Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Supervised Learning: Linear Methods (1/2)

Applied Multivariate Statistics – Spring 2012

Overview

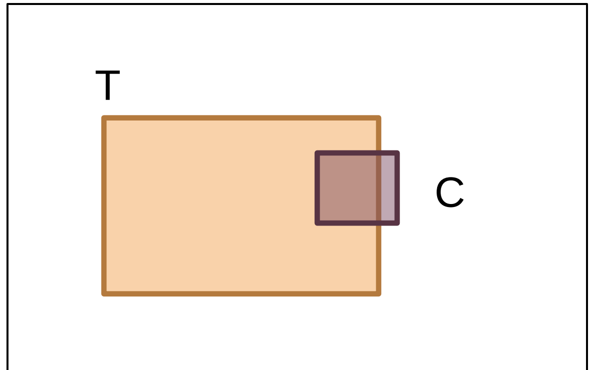
- Review: Conditional Probability
- LDA / QDA: Theory
- Fisher's Discriminant Analysis
- LDA: Example
- Quality control: Testset and Crossvalidation
- Case study: Text recognition

Conditional Probability

T: Med. Test positive

C: Patient has cancer

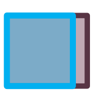
Sample space



(Marginal) Probability:
P(T), P(C)

New sample space:
People with cancer

P(T|C)
large



Conditional Probability:
P(T|C), P(C|T)

New sample space:
People with pos. test

P(C|T)
small

Bayes Theorem:

posterior

$$\boxed{P(C|T)} = \frac{\boxed{P(T|C)}\boxed{P(C)}}{P(T)}$$

prior

Class conditional probability

One approach to supervised learning

$$P(C|X) = \frac{P(C)P(X|C)}{P(X)} \sim P(C)P(X|C)$$

Find some estimate

Prior / prevalence:
Fraction of samples
in that class

Assume:
$X|C \sim N(\mu_c, \Sigma_c)$

Bayes rule:

Choose class where P(C|X) is maximal

(rule is “optimal” if all types of error are equally costly)

Special case: Two classes (0/1)

- choose c=1 if P(C=1|X) > 0.5 or
- choose c=1 if posterior odds P(C=1|X)/P(C=0|X) > 1

In Practice: Estimate $P(C), \mu_C, \Sigma_C$

QDA: Doing the math…

$$\frac{1}{\sqrt{(2\pi)^d|\Sigma_C|}}\exp\left(-\frac{1}{2}(x-\mu_c)^T\Sigma_C^{-1}(x-\mu_c)\right)$$

- $P(C|X) \sim P(C)P(X|C)$
- Use the fact: $\max P(C|X) \Leftrightarrow \max(\log(P(C|X)))$
- $\delta_c(x) = \log(P(C|X)) = \log(P(C)) + \log(P(X|C)) =$

$$= \underbrace{\log(P(C))}_{\text{Prior}} - \frac{1}{2}\underbrace{\log(|\Sigma_C|)}_{\text{Additional term}} - \frac{1}{2}\underbrace{(x-\mu_C)^T\Sigma_C^{-1}(x-\mu_C)}_{\text{Sq. Mahalanobis distance}} + c$$

- Choose class where $\delta_c(x)$ is maximal
- Special case: Two classes
  Decision boundary: Values of x where $\delta_0(x) = \delta_1(x)$ is quadratic in x

- **Quadratic Discriminant Analysis (QDA)**

Simplification

- Assume same covariance matrix in all classes, i.e.

$$X|C \sim N(\mu_c, \Sigma)$$

Fix for all classes

- $\delta_c(x) = \log\big(P(C)\big) - \frac{1}{2}\log(|\Sigma|) - \frac{1}{2}(x - \mu_C)^T\Sigma^{-1}(x - \mu_C) + c =$

Prior

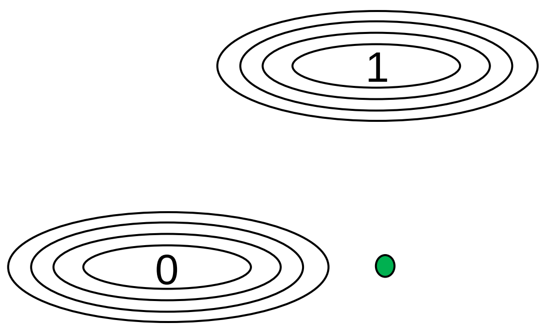
$$= \log\big(P(C)\big) - \frac{1}{2}(x - \mu_C)^T\Sigma^{-1}(x - \mu_C) + d =$$

Sq. Mahalanobis distance

$$(= \log\big(P(C)\big) + x^T\Sigma^{-1}\mu_C - \frac{1}{2}\,\mu_C^T\,\Sigma^{-1}\mu_C)$$

Decision boundary is linear in x

- **Linear Discriminant Analysis (LDA)**



Classify to which class (assume equal prior)?

