Frank Rosenblatt’s Mark I Perceptron at the Cornell Aeronautical Laboratory, Buffalo, New York, circa 1960.
Today’s Schedule

• Course outline and logistics

• An overview of Machine Learning

• Nearest-neighbor classifier
Course outline and logistics
Logistics

• **Instructor:** Asst. Prof. Aykut ERDEM
  aykut@cs.hacettepe.edu.tr
  Office: 111
  Tel: 297 7500, 146

• **Lectures:** Tue 10:00 - 10:45_D8
  Thu 09:00 - 10:45_D10

• **Office hour:** By appointment
• This is a undergraduate-level introductory course in machine learning (ML) which will give a broad overview of many concepts and algorithms in ML.

• The goal is to provide students with a deep understanding of the subject matter and skills to apply these concepts to real world problems.
The course webpage will be updated regularly throughout the semester with lecture notes, programming and reading assignments and important deadlines. http://web.cs.hacettepe.edu.tr/~aykut/classes/spring2015/bbm406/

We will be using Piazza for course related discussions and announcements. Please enroll the class on Piazza by following the link http://piazza.com/class#spring2015/bbm406
Prerequisites

• Basic algorithms, data structures.
• Basic probability and statistics.
• Basic linear algebra and calculus.
• Good programming skills.
Reference Books


• Bayesian Reasoning and Machine Learning, Barber, Cambridge University Press, 2012. (online version available)

• Introduction to Machine Learning (2nd Edition), Alpaydin, MIT Press, 2010

• Pattern Recognition and Machine Learning, Bishop, Springer, 2006

• Machine Learning: A Probabilistic Perspective, Murphy, MIT Press, 2012
Grading Policy

• 30% Problem Sets
• 30% Midterm Exam
• 40% Final Exam
Problem Sets

• There will be 3 problem sets.

• Each one will involve both theoretical and practical exercises.

• Each assignment have to be done individually or in pairs.
KEEP CALM AND DO YOUR PROBLEM SETS
Course Outline

• Week1  Overview of Machine Learning
• Week2  Linear Regression, Least Squares  PS1 out
• Week3  Machine Learning Methodology
• Week4  Statistical Estimation
• Week5  Classification  PS1 due, PS2 out
• Week6  Support Vector Machines
• Week7  Support Vector Machines (cont'd.)
Course Outline (cont’d.)

• Week 8  *Midterm exam*

• Week 9  Decision Tree Learning  *PS2 due, PS3 out*

• Week 10  Ensemble Methods

• Week 11  Instance-based Methods

• Week 12  Neural Networks  *PS3 due*

• Week 13  Unsupervised Learning

• Week 14  Factor Analysis
Machine Learning: An Overview
• “If you were a current computer science student what area would you start studying heavily?”
  – Answer: Machine Learning.
  – “The ultimate is computers that learn”
  – Bill Gates, Reddit AMA

• “Machine learning is the next Internet”
  – Tony Tether, Director, DARPA

• “Machine learning is today’s discontinuity”
  – Jerry Yang, CEO, Yahoo
Learning

What I cannot create, I do not understand.

Know how to solve every problem that has been solved.

Richard Feynman
Two definitions of learning

(1) Learning is the acquisition of knowledge about the world.

Kupfermann (1985)

(2) Learning is an adaptive change in behavior caused by experience.

Shepherd (1988)
Empirical Inference

• Drawing conclusions from empirical data (observations, measurements)

• Example 1: Scientific inference

\[ y = \sum a_i k(x, x_i) + b \]

\[ y = a \cdot x \]

Leibniz, Weyl, Chaitin
Empirical Inference

• Example 2: Perception
Bernhard Schölkopf
Bernhard Schölkopf
Empirical Inference

• Example 2: Perception

"The brain is nothing but a statistical decision organ"

H. Barlow
What is machine learning?
Example: Netflix Challenge

- Goal: Predict how a viewer will rate a movie
- 10% improvement = 1 million dollars
Example: Netflix Challenge

• Goal: Predict how a viewer will rate a movie

• 10% improvement = 1 million dollars

• Essence of Machine Learning:
  • A pattern exists
  • We cannot pin it down mathematically
  • We have data on it
Comparison

• Traditional Programming

  Data → Computer → Output
  Program → Computer

• Machine Learning

  Data → Computer → Program
  Output → Computer
What is Machine Learning?

