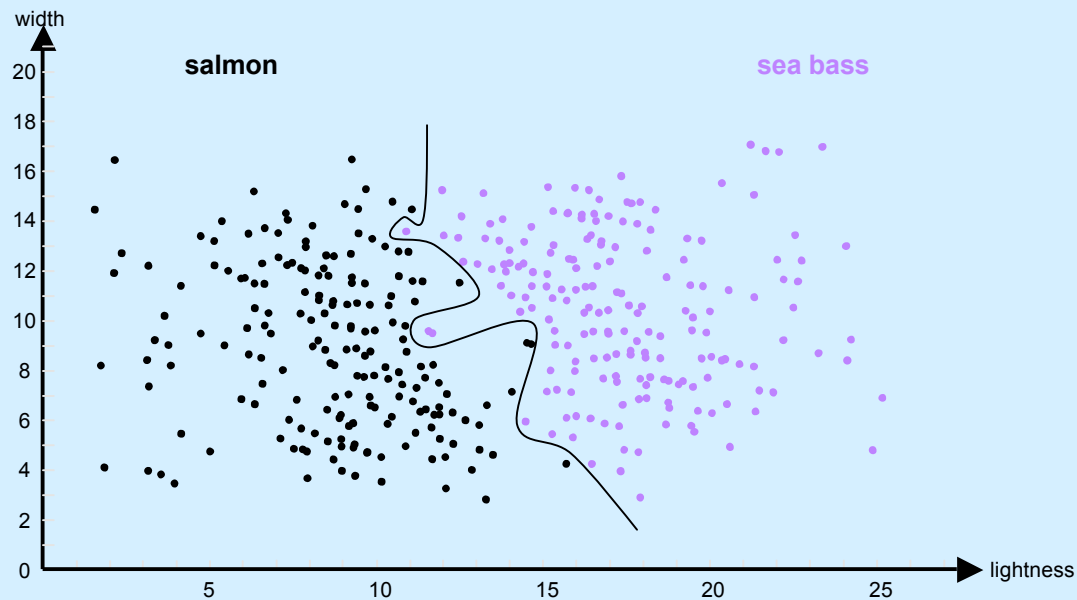


Machine Learning for Context Aware Computing

an example – fish packing plant (7/8)

best decision boundary

- best decision boundary should be the one which provides an optimal performance such as in the following figure:

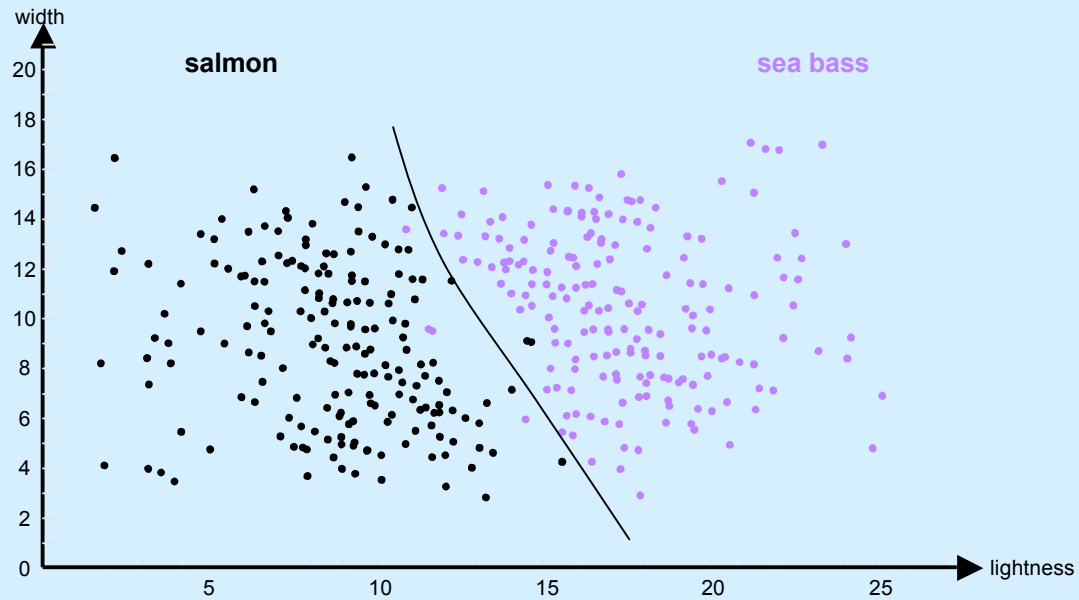


- satisfaction is premature because the central aim of designing a classifier is to correctly classify novel input
- issue of generalization

Machine Learning for Context Aware Computing

an example – fish packing plant (8/8)

