



# Introduction to Statistical Machine Learning

GU 4241

Statistical Machine Learning

Xiaofei Shi



# Basic administrative details:

- All up-to-date information is on Courseworks
- Piazza is used for discussion and would be used for questions / comments and for class quizzes. Make sure you are subscribed.
- Lectures are intended to be self-contained. For supplementary readings, with each lecture, we will have pointers to either online reference materials, or chapters from the following books:
  - Pattern Recognition and Machine Learning, Christopher Bishop.
  - Machine Learning: A probabilistic perspective, Kevin Murphy.
  - The Elements of Statistical Learning: Data Mining, Inference and Prediction, Trevor Hastie, Robert Tibshirani, Jerome Friedman.
  - Machine Learning, Tom Mitchell.



# Basic administrative details:

- Instructor: Xiaofei Shi [xs2427@columbia.edu](mailto:xs2427@columbia.edu)  
Office hours: Tue & Thu 12:30 - 1:30 pm
- TA: Dingyi Fang [df2709@columbia.edu](mailto:df2709@columbia.edu)  
Office hours: Mon 7:00 - 9:00 pm



# Tentative Evaluation Plan

Course grade = 40% Homework + 15% Midterm + 15% Final + 25% Project + 5% Participation

- (40%) Homework: 4 homework in total;
- (15%) Midterm: Tentative: Feb 25th, in class;
- (15%) Final: See school schedule;
- (25%) Project: topics TBD; 10% presentation + 15% final report
- ( 5%) Participation: Piazza, recitations, class survey, etc



# What is machine learning:

- The study of computer algorithms that improve automatically through experience



# What is machine learning:

- The study of computer algorithms that improve automatically through experience
- Tom Mitchell:

“A computer program is said to learn from experience  $E$  with respect to some class of tasks  $T$  and performance measure  $P$  if its performance at tasks in  $T$ , as measured by  $P$ , improves with experience  $E$ .”

  - experience  $E$ : training data
  - task  $T$ : improve decision in prediction, classification, clustering etc
  - performance measure  $P$ : loss function



# What is machine learning:

- The study of computer algorithms that improve automatically through experience

- 



# What is machine learning:

- The study of computer algorithms that improve automatically through experience

- 



- What is its relationship with AI, Data Science, Data Mining and Statistics?





# While there is overlap, there are differences

- Statistics: the goal is the understanding of the data at hand
- Artificial Intelligence: the goal is to build an intelligent agent
- Data Mining: the goal is to extract patterns from large-scale data
- Data Science: the science encompassing collection, analysis, and interpretation of data