• [Arthur Samuel, 1959]
  • Field of study that gives computers
  • the ability to learn without being explicitly programmed

• [Kevin Murphy] algorithms that
  • automatically detect patterns in data
  • use the uncovered patterns to predict future data or other outcomes of interest

• [Tom Mitchell] algorithms that
  • improve their performance (P)
  • at some task (T)
  • with experience (E)
What is Machine Learning?

• If you are a Scientist

• If you are an Engineer / Entrepreneur
  • Get lots of data
  • Machine Learning
  • ???
  • Profit!
Why Study Machine Learning?
Engineering Better Computing Systems

• Develop systems
  • too difficult/expensive to construct manually
  • because they require specific detailed skills/knowledge
    • *knowledge engineering bottleneck*

• Develop systems
  • that adapt and customize themselves to individual users.
  • Personalized news or mail filter
  • Personalized tutoring

• Discover new knowledge from large databases
  • Medical text mining (e.g. migraines to calcium channel blockers to magnesium)
    • *data mining*
• Computational studies of learning may help us understand learning in humans
  • and other biological organisms.

• Hebbian neural learning
  • “Neurons that fire together, wire together.”
Why Study Machine Learning?
The Time is Ripe

• Algorithms
  • Many basic effective and efficient algorithms available.

• Data
  • Large amounts of on-line data available.

• Computing
  • Large amounts of computational resources available.
Where does ML fit in?

Psychology/Physiology:
- biology of learning
- inspiring paradigms
- Ex: neural networks

Applied Maths:
- optimization
- linear algebra
- Ex: convex optim

Applications:
- new challenges
- Ex: ad placement

Computer Science:
- algorithm design
- data structure
- complexity analysis
- Ex: kd tree

Statistics:
- estimation techniques
- theoretical framework
- optimality, efficiency
- Ex: learning theory
A Brief History of AI

A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence.

(John McCarthy)
A Proposal for the

DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE

June 17 - Aug. 16

We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.

The following are some aspects of the artificial intelligence problem:

1) **Automatic Computers**

   If a machine can do a job, then an automatic calculator can be programmed to simulate the machine. The speeds and memory capacities of present computers may be insufficient to simulate many of the higher functions of the human brain, but the major obstacle is not lack of machine capacity, but our inability to write programs taking full advantage of what we have.

2) **How Can a Computer be Programmed to Use a Language**

   It may be speculated that a large part of human thought consists of manipulating words according to rules of reasoning
AI Predictions: Experts

Image Credit: http://intelligence.org/files/PredictingAI.pdf
AI Predictions: Non-Experts
AI Predictions: Failed
Why is AI hard?
What humans see
What computers see
"I saw her duck"
“I saw her duck”
“I saw her duck”
We’ve come a long way… IBM Watson

- What is Jeopardy?

- Challenge:
  - [http://youtu.be/_429UlzN1JM](http://youtu.be/_429UlzN1JM)

- Watson Demo:
  - [http://youtu.be/WFR3lOm_xhE?t=22s](http://youtu.be/WFR3lOm_xhE?t=22s)

- Explanation
  - [http://youtu.be/d_yXV22O6n4?t=4s](http://youtu.be/d_yXV22O6n4?t=4s)

- Future: Automated operator, doctor assistant, finance

  - IBM Watson wins on Jeopardy (February 2011)
  - Watson provides cancer treatment options to doctors in seconds (February 2013)
Why are things working today?

- More compute power
- More data
- Better algorithms/models
Machine Learning
(by examples)
Pose Estimation
Collaborative Filtering

Don’t mix preferences on Netflix!

Amazon books
Imitation Learning in Games

Avatar learns from your behavior

Black & White
Lionsgate Studios
Reinforcement Learning

https://www.youtube.com/watch?v=lle RK hsJBJ0
Spam Filtering

ham

spam
Cheque Reading

segment image

Photograph Front of Check
Place the check on a dark background in a well-lit area, the camera steady and align the check’s edges with the camera.