generalized decision boundary



Machine Learning for Context Aware Computing

index of contents

introduction to ML

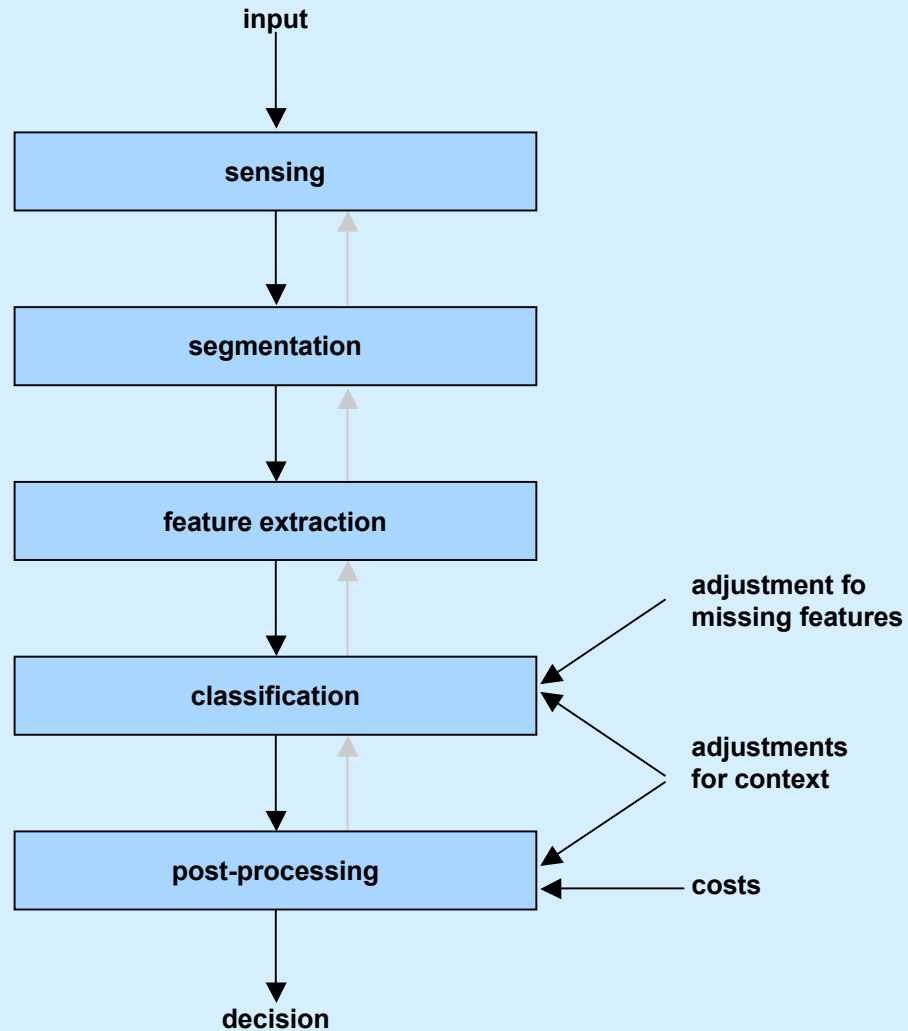
- machine perception
- an example
- > pattern recognition systems
- the design cycle
- learning and adaptation
- conclusion

Bayesian decision theory

Machine Learning for Context Aware Computing

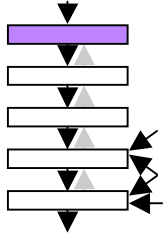
pattern recognition systems (1/3)

overview



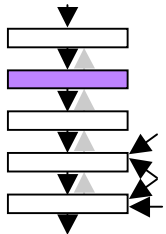
Machine Learning for Context Aware Computing

pattern recognition systems (2/3)



sensing

- use of a transducer (i.e. camera)
- pattern recognition systems depends off:
 - > the bandwidth
 - > the resolution sensitivity distortion of the transducer



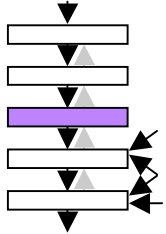
segmentation and grouping

- patterns should:
 - > be well separated
 - > not overlap

Machine Learning for Context Aware Computing

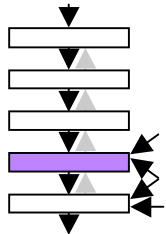
pattern recognition systems (3/3)

feature extraction



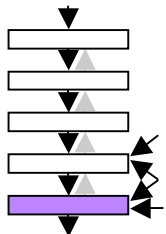
- arbitrary boundary between feature extraction and classification
- invariant features with respect to translation, rotation and scale

classification



- use a feature vector provided by a feature extractor to assign the object to a category

post processing



- exploit context input dependent information other than from the target pattern itself to improve performance

Machine Learning for Context Aware Computing

index of contents

introduction to ML

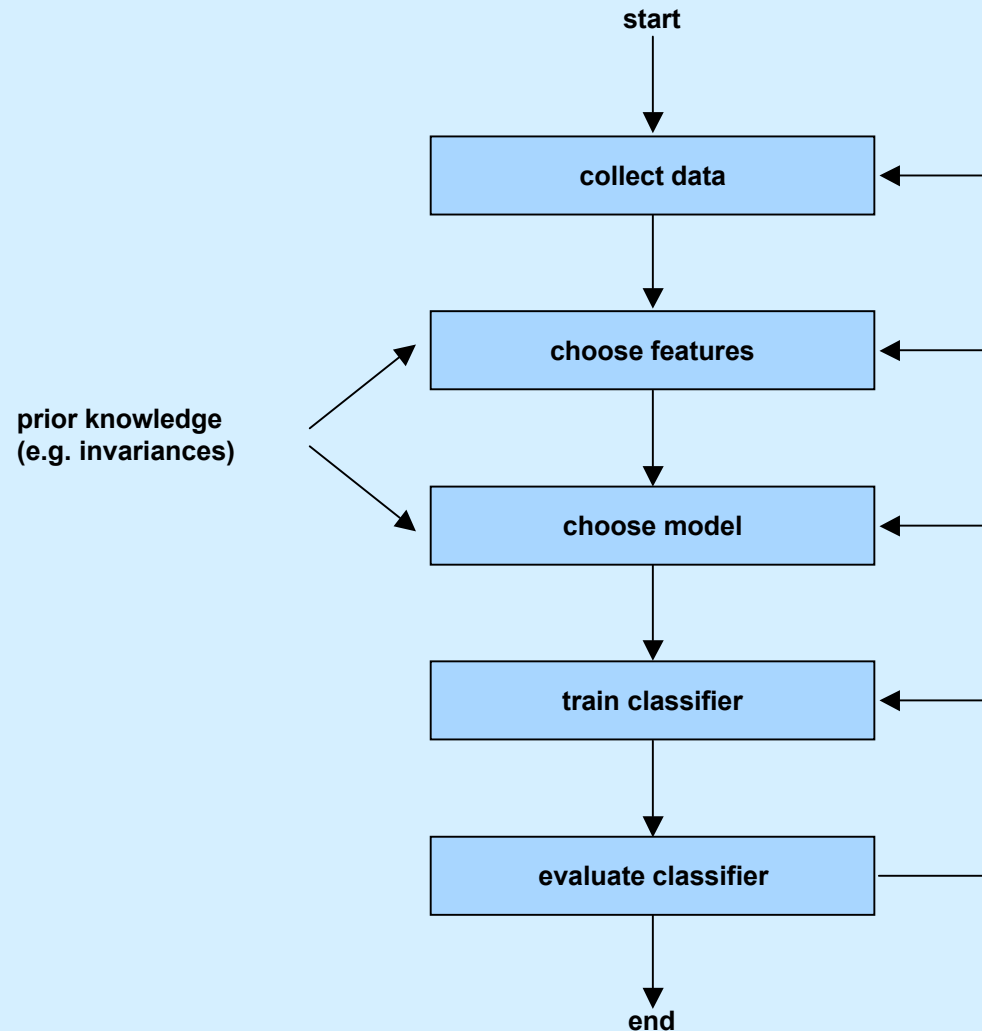
- machine perception
- an example
- pattern recognition systems
- > the design cycle
- learning and adaptation
- conclusion

