Machine Learning

Lecture # 1
Introduction & Fundamentals
My Introduction*

• Did PhD in 2012 from NUST
  – Area of research: Analysis of Fundus images using Image processing and Machine Learning Techniques

• Current Research Areas:
  – Biomedical Image/Signal Analysis (Retina, Cardiac, Dental, EEG, Breath Sounds etc)
  – Biometrics (Dental, Retina, Dorsal hand veins etc)

• Heading BIOMISA and ETL research lab

*www.biomisa.org/usman
*www.biomisa.org
*www.etechlab.pk
Before we start

How many of you are good in probability and linear algebra?

How many of you are familiar with MATLAB and Pattern Recognition?
What is Machine Learning?

• Machine Learning
  – Study of algorithms that
  – improve their performance
  – at some task
  – with experience

• Optimize a performance criterion using example data or past experience.

• Role of Statistics: Inference from a sample

• Role of Computer science: Efficient algorithms
to
  – Solve the optimization problem
  – Representing and evaluating the model for inference
What is Machine Learning?

- Adapt to / learn from data
  - To optimize a performance function

Can be used to:
- Extract knowledge from data
- Learn tasks that are difficult to formalise
- Create software that improves over time
An Introduction to Machine Learning

- According to Herbert Simon, learning is, “Any change in a System that allows it to perform better the second time on repetition of the same task or on another task drawn from the same population.” [G. F. Luger and W. A. Stubblefield, Artificial Intelligence: Structures and Strategies for Complex Problem Solving, The Benjamin/Cummings Publishing Company, Inc. 1989.]
Why “Learn”? 

• Machine learning is programming computers to optimize a performance criterion using example data or past experience.

• Learning is used when:
  – Human expertise does not exist (navigating on Mars),
  – Humans are unable to explain their expertise (speech recognition)
  – Solution changes in time (routing on a computer network)
  – Solution needs to be adapted to particular cases (user biometrics)
What We Talk About When We Talk About “Learning”

• Learning general models from a data of particular examples

• Example in retail: Customer transactions to consumer behavior:

  People who bought “Da Vinci Code” also bought “The Five People You Meet in Heaven” (www.amazon.com)

• Build a model that is a good and useful approximation to the data.
Machine Learning

Applications:

• Retail: Market basket analysis, Customer relationship management (CRM)
• Finance: Credit scoring, fraud detection
• Manufacturing: Optimization, troubleshooting
• Medicine: Medical diagnosis
• Telecommunications: Quality of service optimization
• Web mining: Search engines
Growth of Machine Learning

• Machine learning is preferred approach to
  – Speech recognition, Natural language processing
  – Computer vision
  – Medical outcomes analysis
  – Robot control
  – Computational biology

• This trend is accelerating
  – Improved machine learning algorithms
  – Improved data capture, networking, faster computers
  – New sensors / IO devices
  – Demand for self-customization to user, environment
Categories

• Association Analysis
• Supervised Learning
  – Classification
  – Regression/Prediction
• Unsupervised Learning
• Reinforcement Learning
Learning Associations

• Basket analysis:

\[ P(Y \mid X) \] probability that somebody who buys \( X \) also buys \( Y \) where \( X \) and \( Y \) are products/services.

Example: \( P(\text{bread} \mid \text{cold drink}) = 0.7 \)

<table>
<thead>
<tr>
<th>( TID )</th>
<th>( Items )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Milk</td>
</tr>
<tr>
<td>2</td>
<td>Bread, Diaper, Cold Drink, Eggs</td>
</tr>
<tr>
<td>3</td>
<td>Milk, Diaper, Cold Drink</td>
</tr>
<tr>
<td>4</td>
<td>Bread, Milk, Diaper, Cold Drink</td>
</tr>
<tr>
<td>5</td>
<td>Bread, Milk, Diaper, Water</td>
</tr>
</tbody>
</table>
Classification

- Example: Credit scoring
- Differentiating between low-risk and high-risk customers from their income and savings

**Discriminant**: IF $\text{income} > \theta_1$ AND $\text{savings} > \theta_2$
THEN low-risk ELSE high-risk
Prediction: Regression

- Example: Price of a used car
- $x$: car attributes
  $y$: price
  
  $y = g \left( x \mid \vartheta \right)$
  
  $g \left( \right)$ model, 
  $\vartheta$ parameters
A pattern is the **opposite of a chaos**, it is an entity that can be given a name.
Recognition

