Pattern Recognition

Basic ideas, Bayes' classifier

CS4243
Dr. Terence Sim

Outline

- Example
  - Gender classification
- Basic Ideas
  - Design Cycle
  - Important Questions
- Bayes' Classifier
  - Simple case
  - Generalization

Example

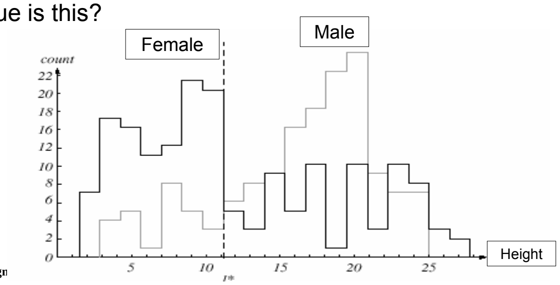
- Gender identification
  - <demo>
- What is it sensing?
- How is it making a decision?
- How well is it doing?

What is it sensing?

- Pattern recognition has many applications:
  - DNA identification: genetic materical → identity
  - Speech recognition: audio → text
  - Face detection: image → location of faces
  - Fingerprint identification: image → identity
  - Optical character recognition (OCR): image → text
- In our case: ??? → {male, female}
  - What is the ???

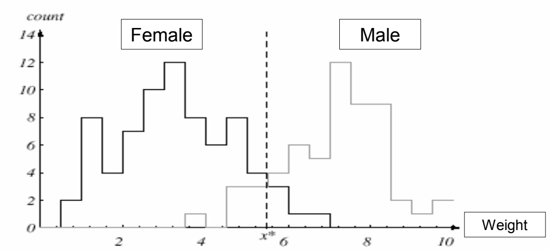
Features

- What features to use?
- Try height
  - Idea: males are generally taller than females
  - Therefore, a large value of height implies male
  - How true is this?



Features

- Height seems to be a poor feature alone.
- Try weight:
  - Idea: males generally weigh more than females
  - Again, how true is this?

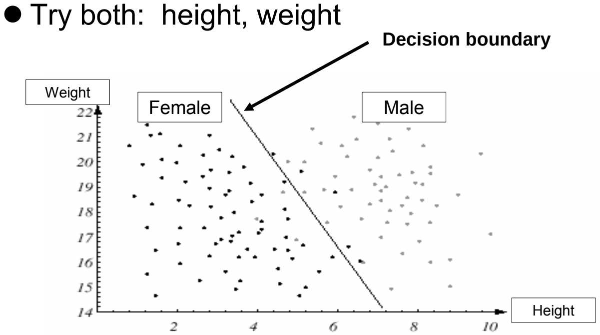


Decision boundary

- Boundary between 2 classes: x*
- Decision rule:
  - If x < x* then decide *Female*
  - Else If x > x* then decide *Male*
  - Else flip a coin

Features

- Try both: height, weight



2 features

- x = [height, weight]^T
- Decision boundary is a line
- Decision rule:
  - If x lies above line, then decide *Male*
  - Else If x lies below line, then decide *Female*
  - Else flip a coin
- But still some errors …

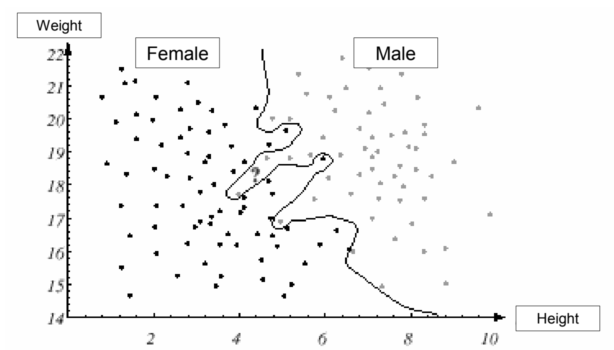
CS4243: Pattern Recognition School *of* Computing 9

More features?

- We might 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”
- Ideally, the best decision boundary should be the one which provides an optimal performance such as in the following figure:

CS4243: Pattern Recognition School *of* Computing 10

Perfect Decision Boundary?



CS4243: Pattern Recognition School *of* Computing 11

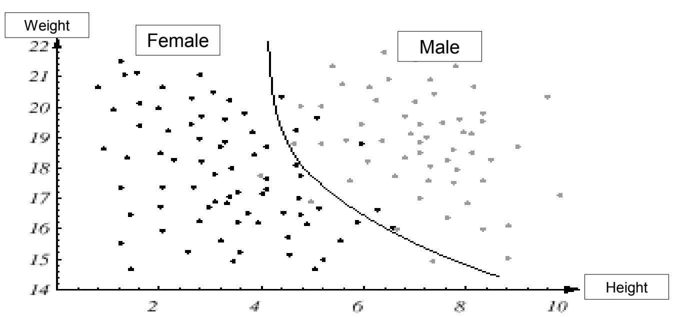
Generalization

- However, our satisfaction is premature because the central aim of designing a classifier is to correctly classify ***novel*** input

Issue of generalization!

