Classification of Data Sets
Robert Stengel
Robotics and Intelligent Systems MAE 345,
Princeton University, 2013

Learning Objectives

• Unsupervised learning
  – Cluster analysis
    • Patterns, Clumps, and Joining

• Supervised learning
  – Graph/tree search
  – Hypothesis testing
  – Linear discriminant
  – Nearest neighbor method

• Estimating classification errors

Classification Objectives

• Class comparison
  – Identify feature sets for predefined classes [Induction]

• Class prediction
  – Develop mathematical function/algorithm that predicts class membership for a novel feature set [Inference]

• Class discovery
  – Identify new classes, sub-classes, or features related to classification objectives [Inference]
Comparison of Typical Gene Pairs in Microarray Data

- Each sample processed by a separate microarray
- Color of dot represents over- or under-expression of an RNA gene transcript in the sample

Class Comparison

- Feature sets for predefined classes
  - Group A samples from tumor tissue
  - Group B samples from normal tissue
    - Genes overexpressed in Group A
    - Genes overexpressed in Group B
Class Prediction

- Algorithm that predicts class membership for a novel feature set
  - Genes of a new sample are analyzed
    - New sample in Group A or Group B?

Class Discovery

- New features revealed in classification
  - New class in universal set?
  - Novel sample type (e.g., antibody) correlates with group?
  - Novel characteristic (e.g., gender, age, or metastasis) correlates with group?
Example for Data Classification

- Data set characterized by two features

Clustering of Data

- What characterizes a cluster?
- How many clusters are there?
Discriminants of Data

Where are the boundaries between sets?

The Data Set Revealed

The discriminant is the Delaware River
Towns and Crossroads of Pennsylvania and New Jersey

Unsupervised Learning

- Learning depends on “closeness” of related features
- Previously unknown correlations or features may be detected
- Meaning of classification occurs after learning via exogenous knowledge
- Same answer given for all questions
Supervised Learning

- Learning depends on prior definition and knowledge of class
- Complex correlation between features is revealed
- Classification is inherent in learning
- Different answers given for different questions

Choosing Features for Classification

- How many?
- How “strong”? 
- Correlation between strong and weak features
- Degree of overlap
- Use of exogenous information for selection
- Statistical significance
- Closeness to boundaries

- To distinguish New Jersey from Pennsylvania, we could consider
  - Longitude
  - Latitude
  - Altitude
  - Temperature
  - Population
  - # of fast-food stores
  - Cultural factors
  - Zip Code
Recall: Membership in a Set

- $A$ = a particular set in $U$
  - defined in a list or rule, or a membership function
- Universal set = all guests at a party
- Particular sets = distinguishing features of guests

Distorted Membership Functions*: Photo

*Photoshop

Ambiguity and uncertainty in data sets to be classified
Distorted Membership Functions*: Map

Characteristics of Classification Features

- **Strong feature**
  - Individual feature provides good classification
  - **Minimal overlap** of feature values in each class
  - Significant difference in **class mean values**
  - Low variance in class

- **Additional features**
  - **Orthogonal feature** (low correlation) adds **new information** to the set
  - **Co-expressed feature** (high correlation) is **redundant**; averaging reduces error
Feature Sets

Best line or curve may classify with significant error

Best plane or surface classifies with equal or less error

Separable Sets

Gene Analysis (2-D)

Bacterial Response to Antibiotics (3-D)
Errors in Classification

• **Expected error in classifier**
  – Minimum possible error with statistically optimal discriminant (e.g., Delaware River) plus
  – Error due to constraint imposed by sub-optimal discriminant (e.g., straight vs. curved line) plus
  – Error due to sampling (i.e., number and distribution of points)

• **Over-/under-fitting**
  – Excessive/inadequate sensitivity to details in training data set
  – Lack of generalization to novel data

• **Validation**
  – Train with less than all available data
  – Reserve some data for evaluation of trained classifier
  – Vary sets used for training and validation
Validation

• Reserve some data for evaluation of trained classifier
• Train with A, test with B
  – A: Training set (or sample)
  – B: Novel set (or sample)
  – Vary sets used for training and validation

• Leave-one-out validation
  – Remove a single sample
  – Train on remaining samples
  – Does the trained classifier identify the single sample?
  – Repeat with all sets, removing all samples, one-by-one

3 x 3 Confusion Matrix

Number of cases predicted to be in each class vs. actual numbers

<table>
<thead>
<tr>
<th>Predicted Class</th>
<th>True Class</th>
<th>Cats</th>
<th>Dogs</th>
<th>Rabbits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cats</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Dogs</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Rabbits</td>
<td>0</td>
<td>1</td>
<td>11</td>
<td></td>
</tr>
</tbody>
</table>