Note: Fidelity cannot act on any written instructions

recognize handwriting
• Raw set of images from several cameras
• Joint layout based on image similarity
Search Ads

why these ads?
Google's self-driving car passes 300,000 miles (Forbes, 8/15/2012)
Given an audio waveform, robustly extract & recognize any spoken words

- Statistical models can be used to
  - Provide greater robustness to noise
  - Adapt to accent of different speakers
  - Learn from training
Natural Language Processing

I need to hide a body
noun, verb, preposition, ...

reservoirs
metal foundries
mines
dumps
swamps
Face Detection

Sudhakar et al., Multi-view Face Detection Using Deep Convolutional Neural Networks, 2015
Topic Models of Text Documents

[Image showing a diagram with keywords and timelines]
Visual Scene Understanding

- skyscraper
- sky
- bell
- dome
- temple
- buildings
- trees
Learning - revisited

prior knowledge → Learning → knowledge

data → Learning
Learning - revisited

Learning

prior knowledge → Learning → knowledge

data → Learning

Crucial open problem: weak intermediate forms of knowledge that support future generalizations.
• Want adaptive robust and fault tolerant systems
• Rule-based implementation is (often)
  - difficult (for the programmer)
  - brittle (can miss many edge-cases)
  - becomes a nightmare to maintain explicitly
  - often doesn’t work too well (e.g. OCR)

• Usually easy to obtain examples of what we want
  IF x THEN DO y
• Collect many pairs \((x_i, y_i)\)
• Estimate function \(f\) such that \(f(x_i) = y_i\) (supervised learning)
• Detect patterns in data (unsupervised learning)
Objectives of Machine Learning

• **Algorithms:** design of efficient, accurate, and general learning algorithms to
  – deal with large-scale problems.
  – make accurate predictions (unseen examples).
  – handle a variety of different learning problems.

• **Theoretical questions:**
  – what can be learned? Under what conditions?
  – what learning guarantees can be given?
  – what is the algorithmic complexity?
• **Example:** an object, instance of the data used.

• **Features:** the set of attributes, often represented as a vector, associated to an example (e.g., height and weight for gender prediction).

• **Labels:** in classification, category associated to an object (e.g., positive or negative in binary classification); in regression real value.

• **Training data:** data used for training learning algorithm (often labeled data).
Definitions and Terminology (cont’d.)

- **Test data**: data used for testing learning algorithm (unlabeled data).

- **Unsupervised learning**: no labeled data.

- **Supervised learning**: uses labeled data.

- **Weakly of semi-supervised learning**: intermediate scenarios.

- **Reinforcement learning**: rewards from sequence of action.
Supervised Learning
Supervised Learning \( y = f(x) \)

- **Binary classification**
  Given \( x \) find \( y \) in \{-1, 1\}

- **Multicategory classification**
  Given \( x \) find \( y \) in \{1, \ldots, k\}

- **Regression**
  Given \( x \) find \( y \) in \( \mathbb{R} \) (or \( \mathbb{R}^d \))

- **Sequence annotation**
  Given sequence \( x_1 \ldots x_l \) find \( y_1 \ldots y_l \)

- **Hierarchical Categorization (Ontology)**
  Given \( x \) find a point in the hierarchy of \( y \) (e.g. a tree)

- **Prediction**
  Given \( x_t \) and \( y_{t-1} \ldots y_1 \) find \( y_t \)

Often with loss \( l(y, f(x)) \)
Binary Classification
Multiclass Classification + Annotation
Regression

Linear:

\[ y = 0.98x - 0.01 \]

\[ r^2 = 0.496 \]

Nonlinear:

<table>
<thead>
<tr>
<th>SPAN</th>
<th>1142</th>
</tr>
</thead>
<tbody>
<tr>
<td>K</td>
<td>0.2153</td>
</tr>
<tr>
<td>PLATEAU</td>
<td>-67.44</td>
</tr>
</tbody>
</table>
Sequence Annotation

given sequence
gene finding
speech recognition
activity segmentation
named entities
Ontology

webpages

Arts
Movies, Television, Music...

Games
Video Games, RPGs, Gambling...

Kids and Teens
Arts, School Time, Teen Life...

Reference
Maps, Education, Libraries...

Shopping
Clothing, Food, Gifts...

World
Català, Dansk, Deutsch, Español, Français, Italiano, 日本語, Nederlands, Polski, Русский, Svenska...

Business
Jobs, Real Estate, Investing...

Health
Fitness, Medicine, Alternative...

Home
Family, Consumers, Cooking...

News
Media, Newspapers, Weather...

Regional
US, Canada, UK, Europe...

Society
People, Religion, Issues...

Sports
Baseball, Soccer, Basketball...

Computers
Internet, Software, Hardware...

Recreation
Travel, Food, Outdoors, Humor...