Bayesian decision theory

Machine Learning for Context Aware Computing

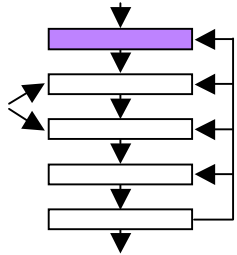
the design cycle (1/3)

overview



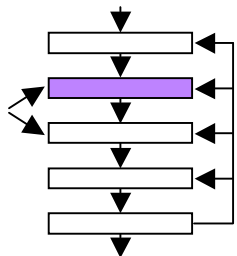
Machine Learning for Context Aware Computing

the design cycle (2/3)



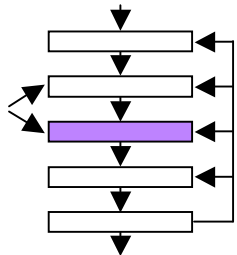
data collection

- how do we know when we have collected an adequately large and representative set of examples for training and testing the system?



feature choice

- depends on the characteristics of the problem domain. Simple to extract, invariant to irrelevant transformation insensitive to noise
- how to combine prior knowledge and empirical data?

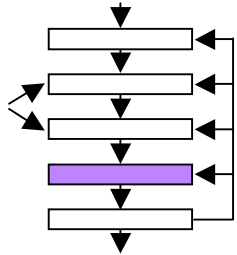


model choice

- unsatisfied with the performance of our fish classifier and want to jump to another class of model

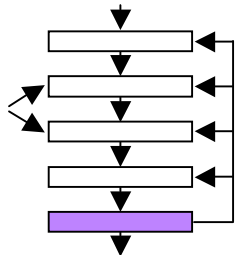
Machine Learning for Context Aware Computing

the design cycle (3/3)



training

- use data to determine the classifier
- different procedures for training classifiers and choosing models



evaluation

- measure the error rate (or performance and switch from one set of features to another one)

solving problems

- no universal methods have been found for solving problems

Machine Learning for Context Aware Computing

index of contents

introduction to ML

- machine perception
- an example
- pattern recognition systems
- the design cycle
- > learning and adaptation
- conclusion

Bayesian decision theory

Machine Learning for Context Aware Computing

learning and adaptation

learning

- pattern recognition problems to hard to guess best classification decision

supervised learning

- teacher provides a category label or cost for each pattern in the training set

unsupervised learning

- the system forms clusters or “natural groupings” of the input patterns

Machine Learning for Context Aware Computing

index of contents

introduction to ML

- machine perception
- an example
- pattern recognition systems
- the design cycle
- learning and adaptation

> conclusion

Bayesian decision theory

Machine Learning for Context Aware Computing

conclusion

overwhelmed

- seems to be overwhelmed by the number, complexity and magnitude of the sub-problems of pattern recognition

problems

- many of these sub-problems can indeed be solved
- mathematical theories solving some problems have been discovered

unresolved problems

- many fascinating unsolved problems still remain

Bayesian decision theory

index of contents

introduction to ML

Bayesian decision theory

> introduction

- Bayesian decision theory / continuous features
- minimum-error-rate classification
- classifiers, discriminant functions and decision surfaces
- the normal density
- discriminant functions for the normal density
- Bayesian decision theory / discrete features

Bayesian decision theory

introduction (1/4)

assumptions

- sequence of types of fish appears to be random
- decision-theoretic terminology: each fish emerges nature is in one or the other of two possible states

state of nature

- ω denote state of nature
- $\omega_1 = \text{sea bass}$ and $\omega_2 = \text{salmon}$

a priori probability

- $P(\omega_1) = \text{priority next fish is a sea bass}$
- the catch of salmon and sea bass is equiprobable:
 $P(\omega_1) = P(\omega_2)$ (uniform priors)
 $P(\omega_1) + P(\omega_2) = 1$ (exclusivity and exhaustivity)

Bayesian decision theory

introduction (2/4)

example

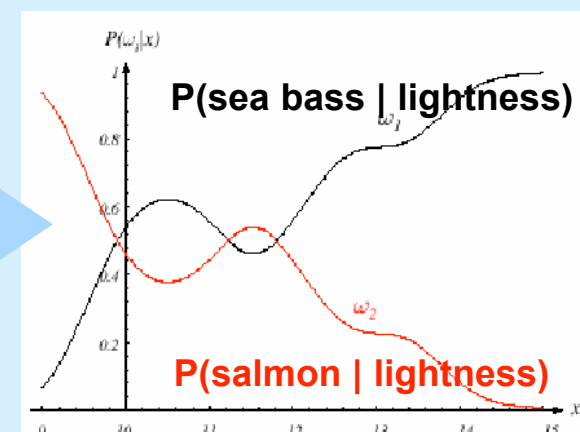
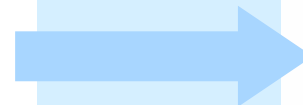
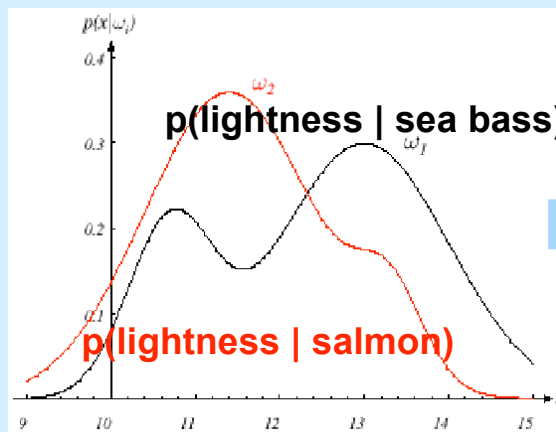
- classification problem of sea bass and salmon by lightness
- assume apriori probabilities are not equal