• Identification of a pattern as a member of a category
Classification

Apples

Oranges
Classification

- You had some training example or ‘training data’
- The examples were ‘labeled’
- You used those examples to make the kid ‘learn’ the difference between an apple and an orange
Classification

Given: training images and their categories

What are the categories of these test images?
Pattern Recognition

Given an input pattern, make a decision about the “category” or “class” of the pattern
Unsupervised Learning

• Learning “what normally happens”
• No output
• Clustering: Grouping similar instances
• Other applications: Summarization, Association Analysis

• Example applications
  – Image compression: Color quantization
  – Bioinformatics: Learning motifs
Reinforcement Learning

• Topics:
  – Policies: what actions should an agent take in a particular situation
  – Utility estimation: how good is a state (→ used by policy)

• Credit assignment problem (what was responsible for the outcome)

• Applications:
  – Game playing
  – Robot in a maze
  – Multiple agents, partial observability, ...
Model Choice

– What type of classifier shall we use? How shall we select its parameters? Is there best classifier...?

– How do we train...? How do we adjust the parameters of the model (classifier) we picked so that the model fits the data?
Features

- **Features**: a set of variables believed to carry discriminating and characterizing information about the objects under consideration

- **Feature vector**: A collection of d features, ordered in some meaningful way into a d-dimensional column vector, that represents the signature of the object to be identified.

- **Feature space**: The d-dimensional space in which the feature vectors lie. A d-dimensional vector in a d-dimensional space constitutes a point in that space.
Features

• Feature Choice
  – Good Features
    • Ideally, for a given group of patterns coming from the same class, feature values should all be similar
    • For patterns coming from different classes, the feature values should be different.
  – Bad Features
    • irrelevant, noisy, outlier?
Features

“Good” features

“Bad” features

Linear separability

Non-linear separability

Highly correlated features

Multi-modal
In this course, we will cover supervised classification and clustering
Classification

Learn a method for predicting the instance class from pre-labeled (classified) instances

Many approaches:
Statistics, Decision Trees, Neural Networks, ...

33
Clustering

Find “natural” grouping of instances given un-labeled data
Supervised Learning

We knew the correct answers
Unsupervised Learning

We need to figure out the patterns
Example Applications
Examples: Medicine

Microscopic tissue data - Cancer Detection
Examples: GIS

- Geographic Information Systems
  - Manipulation of Satellite Imagery
  - Terrain Classification, Meteorology
Examples: Industrial Inspection

- Human operators are expensive, slow and unreliable
- Make machines do the job instead
Examples: HCI

- Try to make human computer interfaces more natural
  - Gesture recognition
  - Facial Expression Recognition
  - Lip reading
Examples: Sign Language/Gesture Recognition

British Sign Language Alphabet
Examples: Facial Expression Recognition

- Implicit customer feedback
Examples: Facial Expression Recognition

- Implicit customer feedback
Examples: Biometrics

- Biometrics - Authentication techniques
- Physiological Biometrics
  - Face, IRIS, DNA, Finger Prints
- Behavioral Biometrics
  - Typing Rhythm, Handwriting, Gait
Face Recognition

Training examples of a person

Test images

AT&T Laboratories, Cambridge UK
http://www.uk.research.att.com/facedatabase.html
Examples: Biometrics – Face Recognition
Faces and Digital Cameras

Setting camera focus via face detection

Camera waits for everyone to smile to take a photo [Canon]

Automatic lighting correction based on face detection
Examples: Biometrics – Finger Print Recognition
Examples: Biometrics – Signature Verification
Examples: Robotics
Examples: Robotics

- AIBO
Examples: Optical Character Recognition

- Convert document image into text
Examples: Optical Character Recognition

- Indexing and Retrieval

- Noisy recognition output

Image Source: CEDAR
Examples: Optical Character Recognition

- License Plate Recognition
Examples: Optical Character Recognition

- Automatic Mail Sorting
Example: Hand-written digits

Data representation: Greyscale images
Task: Classification (0,1,2,3.....9)