# Learning tasks:

ACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAACGCTGAGCAATTCGATAGCAATTC  
GATAACGCTGAGCAATTCGGATAACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAACG  
CTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAATTCGGATATCGATAGCAATTCGATAAATC  
GGATAACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAACGCTGAGCAATTCGATAGC  
AATTCGATAACGCTGAGCAATTCGGATATCGATAGCAATTCGATAACGCTGAGCAACGCTGAGCA  
ATTCGATAGCAATTCGATAACGCTGAGCAATTCGGATAACGCTGAGCAATTCGATAGCAATTCGAT  
AACGCTGAGCAACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAATTCGGATAACGCTG  
AGCAATTCGATAGCAATTCGATAACGCTGAGCTGAGCAATTCGATAGCAATTCGATAACGCTGA  
GCAATTCGGATAACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAATTCGATAACGCTGA  
GATAGCAATTCGATAGCAATTCGATAGCAATTCGATAACGCTGAGCAACGCTGAGCAATTCGAT  
AGCAATTCGATAACGCTGAGCAATTCGGATAACGCTGAGCAATTCGATAGCAATTCGATAACGCT  
GAGCAACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAATTCGGATATCGATAGCAATT  
CGATAACGCTGAGCAACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAATTCGGATAAC  
GCTGAGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAATTCGATAACGCTGAGCAATTC  
CTGATAGCAATTCGATAACGCTGAGCAACGCTGAGCAACGCTGAGCAACGCTGAGCAACGCTGAGCA  
AATTCGGATATCGATAGCAATTCGATAACGCTGAGCAACGCTGAGCAACGCTGAGCAACGCTGAGCA  
ACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAATTCGGATAACGCTGAGCAATTCGAT  
AGCATTTCGATAACGCTGAGCAACGCTGAGCAATTCGATAGCAATTCGATCGGATAACGCTGAGCA  
AATTCGATAGCAATTCGATAACGCTGAGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCA  
ATCGGATAACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAATTCGATAGCAATTCGAT  
AGCAATTCGATAACGCTGAGCAATTCGGATATCGATAGCAATTCGATAACGCTGAGCAATTCGAT  
GCAATTCGATAGCAATTCGATAACGCTGAGCAATTCGATAACGCTGAGCAATTCGATAACGCTGAG  
GATAACGCTGAGCAACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAATTCGATAACGCT  
CTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAATTCGATAACGCTGAGCAATTCGATAACGCT  
TGAGCAATTCGGATAACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAACGCTGAGCAAC  
TTCGATAGCAATTCGATAGCAATTCGATAGCAATTCGATAACGCTGAGCAACGCTGAGCAATTC  
GATAGCAATTCGATAACGCTGAGCAATTCGGATAACGCTGAGCAATTCGATAGCAATTCGATAAC  
GCTGAGCAACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAATTCGGATATCGATAGCA  
ATTCGATAACGCTGAGCAACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAATTCGGAT  
AACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCTGAGCAATTCGATAGCAATTCGATA

Which part is the gene?

Supervised and  
unsupervised learning (can  
also use active learning)

# Learning tasks:

Predict stock price

Help making trading decisions!

Market Summary > Nasdaq Composite  
INDEXNASDAQ: .IXIC

12,888.28 +18.28 (0.14%) ↑

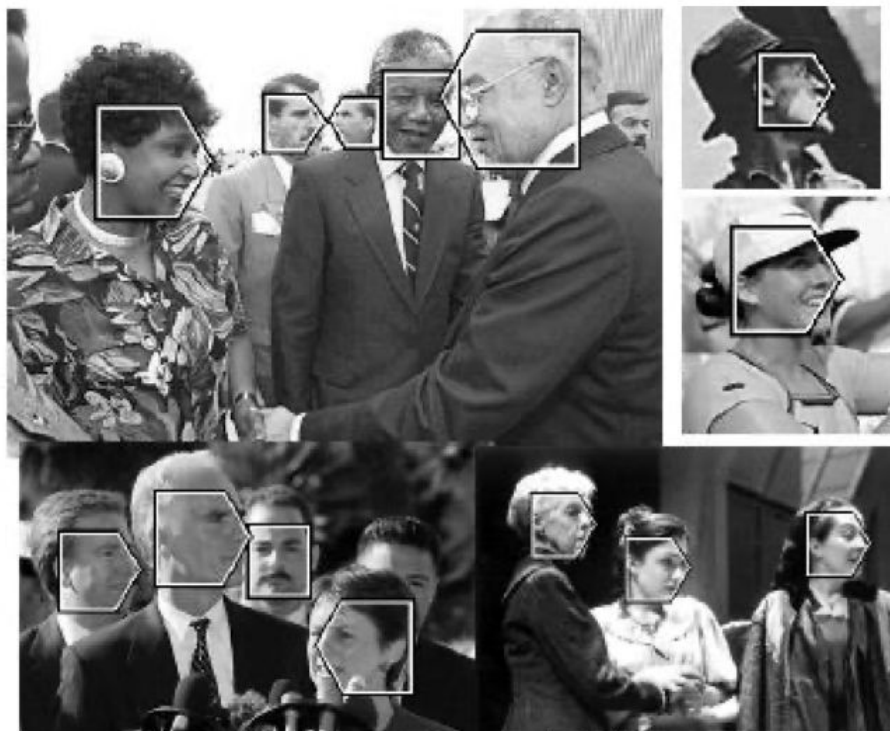
Dec 31, 5:15 PM EST · Disclaimer

1 day 5 days 1 month 6 months YTD 1 year 5 years Max



Supervised  
learning

# Learning tasks:



# Learning tasks:

## NELL: Never-Ending Language Learning

Can computers learn to read? We think so. "Read the Web" is a research project that attempts to create a computer system that learns over time to read the web. Since January 2010, our computer system called NELL (Never-Ending Language Learner) has been running continuously, attempting to perform two tasks each day:

- First, it attempts to "read," or extract facts from text found in hundreds of millions of web pages (e.g., `playsInstrument(George_Harrison, guitar)`).
- Second, it attempts to improve its reading competence, so that tomorrow it can extract more facts from the web, more accurately.





















So far, NELL has accumulated over 50 million candidate beliefs by reading the web, and it is considering these with high confidence. NELL has high confidence in 3,938,530 of these beliefs — these are displayed on this website. It is not perfect, but NELL is learning. You can track NELL's progress below or [@cmunell on Twitter](#), browse and download its [knowledge base](#), read more about our [technical approach](#), or join the [discussion group](#).



semi supervised learning

## Recently-Learned Facts

Refresh

Instance	iteration	date learned	confidence
<a href="#">glass_window_restoration</a> is a <a href="#">household item</a>	1069	03-aug-2017	97.5  
<a href="#">bracelets_curb</a> is a kind of <a href="#">clothing</a>	1069	03-aug-2017	90.9  
<a href="#">hillsborough_lista_d_attesa_crea_un_gruppo_meetup</a> is a <a href="#">visualizable thing</a>	1069	03-aug-2017	99.1  
<a href="#">parison Levitra Viagra Cialis</a> is a <a href="#">drug</a>	1069	03-aug-2017	97.7  
<a href="#">the_democratic_daily</a> is a <a href="#">newspaper</a>	1069	03-aug-2017	100.0  
<a href="#">barcelona_international_airport</a> is an airport <a href="#">in the city barcelona</a>	1073	22-aug-2017	100.0  
<a href="#">john003</a> has brother <a href="#">james</a>	1073	22-aug-2017	100.0  
<a href="#">omaha_world_herald</a> is a newspaper <a href="#">in the city new york</a>	1073	22-aug-2017	93.8  
<a href="#">abc</a> is a company <a href="#">headquartered in the city new york</a>	1073	22-aug-2017	100.0  
<a href="#">arachnids001</a> is an arthropod <a href="#">as well as mites</a> also is	1073	22-aug-2017	93.8  



# Learning tasks:

- Speech recognition, Natural language processing
- Computer vision
- Web forensics
- Medical outcomes analysis Robotics
- Sensor networks
- Social networks
- ...
- Many, many more...



# Data

- Observations:
  - fully observed
  - partially observed: censored data, hidden states, etc...
  - designed experiments
  - actively collected data



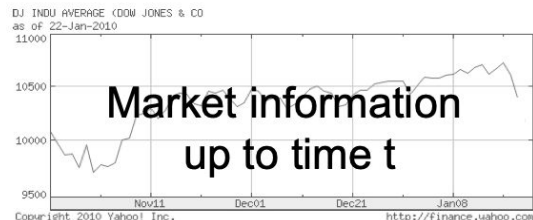
# Task

- Supervised learning, or also called prediction:
  - Regression - given input, estimate output
  - Classification - given input, estimate category
- Unsupervised learning:
  - Data only contains inputs, but no “supervision” in data as to the descriptive outputs
    - density estimation
    - clustering
    - dimensionality reduction



# Supervised learning:

Feature Space  $\mathcal{X}$



Label Space  $\mathcal{Y}$

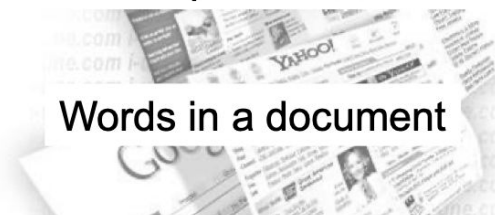
→  
“Sports”  
“News”  
“Science”  
...