- Physical distance in space is equal
- Classify to class 0, since Mahal. Dist. is smaller

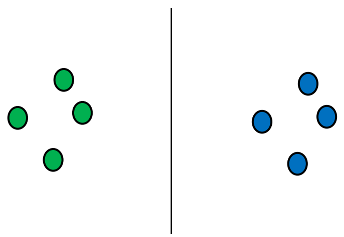
LDA vs. QDA

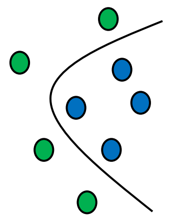
+ Only few parameters to estimate; accurate estimates

- Inflexible
(linear decision boundary)

- Many parameters to estimate; less accurate

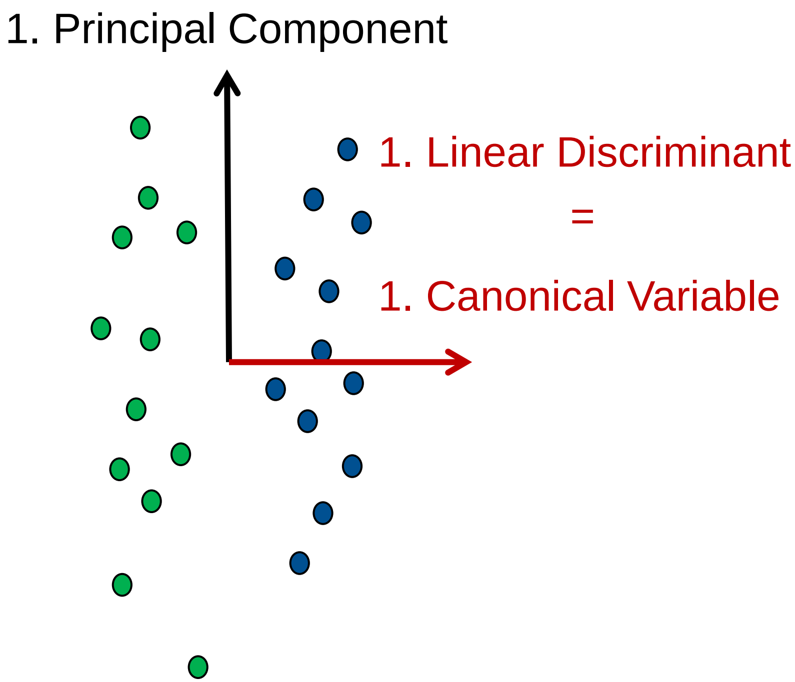
+ More flexible
(quadratic decision boundary)





Fisher's Discriminant Analysis: Idea

Find direction(s) in which groups are separated best

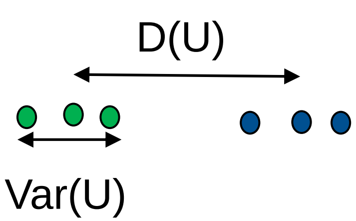


- Class Y, predictors $X = (X_1, \dots, X_d)$
  $\rightarrow U = w^T X$
- Find w so that groups are separated along U best
- Measure of separation: Rayleigh coefficient

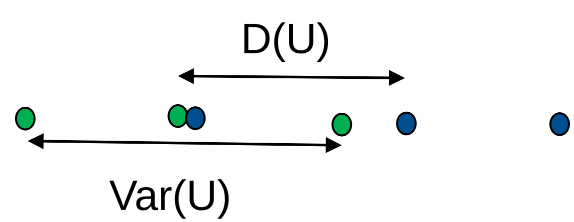
$$J(w) = \frac{D(U)}{Var(U)}$$

where $D(U) = \big(E(U|Y=0) - E(U|Y=1)\big)^2$

- $E[X|Y=j] = \mu_j, Var(X|Y=j) = \Sigma$
  $\Rightarrow E[U|Y=j] = w^T\mu_j, V(U) = w^T\Sigma \mathrm{w}$
- Concept extendable to many groups