• Interpretation: Actually, there are
  – 8 cats: 5 predicted to be cats, 3 to be dogs, and none to be rabbits
  – 6 dogs: 2 predicted to be cats, 3 to be dogs, and 1 to be rabbit
  – 13 rabbits: None predicted to be cats, 2 to be dogs, and 11 to be rabbits
Classification of Overlapping Sets

- Tumor = Positive
- Normal = Negative

- Altering discriminant changes classification errors

- Classification error can never be zero with simple discriminant

False Classification May Be Inevitable
## Categories of Classification Performance

(2 x 2 Confusion Matrix)

<table>
<thead>
<tr>
<th>Predicted Class</th>
<th>Actual Class</th>
<th>Number in Predicted Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>True Positive (TP)</td>
<td>False Positive (FP)</td>
</tr>
<tr>
<td></td>
<td># Predicted(+) = (# TP + # FP)</td>
<td></td>
</tr>
<tr>
<td>Negative</td>
<td>False Negative (FN)</td>
<td>True Negative (TN)</td>
</tr>
<tr>
<td></td>
<td># Predicted(–) = (# FN + # TN)</td>
<td></td>
</tr>
<tr>
<td>Number in Actual Class</td>
<td># Actual(+) = (# TP + # FN)</td>
<td># Actual(–) = (# FP + # TN)</td>
</tr>
</tbody>
</table>

### Measures of Classification Performance

- **Sensitivity** (% / 100) = \( \frac{\# \text{ True Positive}}{\# \text{ Actual Positive}} \)

- **Specificity** (% / 100) = \( \frac{\# \text{ True Negative}}{\# \text{ Actual Negative}} \)
Measures of Classification Performance

**Accuracy** (%/100) = \( \frac{\# \text{ True Positive} + \# \text{ True Negative}}{\# \text{ Actual Positive} + \# \text{ Actual Negative}} \)

**Positive Predictive Value** (%/100) = \( \frac{\# \text{ True Positive}}{\# \text{ Predicted Positive}} \)

**Negative Predictive Value** (%/100) = \( \frac{\# \text{ True Negative}}{\# \text{ Predicted Negative}} \)

Receiver Operating Characteristic* (ROC) Curve

- Comparison of 3 discriminants
- True positive rate (Sensitivity) vs. False positive rate (1 – Specificity) for a varying parameter (e.g., discriminant location)
- Choose discriminant to maximize the area under the ROC curve

*Devised during WWII to evaluate radar target detection*
Unsupervised Learning

• Recognize patterns
• Group data points that are close to each other
  – Hierarchical trees
  – Two-way clustering
  – $k$-means clustering

Cluster Analysis

<table>
<thead>
<tr>
<th>a)</th>
<th>b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>c)</td>
<td>d)</td>
</tr>
<tr>
<td>e)</td>
<td>f)</td>
</tr>
<tr>
<td>g)</td>
<td>h)</td>
</tr>
<tr>
<td>i)</td>
<td>j)</td>
</tr>
</tbody>
</table>
Pattern Recognition

- **Minimum spanning tree**
  - Find smallest total edge length
  - Eliminate inconsistent edges (e.g., A-B at right)
  - Delete noisy points (e.g., bubble chamber track at right)
  - Recognize and span gaps
  - Delete necks by diameter comparison
  - Group by similar density
- ... plus other methods of digital image analysis (shapes, edges, ...)

Distance Measures Between Data Points

- **Distance between real vectors, \( x_1 \) and \( x_2 \):**
  - Euclidean distance
  - Weighted Euclidean distance
  - Squared Euclidean distance
  - Manhattan distance
  - Chebychev distance
- **“Distance” between different categories, \( x_1 \) and \( x_2 \):**
  - Categorical disagreement distance
Hierarchical Trees
(Dendrograms)

- **Top-down evolution**
  - Begin with 2 best clusters
  - Increase number of clusters
    - k-means clustering
    - Self-organizing map

- **k**-means clustering

- **Self-organizing map**

- **Centroid**: \[ \bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} \]

Hierarchical Trees
(Dendrograms)

- **Bottom-up evolution**
  - Find nearest neighbor to each point in data set
  - Link pairs to closest pairs
    - **Single linkage**: distance between nearest neighbors in clusters
    - **Complete linkage**: distance between farthest neighbors in clusters
    - **Pair-group average/centroid**
Dual Hierarchical Trees

- **Two-way joining**
  - Trees derived from two independent variables
  - Cluster by feature and by sample
  - Cluster by different components of measurement