Science
Biology, Psychology, Physics...

Copyright © 2013 Netscape

5,114,083 sites - 96,877 editors - over 1,014,849 categories
tomorrow’s stock price
Unsupervised Learning

We also used our...
Unsupervised Learning

• Given data $x$, ask a good question ... about $x$ or about model for $x$

• **Clustering**
  Find a set of prototypes representing the data

• **Principal Components**
  Find a subspace representing the data

• **Sequence Analysis**
  Find a latent causal sequence for observations
  • Sequence Segmentation
  • Hidden Markov Model (discrete state)
  • Kalman Filter (continuous state)

• **Hierarchical representations**

• **Independent components / dictionary learning**
  Find (small) set of factors for observation

• **Novelty detection**
  Find the odd one out
Clustering

- Documents
- Users
- Webpages
- Diseases
- Pictures
- Vehicles
  ...

98
Variance component model to account for sample structure in genome-wide association studies, Nature Genetics 2010
Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, Nature 2007
Hierarchical Modeling of Document Topics.

We also used our...
Independent Components

find them automatically
Novelty detection

typical

atypical
Important challenges in ML

• How important is the actual learning algorithm and its tuning

• Simple versus complex algorithm

• Overfitting

• Model Selection

• Regularization
Your 1st Classifier: Nearest Neighbor Classifier
• **Definition:** Acquire an operational definition of a general category of objects given positive and negative training examples.

• Also called binary classification, binary supervised learning
• **Instance Space X:** Set of all possible objects describable by attributes (often called features).

• **Concept c:** Subset of objects from X (c is unknown).

• **Target Function f:** Characteristic function indicating membership in c based on attributes (i.e. label) (f is unknown).

• **Training Data S:** Set of instances labeled with target function.

---

<table>
<thead>
<tr>
<th>correct (complete, partial, guessing)</th>
<th>color (yes, no)</th>
<th>original (yes, no)</th>
<th>presentation (clear, unclear, cryptic)</th>
<th>binder (yes, no)</th>
<th>A+</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  complete</td>
<td>yes</td>
<td>yes</td>
<td>clear</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>2  complete</td>
<td>no</td>
<td>yes</td>
<td>clear</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>3  partial</td>
<td>yes</td>
<td>no</td>
<td>unclear</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>4  complete</td>
<td>yes</td>
<td>yes</td>
<td>clear</td>
<td>yes</td>
<td>yes</td>
</tr>
</tbody>
</table>

---
Concept Learning as Learning a Binary Function

• Task:
  – Learn (to imitate) a function \( f: X \rightarrow \{+1,-1\} \)

• Training Examples:
  – Learning algorithm is given the correct value of the function for particular inputs → training examples
  – An example is a pair \((x, y)\), where \(x\) is the input and \(y=f(x)\) is the output of the target function applied to \(x\).

• Goal:
  – Find a function
    \[ h: X \rightarrow \{+1,-1\} \]
  that approximates
  \[ f: X \rightarrow \{+1,-1\} \]
  as well as possible.
Supervised Learning

• Task:
  – Learn (to imitate) a function $f: X \rightarrow Y$

• Training Examples:
  – Learning algorithm is given the correct value of the function for particular inputs → training examples
  – An example is a pair $(x, f(x))$, where $x$ is the input and $y=f(x)$ is the output of the target function applied to $x$.

• Goal:
  – Find a function $h: X \rightarrow Y$
  that approximates $f: X \rightarrow Y$
  as well as possible.
Supervised / Inductive Learning

• **Given**
  • examples of a function \((x, f(x))\)

• **Predict function** \(f(x)\) **for new examples** \(x\)
  • Discrete \(f(x)\): Classification
  • Continuous \(f(x)\): Regression
  • \(f(x) = \) Probability\((x)\): Probability estimation
Appropriate Applications for Supervised Learning

- **Situations where there is no human expert**
  \[ x: \text{Bond graph for a new molecule.} \]
  \[ f(x): \text{Predicted binding strength to AIDS protease molecule.} \]

- **Situations where humans can perform the task but can't describe how they do it.**
  \[ x: \text{Bitmap picture of hand-written character} \]
  \[ f(x): \text{Ascii code of the character} \]

- **Situations where the desired function is changing frequently**
  \[ x: \text{Description of stock prices and trades for last 10 days.} \]
  \[ f(x): \text{Recommended stock transactions} \]