Problem features:
• Highly variable inputs from same class including some “weird” inputs,
• imperfect human classification,
• high cost associated with errors so “don’t know” may be useful.
A classic example of a task that requires machine learning:
It is very hard to say what makes a 2
<table>
<thead>
<tr>
<th>LeNet 5</th>
<th>RESEARCH</th>
</tr>
</thead>
<tbody>
<tr>
<td>answer: 384</td>
<td></td>
</tr>
<tr>
<td>33 88 44</td>
<td></td>
</tr>
<tr>
<td>33188100</td>
<td></td>
</tr>
<tr>
<td>384</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LeNet 5</th>
<th>RESEARCH</th>
</tr>
</thead>
<tbody>
<tr>
<td>answer: 3</td>
<td></td>
</tr>
<tr>
<td>333</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LeNet 5</th>
<th>RESEARCH</th>
</tr>
</thead>
<tbody>
<tr>
<td>answer: 6</td>
<td></td>
</tr>
<tr>
<td>666</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LeNet 5</th>
<th>RESEARCH</th>
</tr>
</thead>
<tbody>
<tr>
<td>answer: 6</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LeNet 5</th>
<th>RESEARCH</th>
</tr>
</thead>
<tbody>
<tr>
<td>answer: 3</td>
<td></td>
</tr>
<tr>
<td>333</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LeNet 5</th>
<th>RESEARCH</th>
</tr>
</thead>
<tbody>
<tr>
<td>answer: 6</td>
<td></td>
</tr>
<tr>
<td>666</td>
<td></td>
</tr>
</tbody>
</table>
Safety and Security

- Autonomous robots
- Driver assistance
- Monitoring pools (Poseidon)
- Pedestrian detection [MERL, Viola et al.]
- Surveillance
Security and Fraud Detection - Case Study

• Credit Card Fraud Detection
• Detection of Money laundering
  – FAIS (US Treasury)
• Securities Fraud
  – NASDAQ KDD system
• Phone fraud
  – AT&T, Bell Atlantic, British Telecom/MCI
• Bio-terrorism detection at Salt Lake Olympics 2002
# Summary of Applications

<table>
<thead>
<tr>
<th>Problem Domain</th>
<th>Application</th>
<th>Input Pattern</th>
<th>Output Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document Image Analysis</td>
<td>Optical Character Recognition</td>
<td>Document Image</td>
<td>Characters/words</td>
</tr>
<tr>
<td>Document Classification</td>
<td>Internet search</td>
<td>Text Document</td>
<td>Semantic categories</td>
</tr>
<tr>
<td>Document Classification</td>
<td>Junk mail filtering</td>
<td>Email</td>
<td>Junk/Non-Junk</td>
</tr>
<tr>
<td>Multimedia retrieval</td>
<td>Internet search</td>
<td>Video clip</td>
<td>Video genres</td>
</tr>
<tr>
<td>Speech Recognition</td>
<td>Telephone directory assistance</td>
<td>Speech waveform</td>
<td>Spoken words</td>
</tr>
<tr>
<td>Natural Language Processing</td>
<td>Information extraction</td>
<td>Sentence</td>
<td>Parts of Speech</td>
</tr>
<tr>
<td>Biometric Recognition</td>
<td>Personal identification</td>
<td>Face, finger print, Iris</td>
<td>Authorized users for access control</td>
</tr>
<tr>
<td>Medical</td>
<td>Computer aided diagnosis</td>
<td>Microscopic Image</td>
<td>Healthy/cancerous cell</td>
</tr>
<tr>
<td>Military</td>
<td>Automatic target recognition</td>
<td>Infrared image</td>
<td>Target type</td>
</tr>
<tr>
<td>Industrial automation</td>
<td>Fruit sorting</td>
<td>Images taken on conveyor belt</td>
<td>Grade of quality</td>
</tr>
<tr>
<td>Bioinformatics</td>
<td>Sequence analysis</td>
<td>DNA sequence</td>
<td>Known types of genes</td>
</tr>
</tbody>
</table>
Resources: Datasets

• UCI Repository:  

• UCI KDD Archive:  

• Statlib:  [http://lib.stat.cmu.edu/](http://lib.stat.cmu.edu/)

• Delve:  [http://www.cs.utoronto.ca/~delve/](http://www.cs.utoronto.ca/~delve/)
Resources: Journals

- Journal of Machine Learning Research
  www.jmlr.org
- Machine Learning
- IEEE Transactions on Neural Networks
- IEEE Transactions on Pattern Analysis and Machine Intelligence
- Annals of Statistics
- Journal of the American Statistical Association
- ...
Resources: Conferences

- International Conference on Machine Learning (ICML)
- European Conference on Machine Learning (ECML)
- Neural Information Processing Systems (NIPS)
- Computational Learning
- International Joint Conference on Artificial Intelligence (IJCAI)
- ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)
- IEEE Int. Conf. on Data Mining (ICDM)
Course Information
• Course Material
  – Lectures slides, assignments (computer/written), solutions to problems, projects, and announcements will be uploaded on course web page.