CS4243: Pattern Recognition School *of* Computing 12

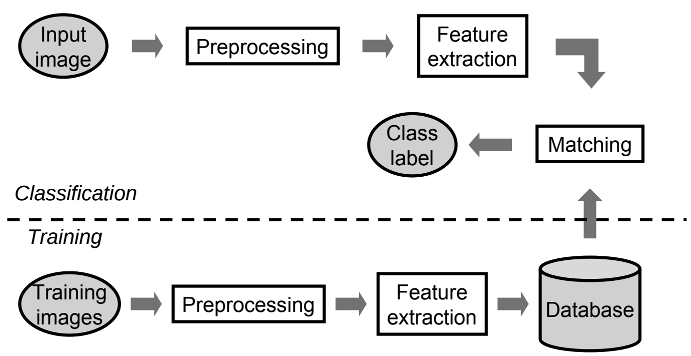
Non-linear boundary



Basic Ideas: Definition

- Let S = $\{\omega_1, \omega_2 \ldots \omega_C\}$ be the set of pre-defined *C* classes
  - e.g. {male, female}
- Let ***x*** be the feature vector in $\boldsymbol{R}^n$
- Classifier is a function $f : \boldsymbol{R}^n \rightarrow S$
  - We say that a classifier *assigns a class label* to the feature vector (pattern)

Basic Ideas: Typical Image PR pipeline

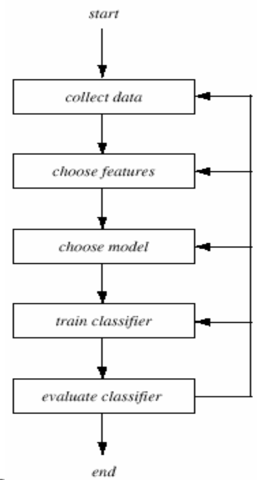


3 Important Questions

- What features are best?
  - _______ knowledge
  - Ask the _________
  - Guess
  - ______ from ______ data
- Given features, how to design classifier?
  - What type of classifier?
  - How to find decision boundary?
- How good is the classifier?
  - How to evaluate performance?

The Design Cycle

- Data collection
- Feature Choice
- Model Choice
- Training
- Evaluation
- Computational Complexity



Issues

- 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.

Issues

- Model Choice
  - Bayes' Classifier, K-nearest neighbor, Fisher's Linear Discriminant, Neural Networks, Support Vector Machines, Decision Trees, etc.
- Training
  - Use data to determine the classifier. Many different procedures for training classifiers and choosing models
  - How do we know we have trained enough?
  - Can we overtrain?

Issues

- Evaluation
  - Measure the error rate
  - Where to get test data?
- Computational Complexity
  - What is the trade-off between computational ease and performance?
  - How does classifier scale as number of features or classes increases?
  - How much storage required?

Bayes' Classifier

Theoretically Optimal Classifier

Statistical PR

- Suppose you have no observation
  - How to classify?
  - You only know the prior probabilities, e.g. males in population = 50.85%
- Decision rule with only the prior information
  - Decide ω_1 if $P(\omega_1) > P(\omega_2)$ otherwise decide ω_2

Bayes' Classifier

- Now suppose you observed X
- How to classify?
- Bayes' classifier says: ***Maximum A Posteriori***

$$\omega^* = \arg\max_{\omega_j} P(\omega_j \mid x)$$

- That is, assign x to label ω_j such that $P(\omega_j \mid x)$ is largest among all $P(\omega_i \mid x)$