$J(w)$ large



$J(w)$ small

LDA and Linear Discriminants

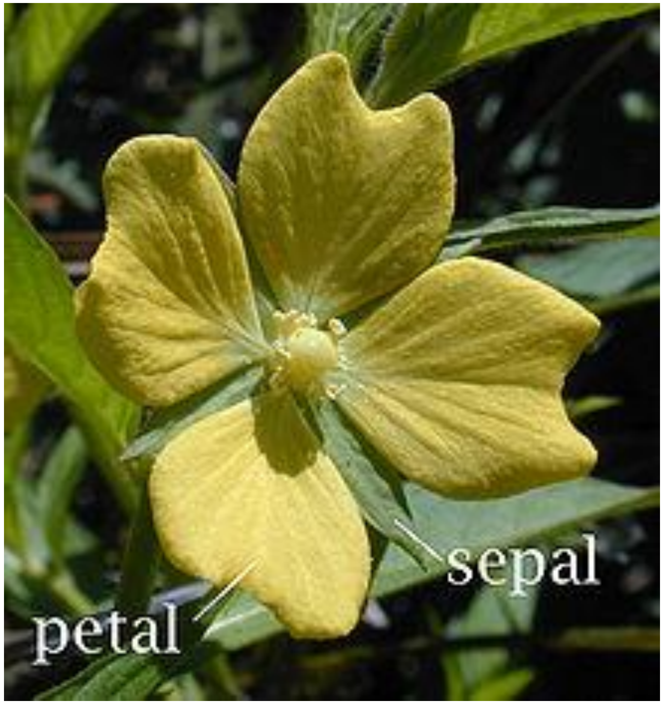
- \- Direction with largest J(w): 1. Linear Discriminant (LD 1)
  \- orthogonal to LD1, again largest J(w): LD 2
  \- etc.
- At most: min(Nmb. dimensions, Nmb. Groups -1) LD’s
  e.g.: 3 groups in 10 dimensions – need 2 LD’s
- Computed using Eigenvalue Decomposition or Singular Value Decomposition
  Proportion of trace: Captured % of variance between group means for each LD
- R: Function «lda» in package MASS does LDA and computes linear discriminants (also «qda» available)

Example: Classification of Iris flowers

Iris setosa

Iris versicolor

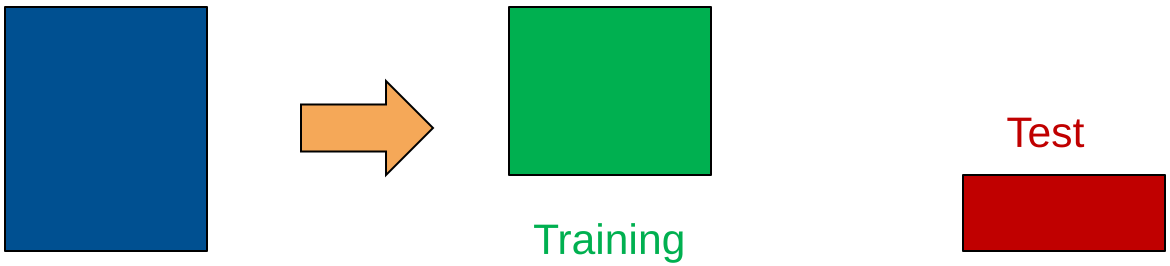
Iris virginica



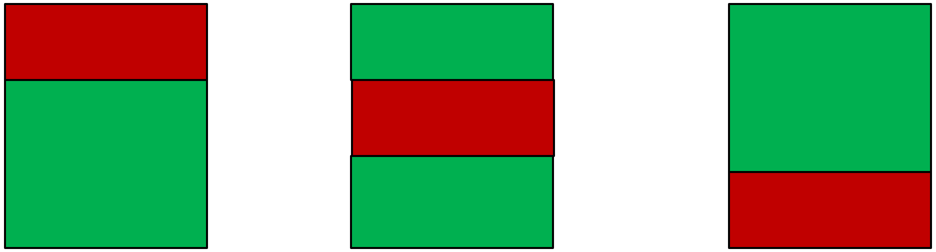
Classify according to sepal/petal length/width

Quality of classification

- Use training data also as test data: Overfitting
  Too optimistic for error on new data
- Separate test data



- Cross validation (CV; e.g. “leave-one-out cross validation):
  Every row is the test case once, the rest in the training data