> i.e. assume $P(\text{sea bass}) > P(\text{salmon})$

> if you don't have a chance to see the fish, every time decide as a sea bass

- if you see the lightness of fish

Question: $P(\text{sea bass} \mid \text{lightness}) = ?$ and
 $P(\text{salmon} \mid \text{lightness}) = ?$



Bayesian decision theory

introduction (3/4)

Bayes' rule

$$P(\Omega_j | x) = \frac{p(x | \Omega_j) \times P(\Omega_j)}{p(x)}$$

where in case of two categories

$$p(x) = \prod_{j=1}^{j=2} p(x | \Omega_j) P(\Omega_j)$$

decision given the posterior probabilities

- x is an observation for which:

$$> P(\Omega_1 | x) > P(\Omega_2 | x) \quad \rightarrow \quad \text{True state of nature} = \Omega_1$$

$$> P(\Omega_1 | x) < P(\Omega_2 | x) \quad \rightarrow \quad \text{True state of nature} = \Omega_2$$

- whenever we observe a particular x, the probability of error is :

$$> P(\text{error} | x) = P(\Omega_1 | x) \text{ if we decide } \Omega_2$$

$$> P(\text{error} | x) = P(\Omega_2 | x) \text{ if we decide } \Omega_1$$

- maximum a posterior classifier or Bayes classifier

Bayesian decision theory

introduction (4/4)

minimizing the probability of error

- decide \square_1 if $P(\square_1 | x) > P(\square_2 | x)$; otherwise decide \square_2

$$P(\text{error} | x) = \min [P(\square_1 | x), P(\square_2 | x)]$$

(Bayes decision)

maximum-likelihood classifier

- $P(\square_1) = P(\square_2)$
- simpler decision rule
- decide \square_1 if $p(x | \square_1) > p(x | \square_2)$; otherwise decide \square_2

Bayesian decision theory

index of contents

introduction to ML

Bayesian decision theory

- introduction
- > Bayesian decision theory / continuous features
- minimum-error-rate classification
- classifiers, discriminant functions and decision surfaces
- the normal density
- discriminant functions for the normal density
- Bayesian decision theory / discrete features

Bayesian decision theory

Bayesian decision theory / continuous features (1/4)

generalization of the preceding ideas

- use of more than one feature
- use more than two states of nature
- allow actions and not only decide the state of nature
- introduce a loss of function which is more general than the probability of error

feature space

- replace scalar x by the feature vector \mathbf{x}

loss function

- states how costly each action is
- let us treat situations in which some kinds of classification mistakes are more costly than others

Bayesian decision theory

Bayesian decision theory / continuous features (2/4)

definitions

- let $\{\omega_1, \omega_2, \dots, \omega_c\}$ be the set of c states of nature (or “categories”)
- let $\{\alpha_1, \alpha_2, \dots, \alpha_a\}$ be the set of possible actions
- let $\lambda(\alpha_i | \omega_j)$ be the loss incurred for taking action α_i when the state of nature is ω_j
- a posterior probability $P(\omega_j | \mathbf{x}) = \frac{p(\mathbf{x} | \omega_j) \times P(\omega_j)}{p(\mathbf{x})}$

risk := expected values (cost)

- expected loss by taking action α_i :

$$R(\alpha_i | \mathbf{x}) = \sum_{j=1}^{j=c} \lambda(\alpha_i | \omega_j) P(\omega_j | \mathbf{x})$$

select the action α_i for which $R(\alpha_i | \mathbf{x})$ is minimum

- R is minimum and R in this case is called the Bayes risk = best performance that can be achieved!

Bayesian decision theory

Bayesian decision theory / continuous features (3/4)

two-category classification

- α_1 : deciding ω_1
- α_2 : deciding ω_2
- $\alpha_{ij} = \alpha(\omega_i | \omega_j)$
- loss incurred for deciding ω_i when the true state of nature is ω_j

conditional risk

- $R(\alpha_1 | x) = \alpha_{11}P(\omega_1 | x) + \alpha_{12}P(\omega_2 | x)$
- $R(\alpha_2 | x) = \alpha_{21}P(\omega_1 | x) + \alpha_{22}P(\omega_2 | x)$

rule

- if $R(\alpha_1 | x) < R(\alpha_2 | x)$ action α_1 : “decide ω_1 ” is taken
- this results in the equivalent rule:
decide ω_1 if:
 $(\alpha_{21} - \alpha_{11}) P(x | \omega_1) > (\alpha_{12} - \alpha_{22}) P(x | \omega_2)$ and decide ω_2 otherwise

Bayesian decision theory

Bayesian decision theory / continuous features (4/4)

likelihood

- the preceding rule is equivalent to the following rule:

$$\frac{p(x | \omega_1)}{p(x | \omega_2)} > \frac{\lambda_{12} + \lambda_{22}}{\lambda_{21} + \lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)}$$

- then take action ω_1 (decide ω_1), otherwise action ω_2 (decide ω_2)

likelihood ratio

- likelihood ratio for class-conditional probability density function
- decision boundary determined by threshold λ_{ω_1}