→  
Share Price  
“\$ 24.50”

**Task:** Given  $X \in \mathcal{X}$ , predict  $Y \in \mathcal{Y}$ .

# Supervised learning:

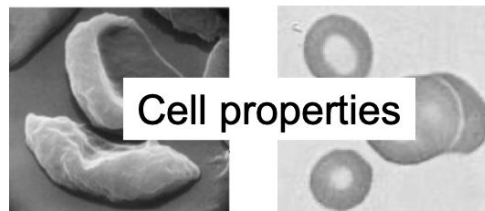
Feature Space  $\mathcal{X}$



Label Space  $\mathcal{Y}$



"Sports"  
"News"  
"Science"  
...



"Anemic cell"  
"Healthy cell"

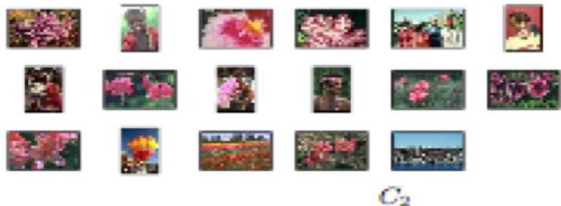
**Task:** Given  $X \in \mathcal{X}$ , predict  $Y \in \mathcal{Y}$ .



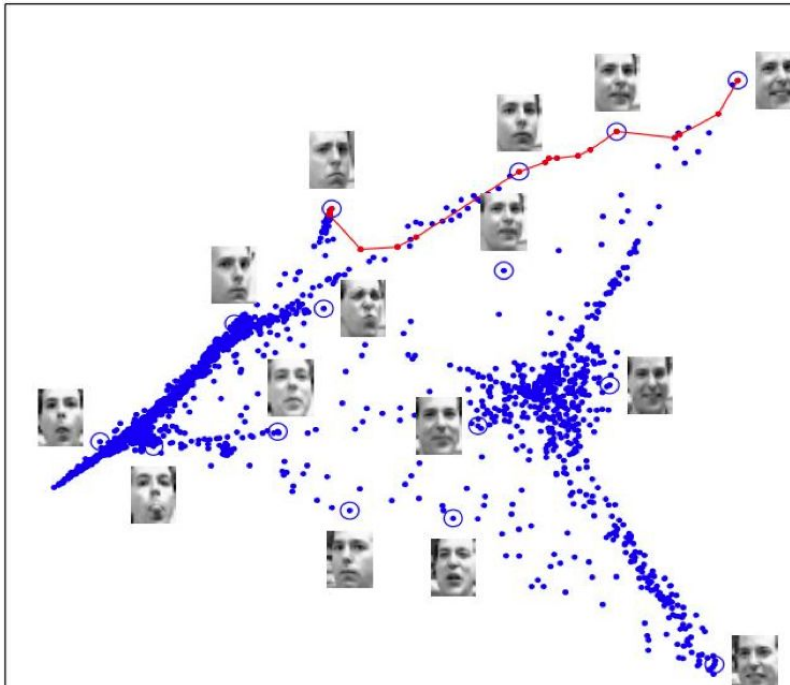
# Unsupervised learning:

Group similar things e.g. images

[Goldberger et al.]



# Unsupervised learning:



Facial recognition  
dimensionality reduction

[Saul & Roweis '03]



# Task

- Supervised learning:  
Given a set of features and labels learn a model that will predict a label to a new feature set
- Unsupervised learning  
Discover patterns in data
- Reasoning under uncertainty  
Determine a model of the world either from samples or as you go along
- Active learning  
Select not only model but also which examples to use



# Algorithm



- Model-based methods
  - probabilistic model of the data
  - parametric models
  - non-parametric models
- Model-free methods