Bayes' Classifier

- Bayes' Rule: $$P(A \mid B) = \frac{P(B \mid A) \times P(A)}{P(B)}$$
- So

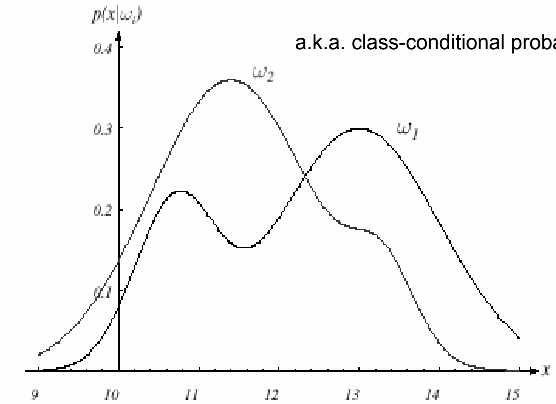
$$\omega^* = \arg\max_{\omega_j} \underbrace{P(\omega_j \mid x)}_{\textit{Posterior}}$$

$$= \arg\max_{\omega_j} \frac{\overbrace{P(x \mid \omega_j)}^{\textit{Likelihood}} \bullet \overbrace{P(\omega_j)}^{\textit{Prior}}}{\underbrace{P(x)}_{\textit{Evidence}}}$$

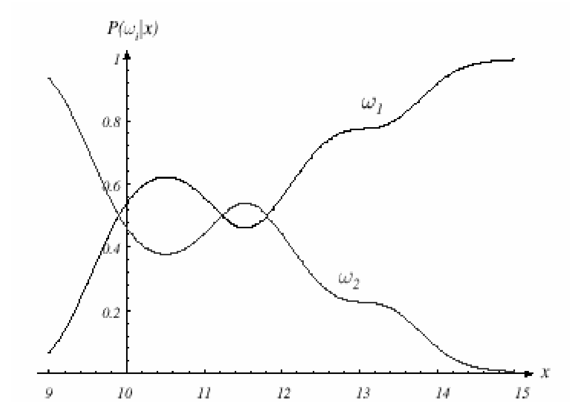
$$= \arg\max_{\omega_j} P(x \mid \omega_j) \bullet P(\omega_j)$$

Likelihood: learn from training data

a.k.a. class-conditional probability



Maximum A Posteriori



Special case

- Equal priors $P(\omega_1) = P(\omega_2) = \cdots = P(\omega_C) = \frac{1}{C}$
- Then $\omega^* = \arg\max_{\omega_j} P(x \mid \omega_j) \bullet \cancel{P(\omega_j)}$

Maximum Likelihood

Special case: only 2 classes

- Decide ω_1 *if* $P(\omega_1 \mid x) > P(\omega_2 \mid x)$; otherwise decide ω_2

Alternatively:

- Decide ω_1 *if* $g(x) > 0$ otherwise decide ω_2
- Where $g(x) = P(\omega_1 \mid x) - P(\omega_2 \mid x)$
  - *g(x) is called a* ***Discriminant Function***

Generalizing Bayes'

- Allowing actions other than classification primarily allows the possibility of rejection.
- Refusing to make a decision in close or bad cases!
- The loss function states how costly each action taken is.

Generalizing Bayes'

Let $\{\omega_1, \omega_2, \ldots, \omega_c\}$ be the set of C classes

Let $\{\alpha_1, \alpha_2, \ldots, \alpha_a\}$ be the set of possible actions

Let $\lambda(\alpha_i \mid \omega_j)$ be the loss incurred for taking

action α_i when the class is ω_j

Overall Risk

R = *Sum of all* $\underbrace{R(\alpha_i \mid x)}_{\textbf{Conditional risk}}$ *for i = 1,…,a*

Minimizing R $\Longleftrightarrow$ Minimizing $R(\alpha_i \mid x)$ *for i = 1,…, a*

$$R(\alpha_i \mid x) = \sum_{j=1}^{j=C} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid x)$$

for i = 1,…,a

Bayes' Risk

Select the action α_i for which $R(\alpha_i \mid x)$ is minimum

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

2-class classification

α_1 : *deciding* ω_1

α_2 : *deciding* ω_2

$\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$

loss incurred for deciding ω_i when the true class is ω_j

Conditional risk:

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

Decision Rule

Our rule is the following:

if $R(\alpha_1 \mid x) < R(\alpha_2 \mid x)$

action α_1: "decide ω_1" is taken

This results in the equivalent rule :

decide ω_1 if:

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

and decide ω_2 otherwise

Likelihood Ratio

The preceding rule is equivalent to the following rule:

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

Then take action α_1 (decide ω_1)

Otherwise take action α_2 (decide ω_2)

Note: right-hand side independent of input *x*

Note: if $\lambda_{21} = \lambda_{12} = 1$ and $\lambda_{11} = \lambda_{22} = 0$, then MAP!

Summary

- Pattern Recognition or Classification means assigning class label to input pattern.
- Choosing features is an art!
- Bayes' Classifier is theoretically optimum
  - Provided you know the priors and likelihoods!
  - It takes into account cost (loss) of making decisions.
  - Bayes' is an example of Statistical PR.
- Next week: other PR methods