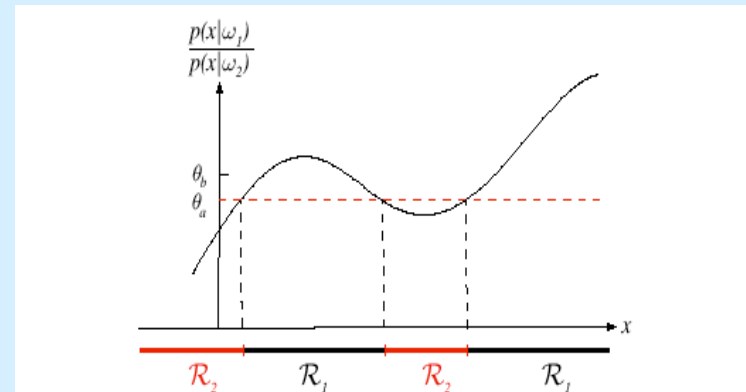


FIGURE 2.3. The likelihood ratio $p(x|\omega_1)/p(x|\omega_2)$ for the distributions shown in Fig. 2.1. If we employ a zero-one or classification loss, our decision boundaries are determined by the threshold θ_a . If our loss function penalizes miscategorizing ω_2 as ω_1 patterns more than the converse, we get the larger threshold θ_b , and hence \mathcal{R}_1 becomes smaller. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Bayesian decision theory

index of contents

introduction to ML

Bayesian decision theory

- introduction
- Bayesian decision theory / continuous features
 - > minimum-error-rate classification
- classifiers, discriminant functions and decision surfaces
- the normal density
- discriminant functions for the normal density
- Bayesian decision theory / discrete features

Bayesian decision theory

minimum-error-rate classification (1/2)

actions are decisions on classes

- if action α_i is taken and the true state of nature is ω_j then:
the decision is correct if $i = j$ and in error if $i \neq j$

decision rule

- seek a decision rule that minimizes the *probability of error*
which is the *error rate*

introduction of the zero-one loss function

$$L(\alpha_i, \omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \quad i, j = 1, \dots, c$$

- therefore, the conditional risk is:

$$\begin{aligned} R(\alpha_i | x) &= \sum_{j=1}^{j=c} L(\alpha_i | \omega_j) P(\omega_j | x) \\ &= \sum_{j \neq i} P(\omega_j | x) = 1 - P(\omega_i | x) \end{aligned}$$

Bayesian decision theory

minimum-error-rate classification (2/2)

minimizing the risk requires maximization of $P(\omega_i | x)$

for minimum error rate

- Decide ω_i if $P(\omega_i | x) > P(\omega_j | x) \quad \omega_j \neq i$

Bayesian decision theory

index of contents

introduction to ML

Bayesian decision theory

- introduction
- Bayesian decision theory / continuous features
- minimum-error-rate classification
- > classifiers, discriminant functions and decision surfaces
- the normal density
- discriminant functions for the normal density
- Bayesian decision theory / discrete features

Bayesian decision theory

classifiers, discriminant functions and decision surfaces (1/4)

the multi-category case

- set of discriminant functions $g_i(\mathbf{x}), i = 1, \dots, c$
- the classifier assigns a feature vector \mathbf{x} to class ω_i if:
 $g_i(\mathbf{x}) > g_j(\mathbf{x}) \quad \forall j \neq i$

functional structure of general statistical pattern classifier

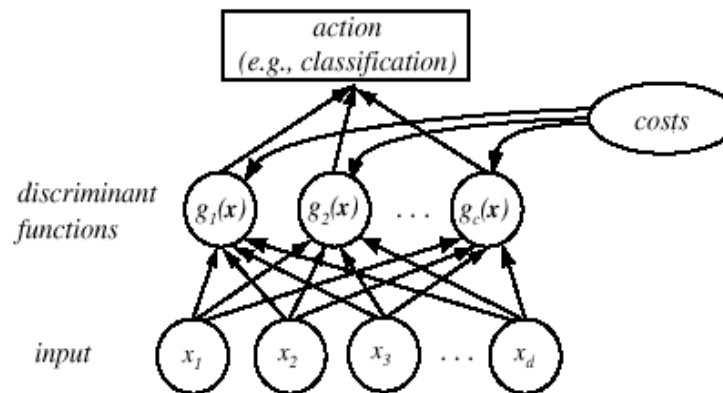


FIGURE 2.5. The functional structure of a general statistical pattern classifier which includes d inputs and c discriminant functions $g_i(\mathbf{x})$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly. The arrows show the direction of the flow of information, though frequently the arrows are omitted when the direction of flow is self-evident. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Bayesian decision theory

classifiers, discriminant functions and decision surfaces (2/4)

Bayes classifier can be represented in this way

- $g_i(\mathbf{x}) = -R(\omega_i | \mathbf{x})$ minimum conditional risk
- $g_i(\mathbf{x}) = P(\omega_i | \mathbf{x})$ minimum error-rate

the selection of a discriminant function is not unique

for minimum error classifier, one may choose:

$$g_i(\mathbf{x}) = p(\mathbf{x} | \omega_i) P(\omega_i)$$

$$g_i(\mathbf{x}) = \ln P(\mathbf{x} | \omega_i) + \ln P(\omega_i)$$

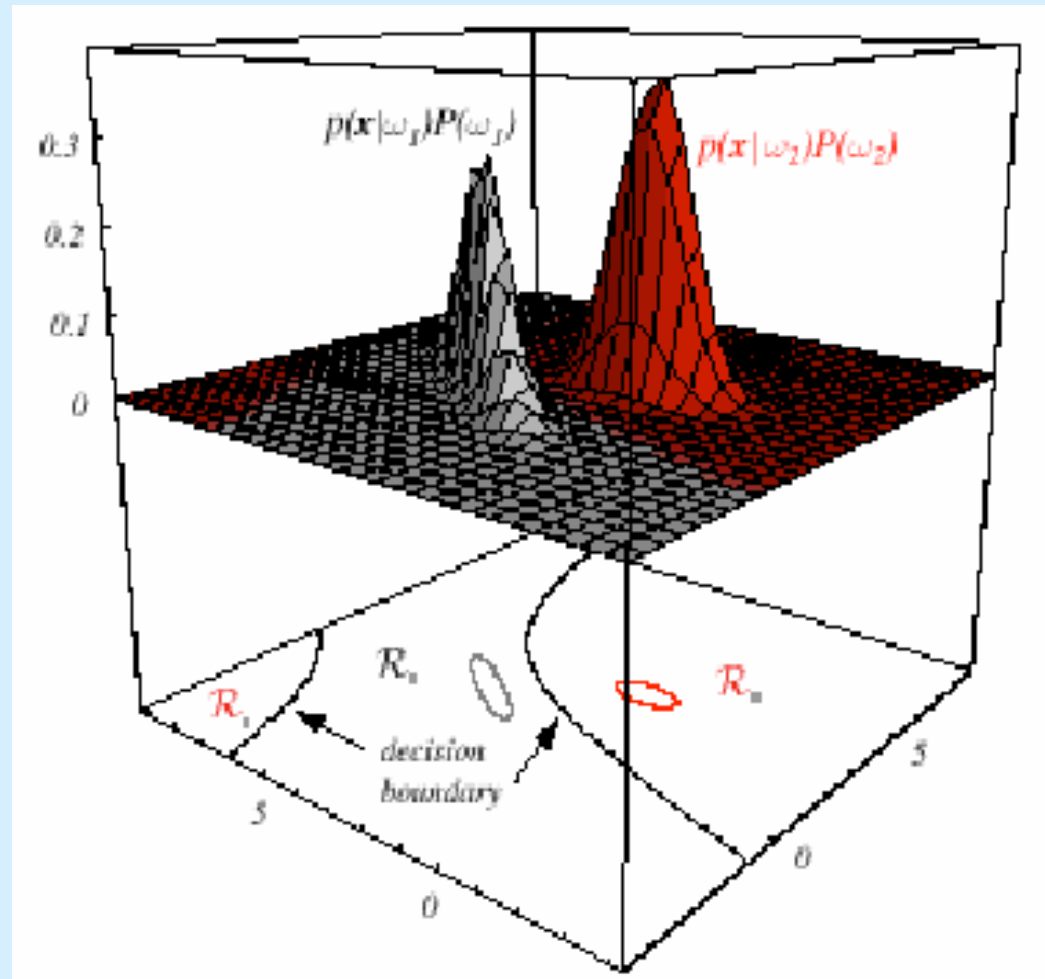
discriminant functions

- discriminant functions can be in different forms, but the effect of decision rules is the same: decision boundaries
- decide \mathbf{x} is in R_i if: $g_i(\mathbf{x}) > g_j(\mathbf{x}) \quad \forall j \neq i$

Bayesian decision theory

classifiers, discriminant functions and decision surfaces (4/4)

two-dimensional two-category classifier



Bayesian decision theory

index of contents

introduction to ML

Bayesian decision theory

- introduction
- Bayesian decision theory / continuous features
- minimum-error-rate classification
- classifiers, discriminant functions and decision surfaces
- > the normal density
- discriminant functions for the normal density
- Bayesian decision theory / discrete features

Bayesian decision theory

the normal density (1/2)

univariate normal density

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

- $\mu = (\mu_1, \mu_2, \dots, \mu_d)^t$
- $\sigma^2 =$ expected squared deviation or variance

multivariate normal density in d dimensions

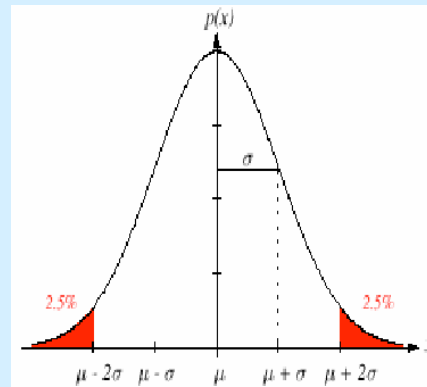
$$P(x) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^t \Sigma^{-1} (x-\mu)\right]$$

- $x = (x_1, x_2, \dots, x_d)^t$ (t stands for the transpose vector form)
- $\Sigma = d$ -by- d covariance matrix
- $|\Sigma|$ and Σ^{-1} are determinant and inverse respectively

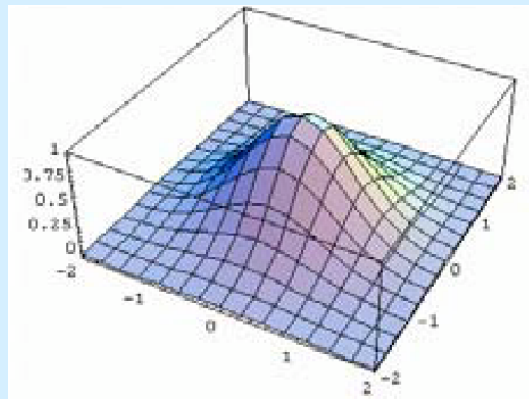
Bayesian decision theory

the normal density (2/2)

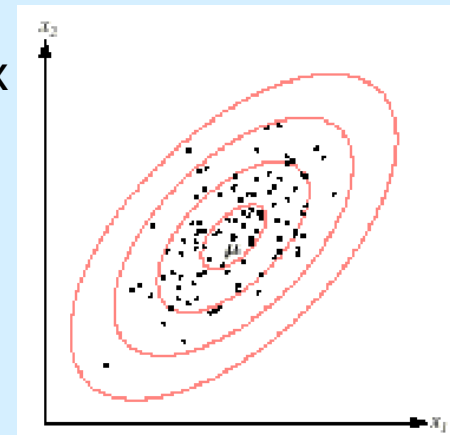
univariate normal distribution



multivariate normal distribution



covariance matrix
determines the
shape of
Gaussian curve



Bayesian decision theory

index of contents

introduction to ML

Bayesian decision theory

- introduction
- Bayesian decision theory / continuous features
- minimum-error-rate classification
- classifiers, discriminant functions and decision surfaces
- the normal density
- > discriminant functions for the normal density
- Bayesian decision theory / discrete features