Measures for prediction error

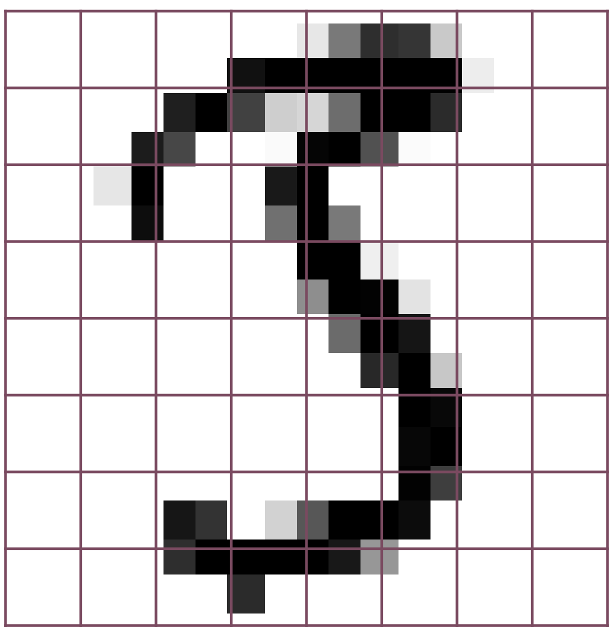
- Confusion matrix (e.g. 100 samples)

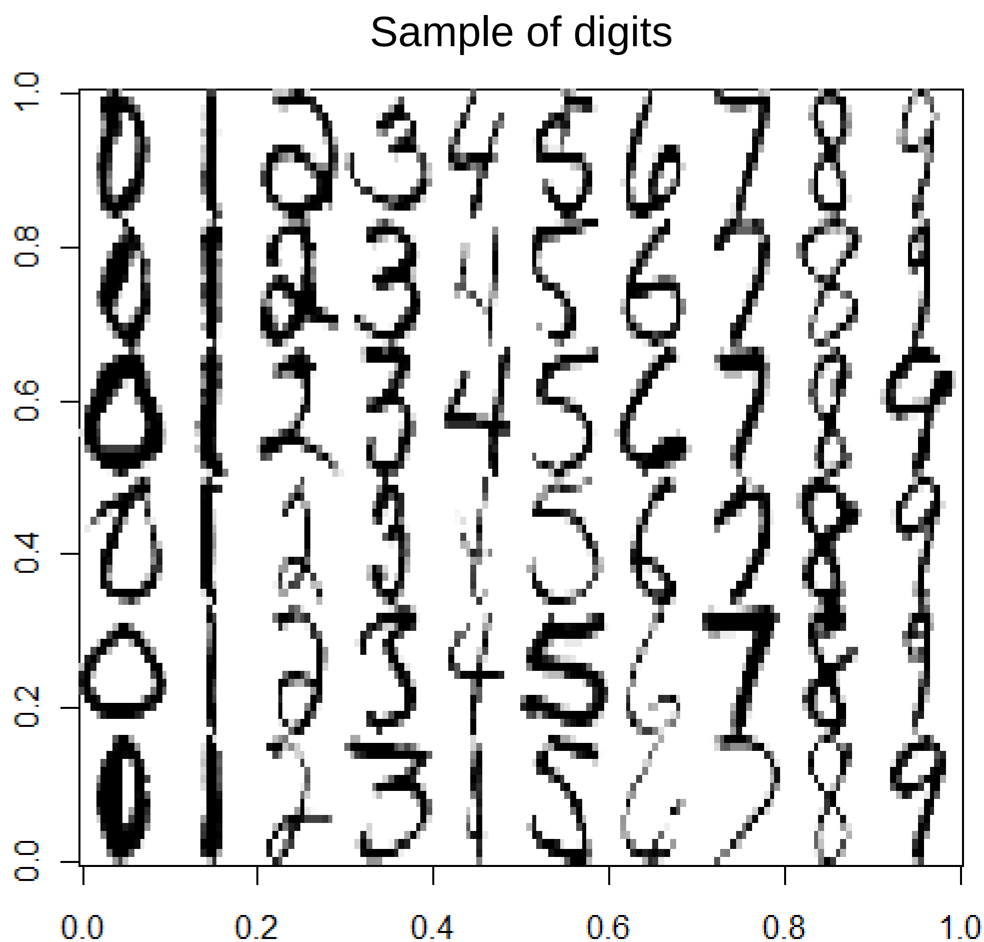
	Truth = 0	Truth = 1	Truth = 2
Estimate = 0	23	7	6
Estimate = 1	3	27	4
Estimate = 2	3	1	26

- Error rate:
  1 – sum(diagonal entries) / (number of samples) =
  = 1 – 76/100 = 0.24
- We expect that our classifier predicts 24% of new observations incorrectly (this is just a rough estimate)

Example: Digit recognition

- 7129 hand-written digits
- Each (centered) digit was put in a 16*16 grid
- Measure grey value in each part of the grid, i.e. 256 grey values


Example with 8*8 grid



Concepts to know

- Idea of LDA / QDA
- Meaning of Linear Discriminants
- Cross Validation
- Confusion matrix, error rate

R functions to know

- lda