http://biomisa.org/usman/machine-learning
Course Information

Books

- Pattern Classification, Duda, Hart & Stork
- Introduction to Machine Learning, Alphaydin

Other reference books are mentioned in course outline on the course web page.
Course Contents

• Introduction to Machine Learning
• K- Nearest Neighbor (KNN)
• Minimum Distance classifier
• Single and Multi layer Neural Network (Perceptron)
• Decision Trees
• Naive Bayes Classifier
• Bayesian decision Theory
• Maximum likelihood Estimation
• Gaussian Mixture Models (GMM) and Expectation maximization
• Hidden Markov Models (HMM)
• Clustering
• Fisher Discriminate Analysis (FDA) and ranksum based feature selection
• Introduction to Fuzzy logic and fuzzy C Means
• Introduction to recommender systems
Grading Policy

Mid Term Exam: 20%
Quizzes (4-6): 10%
Computer and numerical assignments: 15%
Final Project: 10%
Final Exam: 45%
A Classification Problem Example
Given a collection of annotated data. In this case 5 instances of **Katydids** and five of **Grasshoppers**, decide what type of insect the unlabeled example is.

Katydid or Grasshopper?
For any domain of interest, we can measure features

Color \{Green, Brown, Gray, Other\} 

Has Wings?

Abdomen Length

Thorax Length

Antennae Length

Spiracle Diameter

Leg Length

Mandible Size
We can store features in a database.

The classification problem can now be expressed as:

• Given a training database (My_Collection), predict the class label of a previously unseen instance

<table>
<thead>
<tr>
<th>Insect ID</th>
<th>Abdomen Length</th>
<th>Antennae Length</th>
<th>Insect Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.7</td>
<td>5.5</td>
<td>Grasshopper</td>
</tr>
<tr>
<td>2</td>
<td>8.0</td>
<td>9.1</td>
<td>Katydid</td>
</tr>
<tr>
<td>3</td>
<td>0.9</td>
<td>4.7</td>
<td>Grasshopper</td>
</tr>
<tr>
<td>4</td>
<td>1.1</td>
<td>3.1</td>
<td>Grasshopper</td>
</tr>
<tr>
<td>5</td>
<td>5.4</td>
<td>8.5</td>
<td>Katydid</td>
</tr>
<tr>
<td>6</td>
<td>2.9</td>
<td>1.9</td>
<td>Grasshopper</td>
</tr>
<tr>
<td>7</td>
<td>6.1</td>
<td>6.6</td>
<td>Katydid</td>
</tr>
<tr>
<td>8</td>
<td>0.5</td>
<td>1.0</td>
<td>Grasshopper</td>
</tr>
<tr>
<td>9</td>
<td>8.3</td>
<td>6.6</td>
<td>Katydid</td>
</tr>
<tr>
<td>10</td>
<td>8.1</td>
<td>4.7</td>
<td>Katydids</td>
</tr>
</tbody>
</table>

previously unseen instance = 

| 11 | 5.1 | 7.0 | ?????? |
We will return to the previous slide in two minutes. In the meantime, we are going to play a quick game.

I am going to show you some classification problems which were shown to pigeons!

Let us see if you are as smart as a pigeon!
Pigeon Problem 1

Examples of class A

- 3
- 4
- 1.5
- 5
- 6
- 8
- 2.5

Examples of class B

- 5
- 2.5
- 5
- 2
- 8
- 3
- 4.5
- 3
### Pigeon Problem 1

#### Examples of class A

| 3  | 4  |
| 1.5 | 5  |
| 6  | 8  |
| 2.5 | 5  |

#### Examples of class B

| 5  | 2.5 |
| 5  | 2  |
| 8  | 3  |
| 4.5 | 3  |

What class is this object?

What about this one, A or B?

![Bar charts for class A and B examples]

- Class A: 3, 4, 1.5, 6, 2.5, 8, 2.5, 5
- Class B: 5, 2.5, 5, 2, 8, 3, 4.5, 3

![Bar charts for question objects]

- Object 1: 8, 1.5
- Object 2: 4.5, 7
Pigeon Problem 1

Examples of class A

3  4
1.5  5
6  8
2.5  5

Examples of class B

5  2.5
5  2
8  3
4.5  3

This is a B!