Bayesian decision theory

discriminant functions for the normal density (1/4)

minimum error-rate classification can be achieved by the discriminant function

$$g_i(x) = \ln P(x | \omega_i) + \ln P(\omega_i)$$

case of multivariate normal density, discriminant function is

$$g_i(x) = -\frac{1}{2} (x - \mu_i)^t \Sigma_i^{-1} (x - \mu_i) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_i| + \ln P(\omega_i)$$

case 1: $\Sigma_i = \Sigma$ (independence, equal Σ)

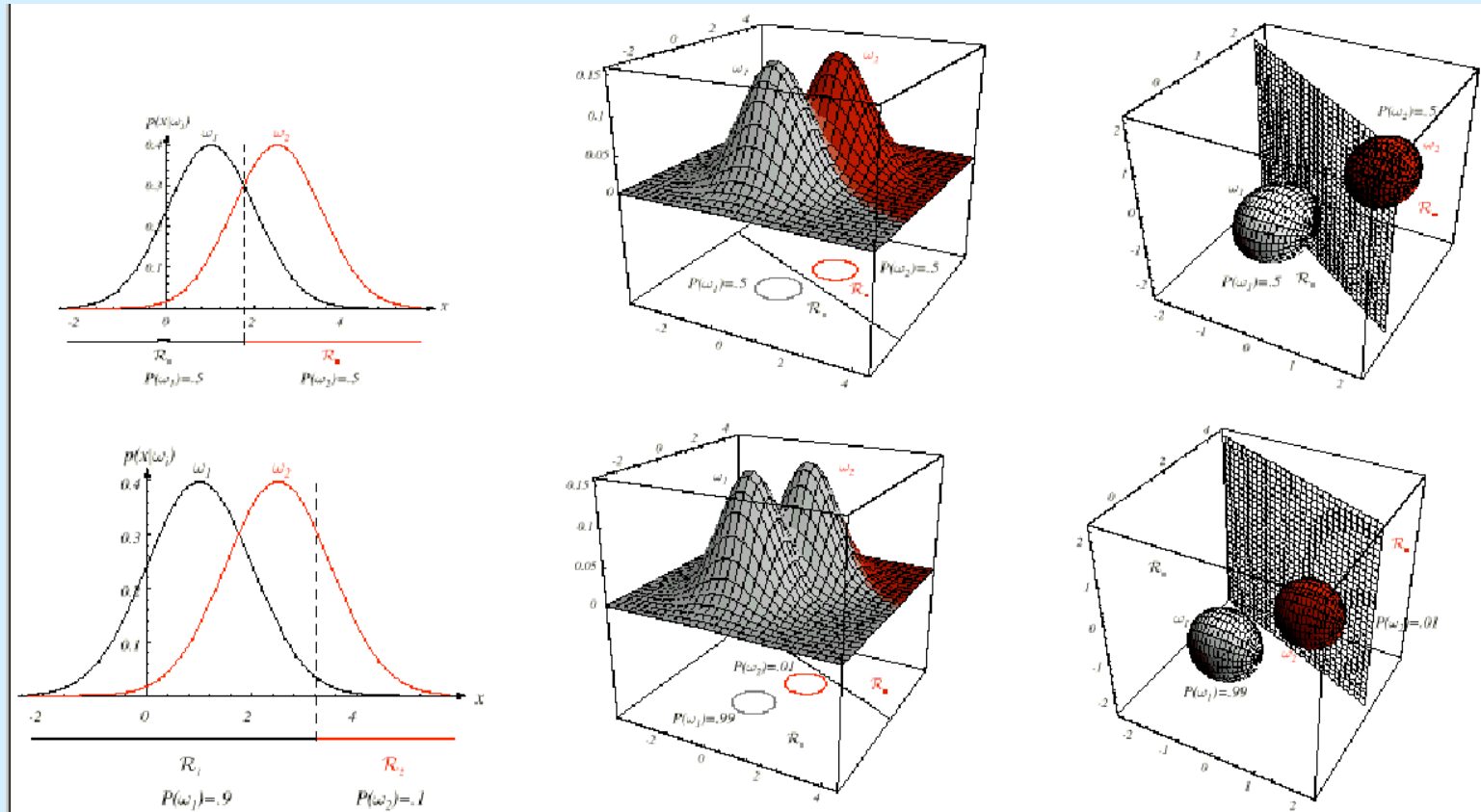
$$g_i(x) = -\frac{1}{2\Sigma^2} [x^t x - 2\mu_i^t x + \mu_i^t \mu_i] + \ln P(\omega_i)$$

$$g_i(x) = \frac{\mu_i^t}{\Sigma^2} x - \frac{1}{2\Sigma^2} \mu_i^t \mu_i + \ln P(\omega_i)$$

Bayesian decision theory

discriminant functions for the normal density (2/4)

case 1: $\sigma_i = \sigma_1$ (independence, equal σ)

