Supervised Learning: Linear Methods (1/2)

Applied Multivariate Statistics – Spring 2013
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

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

New sample space: People with cancer
P(T|C) large

New sample space: People with pos. test
P(C|T) small

Bayes Theorem:

\[ P(C|T) = \frac{P(T|C)P(C)}{P(T)} \]

Class conditional probability
posterior
prior
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...

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

- \( P(C|X) \sim P(C)P(X|C) \)
- Use the fact: \( \max P(C|X) \Leftrightarrow \max(\log(P(C|X))) \)
- \( \delta_c(x) = \log(P(C|X)) = \log(P(C)) + \log(P(X|C)) = \log(P(C)) - \frac{1}{2} \log(|\Sigma_c|) - \frac{1}{2}(x - \mu_c)^T \Sigma^{-1}_c (x - \mu_c) + c \)
  - Prior
  - Additional term
  - Sq. Mahalanobis distance
- 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.**
  \[ X|C \sim N(\mu_c, \Sigma) \]

- **Linear Discriminant Analysis (LDA)**
  - Decision boundary is linear in x
  - Fix for all classes
  - Prior
  - Square Mahalanobis distance
  - Physical distance in space is equal
  - Classify to class 0, since Mahal. Dist. is smaller

\[ \delta_c(x) = \log(P(C)) - \frac{1}{2} \log(|\Sigma|) - \frac{1}{2} (x - \mu_c)^T \Sigma^{-1} (x - \mu_c) + c = \]

\[ = \log(P(C)) - \frac{1}{2} (x - \mu_c)^T \Sigma^{-1} (x - \mu_c) + d = \]

\[ (= \log(P(C)) + x^T \Sigma^{-1} \mu_c - \frac{1}{2} \mu_c^T \Sigma^{-1} \mu_c) \]
LDA vs. QDA

LDA:
+ Only few parameters to estimate; accurate estimates
- Inflexible (linear decision boundary)

QDA:
- 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

1. Principal Component

1. Linear Discriminant = 1. Canonical Variable

Class Y, predictors $X = (X_1, ..., X_d)$

$\rightarrow U = w^T X$

Find $w$ so that groups are separated along $U$ best

Measure of separation: Rayleigh coefficient

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

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

$E[X|Y = j] = \mu_j, \text{Var}(X|Y = j) = \Sigma$

$\Rightarrow E[U|Y = j] = w^T\mu_j, V(U) = w^T\Sigma w$

Concept extendable to many groups
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: \( \min(\text{Nmb. dimensions}, \text{Nmb. Groups} - 1) \) 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)

<table>
<thead>
<tr>
<th></th>
<th>Truth = 0</th>
<th>Truth = 1</th>
<th>Truth = 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate = 0</td>
<td>23</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>Estimate = 1</td>
<td>3</td>
<td>27</td>
<td>4</td>
</tr>
<tr>
<td>Estimate = 2</td>
<td>3</td>
<td>1</td>
<td>26</td>
</tr>
</tbody>
</table>

- Error rate:
  \[ 1 - \frac{\text{sum(diagonal entries)}}{\text{(number of samples)}} = 1 - \frac{76}{100} = 0.24 \]

- 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
Concepts to know

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

- lda