Here is the rule. If the left bar is smaller than the right bar, it is an A, otherwise it is a B.
Pigeon Problem 2

Examples of class A

<table>
<thead>
<tr>
<th>4</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Examples of class B

<table>
<thead>
<tr>
<th>5</th>
<th>2.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>2.5</td>
<td>3</td>
</tr>
</tbody>
</table>

Oh! This one's hard!

Even I know this one
Pigeon Problem 2

The rule is as follows, if the two bars are equal sizes, it is an A. Otherwise it is a B.

So this one is an A.
Pigeon Problem 3

Examples of class A

4  4
1  5
6  3

Examples of class B

5  6
7  5
4  8
7  7

This one is really hard! What is this, A or B?
Pigeon Problem 3

Examples of class A

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
</tr>
</tbody>
</table>

Examples of class B

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
</tr>
</tbody>
</table>

It is a B!

The rule is as follows, if the sum of the two bars is less than or equal to 10, it is an A. Otherwise it is a B.
Pigeon Problem 1

Examples of class A

<table>
<thead>
<tr>
<th>Left Bar</th>
<th>Right Bar</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>1.5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>2.5</td>
<td>5</td>
</tr>
</tbody>
</table>

Examples of class B

<table>
<thead>
<tr>
<th>Left Bar</th>
<th>Right Bar</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>2.5</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>4.5</td>
<td>3</td>
</tr>
</tbody>
</table>

Here is the rule again. If the left bar is smaller than the right bar, it is an A, otherwise it is a B.
Let me look it up... here it is. The rule is, if the two bars are equal sizes, it is an **A**. Otherwise it is a **B**.
The rule again: if the square of the sum of the two bars is less than or equal to 100, it is an A. Otherwise it is a B.
Grasshoppers

Katydid Length

Antenna Length

Abdomen Length
We can “project” the previously unseen instance into the same space as the database.

We have now abstracted away the details of our particular problem. It will be much easier to talk about points in space.

Katydidss
Grasshoppers
Simple Linear Classifier

If previously unseen instance above the line then
class is **Katydid**
else
class is **Grasshopper**

- **Katydids**
- **Grasshoppers**

R.A. Fisher
1890-1962
Which of the “Pigeon Problems” can be solved by the Simple Linear Classifier?

1) Perfect  
2) Useless  
3) Pretty Good

Problems that can be solved by a linear classifier are called linearly separable.
MULTI-CLASS CLASSIFICATION
A Famous Problem

R. A. Fisher’s Iris Dataset.

- 3 classes
- 50 of each class

The task is to classify Iris plants into one of 3 varieties using the Petal Length and Petal Width.

Iris Setosa
Iris Versicolor
Iris Virginica
We can generalize the piecewise linear classifier to N classes, by fitting N-1 lines. In this case we first learned the line to (perfectly) discriminate between **Setosa** and **Virginica/Versicolor**, then we learned to approximately discriminate between **Virginica** and **Versicolor**.
Binary classification:

Multi-class classification:
One-vs-all (one-vs-rest):

Class 1: △
Class 2: □
Class 3: ✗
SPLITTING OF TRAINING AND TEST DATA
Dividing Up Data

• We need independent data sets to train, set parameters, and test performance
• Thus we will often divide a data set into three
  – Training set
  – Parameter selection set
  – Test set
• These must be independent
• Data set 2 is not always necessary
## Dataset

<table>
<thead>
<tr>
<th>Inputs</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>95</td>
</tr>
<tr>
<td>33</td>
<td>90</td>
</tr>
<tr>
<td>78</td>
<td>70</td>
</tr>
<tr>
<td>70</td>
<td>45</td>
</tr>
<tr>
<td>80</td>
<td>18</td>
</tr>
<tr>
<td>35</td>
<td>65</td>
</tr>
<tr>
<td>45</td>
<td>70</td>
</tr>
<tr>
<td>31</td>
<td>61</td>
</tr>
<tr>
<td>50</td>
<td>63</td>
</tr>
<tr>
<td>98</td>
<td>80</td>
</tr>
<tr>
<td>73</td>
<td>81</td>
</tr>
<tr>
<td>50</td>
<td>18</td>
</tr>
<tr>
<td>Inputs</td>
<td>Labels</td>
</tr>
<tr>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>15</td>
<td>95</td>
</tr>
<tr>
<td>33</td>
<td>90</td>
</tr>
<tr>
<td>78</td>
<td>70</td>
</tr>
<tr>
<td>70</td>
<td>45</td>
</tr>
<tr>
<td>80</td>
<td>18</td>
</tr>
<tr>
<td>35</td>
<td>65</td>
</tr>
<tr>
<td>45</td>
<td>70</td>
</tr>
<tr>
<td>31</td>
<td>61</td>
</tr>
<tr>
<td>50</td>
<td>63</td>
</tr>
<tr>
<td>98</td>
<td>80</td>
</tr>
<tr>
<td>73</td>
<td>81</td>
</tr>
<tr>
<td>50</td>
<td>18</td>
</tr>
</tbody>
</table>