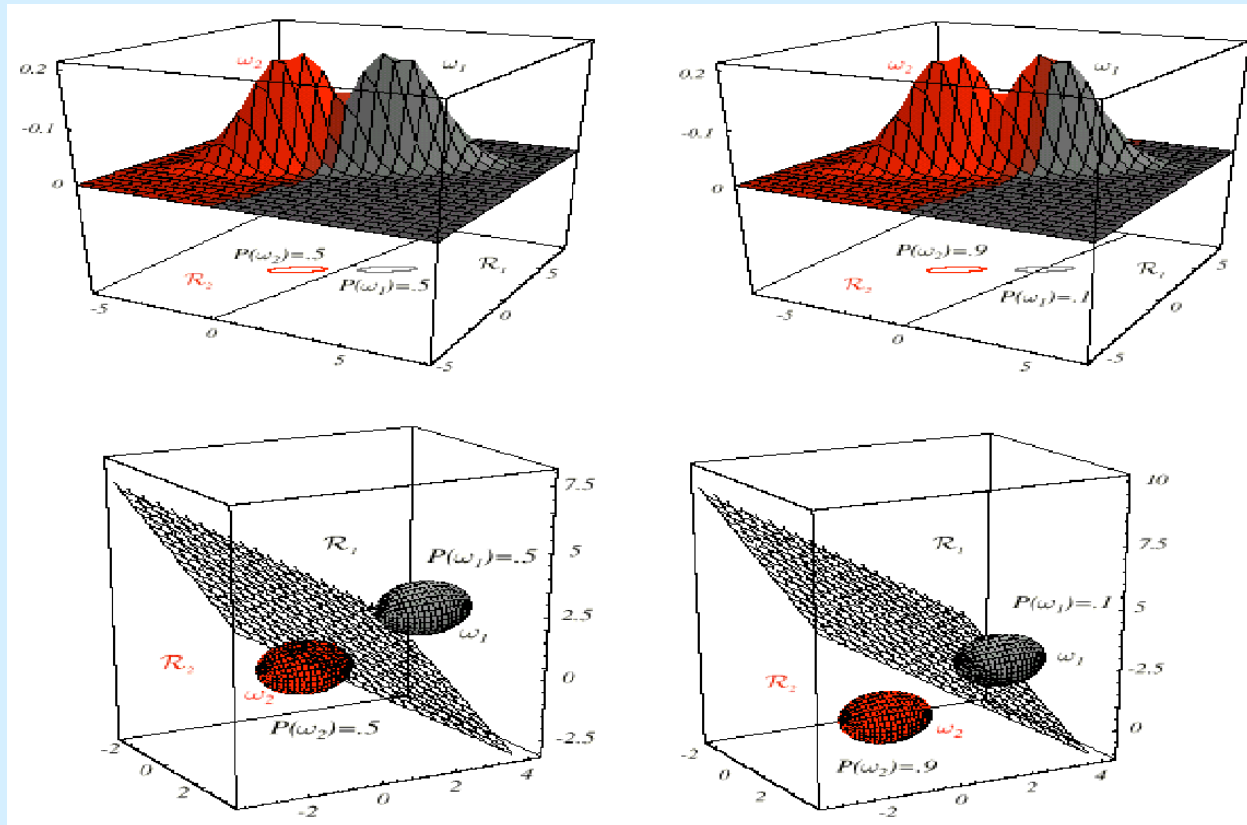


Bayesian decision theory

discriminant functions for the normal density (3/4)

case 2: $\Sigma_i = \Sigma$ (covariance of all classes are identical but arbitrary!)

$$g_i(\vec{x}) = -\frac{1}{2}[(\vec{x} - \vec{\mu}_i)' \Sigma^{-1}(\vec{x} - \vec{\mu}_i)] + \log P(\omega_i)$$

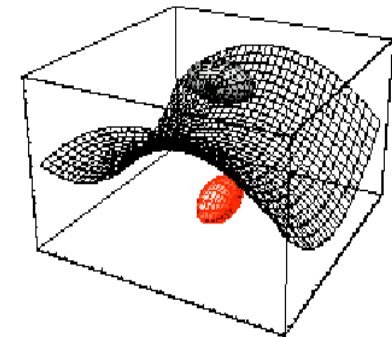
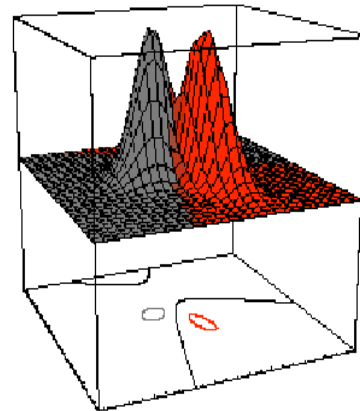
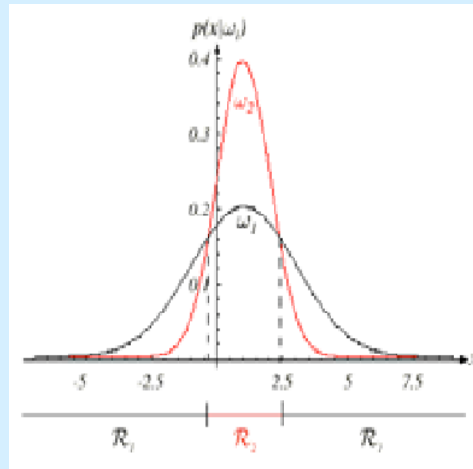


Bayesian decision theory

discriminant functions for the normal density (4/4)

case 3: $\Sigma_i = \text{arbitrary}$

$$g_i(x) = -\frac{1}{2}(x - \mu_i)^t \Sigma_i^{-1} (x - \mu_i) - \frac{d}{2} \ln |\Sigma_i| - \frac{1}{2} \ln P(\omega_i)$$



Bayesian decision theory

index of contents

introduction to ML

Bayesian decision theory

- introduction
- Bayesian decision theory / continuous features
- minimum-error-rate classification
- classifiers, discriminant functions and decision surfaces
- the normal density
- discriminant functions for the normal density
- > Bayesian decision theory / discrete features

Bayesian decision theory

Bayesian decision theory / discrete features (1/2)

components of x are binary or integer valued, x can take only one of m discrete values

- V_1, V_2, \dots, V_m

$\int p(x | \omega_i) dx$ replaced by $\sum_x P(x | \omega_i)$

- fundamental Bayes decision rule remains the same

case of independent binary features in 2 category problem

- $x = [x_1, x_2, \dots, x_d]^t$ where each x_i is either 0 or 1, with probabilities

$$> p_i = P(x_i = 1 | \omega_1)$$

$$> q_i = P(x_i = 1 | \omega_2)$$

Bayesian decision theory

Bayesian decision theory / discrete features (2/2)

the discriminant function in this case is

$$g(x) = \sum_{i=1}^d w_i x_i + w_0$$

where:

$$w_i = \ln \frac{p_i(1 - q_i)}{q_i(1 - p_i)} \quad i = 1, \dots, d$$

and:

$$w_0 = \sum_{i=1}^d \ln \frac{1 - p_i}{1 - q_i} + \ln \frac{P(\Omega_1)}{P(\Omega_2)}$$

decide Ω_1 if $g(x) > 0$ and Ω_2 if $g(x) \leq 0$