**Supervised Learning**
Discriminant Analysis

- **Hypothesis test**
  - Are 2 given populations different?
- **Linear discriminant**
  - What is(are) the best line(s)/plane(s)/hyperplane(s) for separating 2 (or $k$) populations?

\[ y = mx + b \]

---

Linear Discriminant

- What is(are) the best line(s)/plane(s)/hyperplane(s) for separating 2 (or $k$) populations?
  - Fisher’s linear discriminant
  - Gradient descent
  - Perceptron
- **Nonseparable sets**
  - Minimum square error

[Click for more information](http://en.wikipedia.org/wiki/Linear_discriminant_analysis)
Discriminant Analysis

- Nearest neighbor method
  - Ignore all points except those closest to evolving discriminant
    - Support vector machine
  - Reshape the space by transformation

Simple Hypothesis Test: $t$ Test

- $t$ compares mean values of two data sets
  - $t$ is reduced by uncertainty in the data sets ($s$)
  - $t$ is increased by number of points in the data sets ($n$)

$$t = \frac{(m_A - m_B)}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}$$

- $m = \text{mean value of data set}$
- $\sigma = \text{standard deviation of data set}$
- $n = \text{number of points in data set}$

- $|t| > 3$, $m_A \neq m_B$ with $\approx 99.7\%$ confidence (error probability $\leq 0.003$ for Gaussian distributions) [$n > 25$]
Analysis of Variance

Variance

\[ \sigma_x^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N - 1} \]

F Statistic

\[ F_{AB} = \frac{\sigma_{x_1}^2}{\sigma_{x_2}^2} = \frac{\sigma_A^2}{\sigma_B^2} \]

- **F test of two populations**
  - Populations are equivalent if
    \[ F_{\text{min}} < F_{AB} < F_{\text{max}} \text{ or } F_{AB} \approx 1 \]
  - Populations are strongly equivalent if
    \[ F_{AB} \approx 1 \text{ and } t_{AB} \approx 0 \]

DNA Microarrays

- Photolithography deposits known 25-mer DNA sequences (oligonucleotides) at known locations (features, or probes) on chip
- 10-20 probes (base pairs) per gene
- Perfect and mismatched features for each gene in separate probes
Microarray Application

cDNA produced from sample RNA, labeled, and hybridized to the array
Array is washed, stained, scanned, and quantified

Example of Gene-by-Gene Tumor/Normal Classification by $t$ Test
(Data from Alon et al, 1999)

- 1,151 genes are over/under-expressed in tumor/normal comparison, $p \leq 0.003$
- Genetically dissimilar samples are apparent

“Cancer-positive gene sets”

“Cancer-negative gene sets”
Different Correlation Matrices from Same Data Set

- Gene correlation \( (D = BB^T) \)

- Sample correlation \( (E = C^T C) \)

Ensemble Mean Values

- Treat each probe set (row) as a redundant, corrupted measurement of the same tumor/normal indicator:

\[
    z_{ij} = k_i y + \varepsilon_{ij}, \quad i = 1, m, \quad j = 1, n
\]

- Compute column averages for each sample sub-group (i.e., sum each column and divide by \( n \)):

\[
    \hat{z}_j = \frac{1}{n} \sum_{i=1}^{n} z_{ij}
\]

- Feature space is reduced from (\# samples x \# genes) to (\# samples)
- Statistics of random variable sums are normal by central limit theorem
Two-Feature Discriminants for Class Prediction and Evaluation
(Alon, Notterman, 1999, data)

• Scatter plot presents average value of up genes vs. average value of down genes for each sample

\[
\hat{z}_{up_j}, \hat{z}_{down_j}
\]

• Classification based on ensemble averages

• Mislabeled samples are identifiable

---

Clustering of Sample Averages for Primary Colon Cancer vs. Normal Mucosa
(PPG data, 2004)

- 144 samples, 3,437 probe sets analyzed
- 47 primary colon cancer
- 22 normal mucosa
- Affymetrix HGU-133A GeneChip
- All transcripts “Present” in all samples

# Probe sets
Up: 1067
Down: 290
Constant: 19
Clustering of Sample Averages for Primary Polyp vs. Normal Mucosa

- 21 primary polyp
- 22 normal mucosa

# Probe sets
Up: 1014
Down: 219
Constant: 40

Clustering of Sample Averages for Primary Polyp vs. Primary Colon Cancer

- 21 primary polyp
- 47 primary colon cancer

# Probe sets
Up: 273
Down: 126
Constant: 172
“Big Data” and Data Mining

Multi-dimensional classification

Ray of hope ...? ... or infinite harm?

http://en.wikipedia.org/wiki/Big_data

Next Time:
Neural Networks