- **Situations where each user needs a customized function} \]
  \[ x: \text{Incoming email message.} \]
  \[ f(x): \text{Importance score for presenting to user (or deleting without presenting).} \]
Nearest Neighbors

- Table lookup
  For previously seen instance remember label
- Nearest neighbor
  - Pick label of most similar neighbor
  - Slight improvement - use k-nearest neighbors
  - Really useful baseline!
  - Easy to implement for small amounts of data. Why?
K-Nearest Neighbor (kNN)

• Given: Training data \((x_1, y_1), ..., (x_n, y_n)\)
  – Attribute vectors: \(x_i \in X\)
  – Labels: \(y_i \in Y\)

• Parameter:
  – Similarity function: \(K : X \times X \to R\)
  – Number of nearest neighbors to consider: \(k\)

• Prediction rule
  – New example \(x'\)
  – K-nearest neighbors: \(k\) train examples with largest \(K(x_i, x')\)

\[
h(x') = \arg \max_{y \in Y} \left\{ \sum_{i \in \text{knn}(x')} 1[y_i = y] \right\}
\]
K-Nearest Neighbor (kNN)

- How will new examples be classified?
  - Similarity function?
  - Value of $k$?

$$h(x') = \arg \max_{y \in Y} \left\{ \sum_{i \in knn(x')} 1[y_i=y] \right\}$$
4-Nearest Neighbors
4-Nearest Neighbors Sign
If we get more data

- 1 Nearest Neighbor
  - Converges to perfect solution if clear separation
  - Twice the minimal error rate $2p(1-p)$ for noisy problems
- $k$-Nearest Neighbor
  - Converges to perfect solution if clear separation (but needs more data)
  - Converges to minimal error $\min(p, 1-p)$ for noisy problems if $k$ increases
Weighted K-Nearest Neighbor

- Given: Training data \((x_1, y_1), \ldots, (x_n, y_n)\)
  - Attribute vectors: \(x_i \in X\)
  - Target attribute \(y_i \in Y\)

- Parameter:
  - Similarity function: \(K : X \times X \to R\)
  - Number of nearest neighbors to consider: \(k\)

- Prediction rule
  - New example \(x'\)
  - K-nearest neighbors: \(k\) train examples with largest \(K(x_i, x')\)

\[
h(x') = \arg \max_{y \in Y} \left\{ \sum_{i \in \text{knn}(x')} 1_{[y_i = y]} K(x_i, x') \right\}
\]
Nearest Neighbors in Visual Data
Where in the World? [Hays & Efros, CVPR 2008]

A nearest neighbor recognition example
Where in the World? [Hays & Efros, CVPR 2008]
Where in the World? [Hays & Efros, CVPR 2008]
6+ million geotagged photos by 109,788 photographers

Annotated by Flickr users
6+ million geotagged photos by 109,788 photographers

Annotated by Flickr users
Scene Matches

Madrid
England
France
Paris
Croatia
Heidelberg
Macau
Malta
Cairo
Italy
Italy
Italy
Latvia
Europe
Barcelona
Austria
Scene Matches
Scene Matches

Philippines

Houston

Thailand

Houston

Maldives

Philippines

New Zealand

Bermuda

Palau

Mexico 2

Brazil

Mendoza

Brazil

Thailand

Arkansas

Hawaii
The Importance of Data

Percentage of Geolocations within 200km

First Nearest Neighbor Scene Match
Chance – Random Scenes

Database size (thousands of images, log scale)
Scene Completion  [Hays & Efros, SIGGRAPH07]

Original

Input

Scene Matches

Output
... 200 total
Context Matching
Graph cut + Poisson blending

Hays and Efros, SIGGRAPH 2007
Nearest neighbor search in higher dimensions

Linear Search:
  e.g. scanning 4.5M images!

k-D trees:
  axis parallel partitions of the data
  Only effective in low-dimensional data

Large Scale Approximate Indexing
  Locality Sensitive Hashing (LSH), Spill-Tree, NV-Tree
  All above run on a single machine with all data in memory, and scale to big data

Web-scale Approximate Indexing
  Parallel variant of Spill-tree, NV-tree on distributed systems,
  Scale to Billions of images in disks on multiple machines
Overview of Nearest Neighbors

• Very simple method

• Retain all training data
  - Can be slow in testing
  - Finding NN in high dimensions is slow

• Metrics are very important

• Good baseline