50:50 split

<table>
<thead>
<tr>
<th>15</th>
<th>95</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>33</td>
<td>90</td>
<td>1</td>
</tr>
<tr>
<td>78</td>
<td>70</td>
<td>0</td>
</tr>
<tr>
<td>70</td>
<td>45</td>
<td>0</td>
</tr>
<tr>
<td>80</td>
<td>18</td>
<td>0</td>
</tr>
<tr>
<td>35</td>
<td>65</td>
<td>1</td>
</tr>
</tbody>
</table>

Can be 70:30 or any other
Estimating the Generalisation Error

• We have a dilemma if we have limited data
  – We want to use as much data as possible for training
  – We need lots of data for estimating the generalisation error

• Obtaining a good estimate of generalisation performance is important for selecting the best parameter values
Cross Validation

• We can solve our dilemma by repeating the training many times on different partitioning
• This is known as K-fold cross validation

\[ D = \{ D_i \}_{i=1}^{P} \quad D_i = (x_i, y_i) \]
Cross Validation

\[ E_g = 5.1 \]
Cross Validation

\[ E_g = 3.7 \]
Cross Validation

$E_g = 4.6$
Cross Validation

$E_g = 4.6$
Cross Validation

\[ E_g = 3.3 \]
Cross Validation

\[
\langle E_g \rangle = \frac{5.1 + 3.7 + 4.6 + 4.6 + 3.3}{5} = 4.3
\]
Cross Validation

Test Set  
Training Set

10-fold cross-validation

\[ E_g = 5.8 \]
Cross Validation

10-fold cross-validation

\[ E_g = 1.8 \]
Cross Validation

$E_g = 4.8$
Cross Validation

\[ E_g = 3.6 \]
Cross Validation

\[ E_g = 7.4 \]
Cross Validation

$$E_g = 0.99$$
Cross Validation

\[ E_g = 4.5 \]
Cross Validation

\[ E_g = 5.4 \]
Cross Validation

$E_g = 6.2$
Cross Validation

$E_g = 2.7$
Cross Validation

\[
\langle E_g \rangle = \frac{5.8 + 1.8 + 4.8 + 3.6 + 7.4 + 0.99 + 4.5 + 5.4 + 6.2 + 2.7}{10} = 4.3
\]
Cross Validation

Test

Leave-one-out cross-validation
Cross Validation

Leave-one-out cross-validation
Cross Validation

Leave-one-out cross-validation
Cross Validation

Leave-one-out cross-validation
Cross Validation

Test

Leave-one-out cross-validation
Cross Validation

$\langle E_g \rangle = 3.9$

- Leave-one-out cross-validation is extreme case
Price of Cross Validation

• Cross-validation is computationally expensive (K-fold cross-validation requires K times as much work)
• There are attempts at estimating generalisation error more cheaply (boot-strapping) methods, but these are not very accurate
• Cross-validation is only necessary when you have little data
• Re-running code is usually cheap compared with writing the code
PERFORMANCE MEASUREMENTS
R.O.C. Analysis

False positives – i.e. falsely predicting an event
False negatives – i.e. missing an incoming event

Similarly, we have “true positives” and “true negatives”
Accuracy Measures

• Accuracy
  \[- = \frac{TP+TN}{P+N}\]

• Sensitivity or true positive rate (TPR)
  \[- = \frac{TP}{TP+FN} = \frac{TP}{P}\]

• Specificity or TNR
  \[- = \frac{TN}{FP+TN} = \frac{TN}{N}\]

• Positive Predictive value (Precision) (PPV)
  \[- = \frac{Tp}{Tp+Fp}\]

• Recall
  \[- = \frac{Tp}{Tp+Fn}\]
Acknowledgements

- Introduction to Machine Learning, Alpaydin
- Statistical Pattern Recognition: A Review – A.K Jain et al., PAMI (22) 2000
- Pattern Recognition and Analysis Course – A.K. Jain, MSU
- *Pattern Classification*” by Duda et al., John Wiley & Sons.
- Some material adapted from Dr Ali Hassan’s slides