# Model-based algorithm

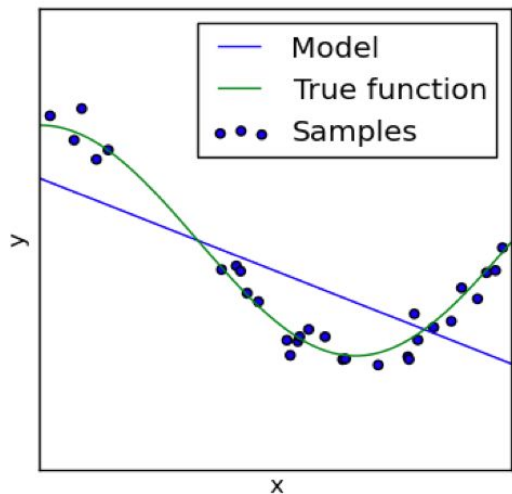


- Learning: from data to model
  - build a model to summarize or to generate the data
  - get estimation on model parameters
  - thus know how to generate future data
- Inference: from model to data
  - Given the model, how can we answer questions relevant to us



# Parametric model

- Fixed size model that the number of parameters does not grow with the data
- More data  $\rightarrow$  better fit of the model

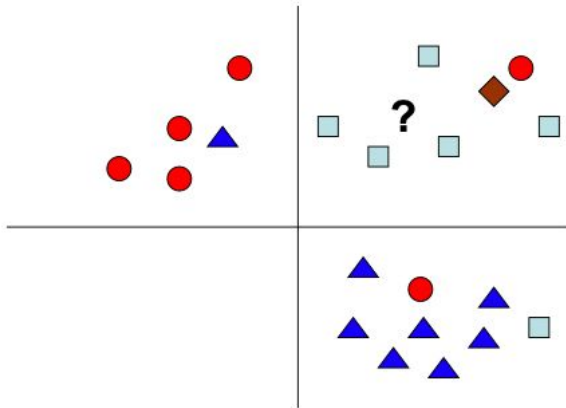


Fitting a simple line (2 params)  
to a bunch of one-dim. samples

Model: data = point on line + noise

# Non-parametric model

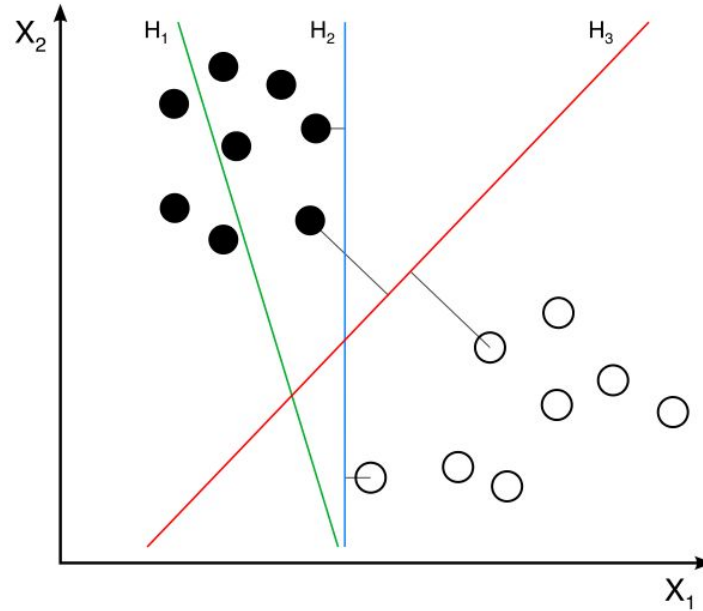
- The number of parameters grows with the data
- More data → a more complex model



- What is the class of the ?  
Input
- Can use the other points (k nearest neighbors) but the number of points to search scales with the input data

# Discriminative model

- Find the best line that separates black points from white points
- No generative assumption





# Common topics

- Mathematical framework:  
Well defined concepts based on explicit assumptions
- Representation:  
How to encode/decode text? Images?
- Model selection  
Which model should we use? How complex should it be?
- Use of prior knowledge  
How do we take our beliefs into consideration? How much can we assume?



# References

- Tom Mitchell: Machine Learning, Chapter 6
- Kevin Murphy: Machine Learning: A probabilistic perspective, Chapter 1
- Ziv Bar-Joseph, Tom Mitchell, Pradeep Ravikumar and Aarti Singh: CMU 10-701