Day -1
Introduction to Machine Learning

Why "Learn"?

- Machine learning is programming computers to optimize a performance criterion using example data or past experience.
- There is no need to "learn" to calculate payroll
- 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
- Data is cheap and abundant (data warehouses, data marts); knowledge is expensive and scarce.
- 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.

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

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
  - Software too complex to write by hand
  - New sensors / IO devices
  - Demand for self-customization to user, environment
  - It turns out to be difficult to extract knowledge from human experts→*failure of expert systems in the 1980's.*

Learning Associations

- Basket analysis:

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

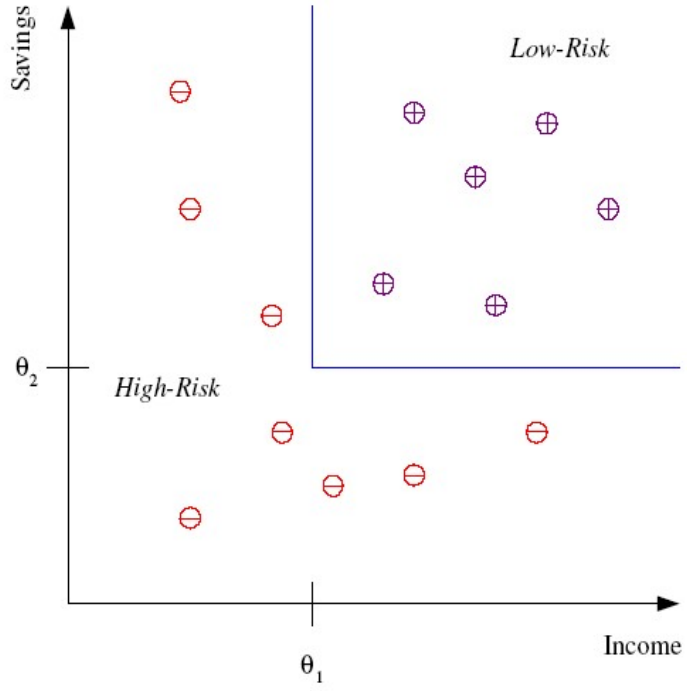
  Example: $P(\text{chips} \mid \text{beer}) = 0.7$

Market-Basket transactions

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Classification

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



Discriminant: IF *income* $> \theta_1$ AND *savings* $> \theta_2$
THEN low-risk ELSE high-risk

Model

Classification: Applications

- Pattern recognition
- Face recognition: Pose, lighting, occlusion (glasses, beard), make-up, hair style
- Character recognition: Different handwriting styles.
- Speech recognition: Temporal dependency.
  - Use of a dictionary or the syntax of the language.
  - Sensor fusion: Combine multiple modalities; eg, visual (lip image) and acoustic for speech
- Medical diagnosis: From symptoms to illnesses
- Web Advertizing: Predict if a user clicks on an ad on the Internet.

Face Recognition

Training examples of a person

Test images

AT&T Laboratories, Cambridge UK
http://www.uk.research.att.com/facedatabase.html

Prediction: Regression

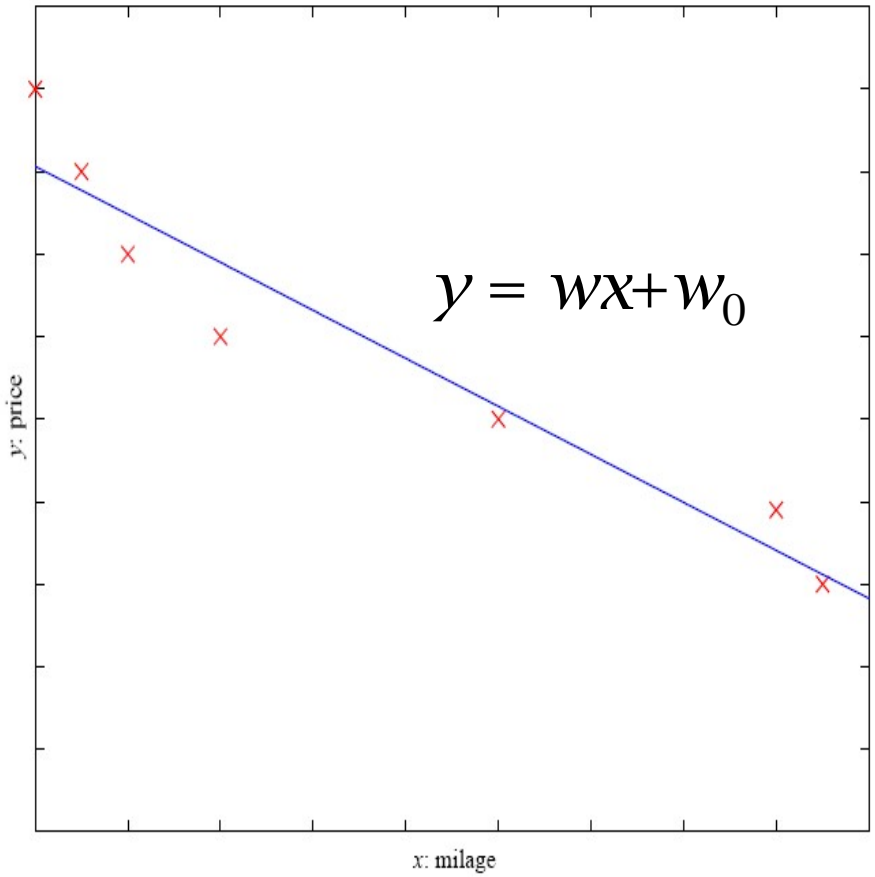
- Example: Price of a used car
- x : car attributes

  y : price

  $$y = g(x \mid \vartheta)$$

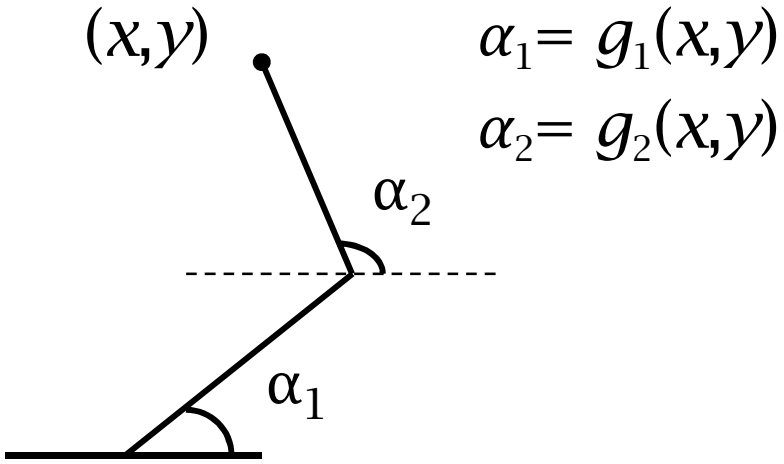
  $g(\ )$ model,

  ϑ parameters



Regression Applications

- Navigating a car: Angle of the steering wheel (CMU NavLab)
- Kinematics of a robot arm



$$\alpha_1 = g_1(x,y)$$
$$\alpha_2 = g_2(x,y)$$

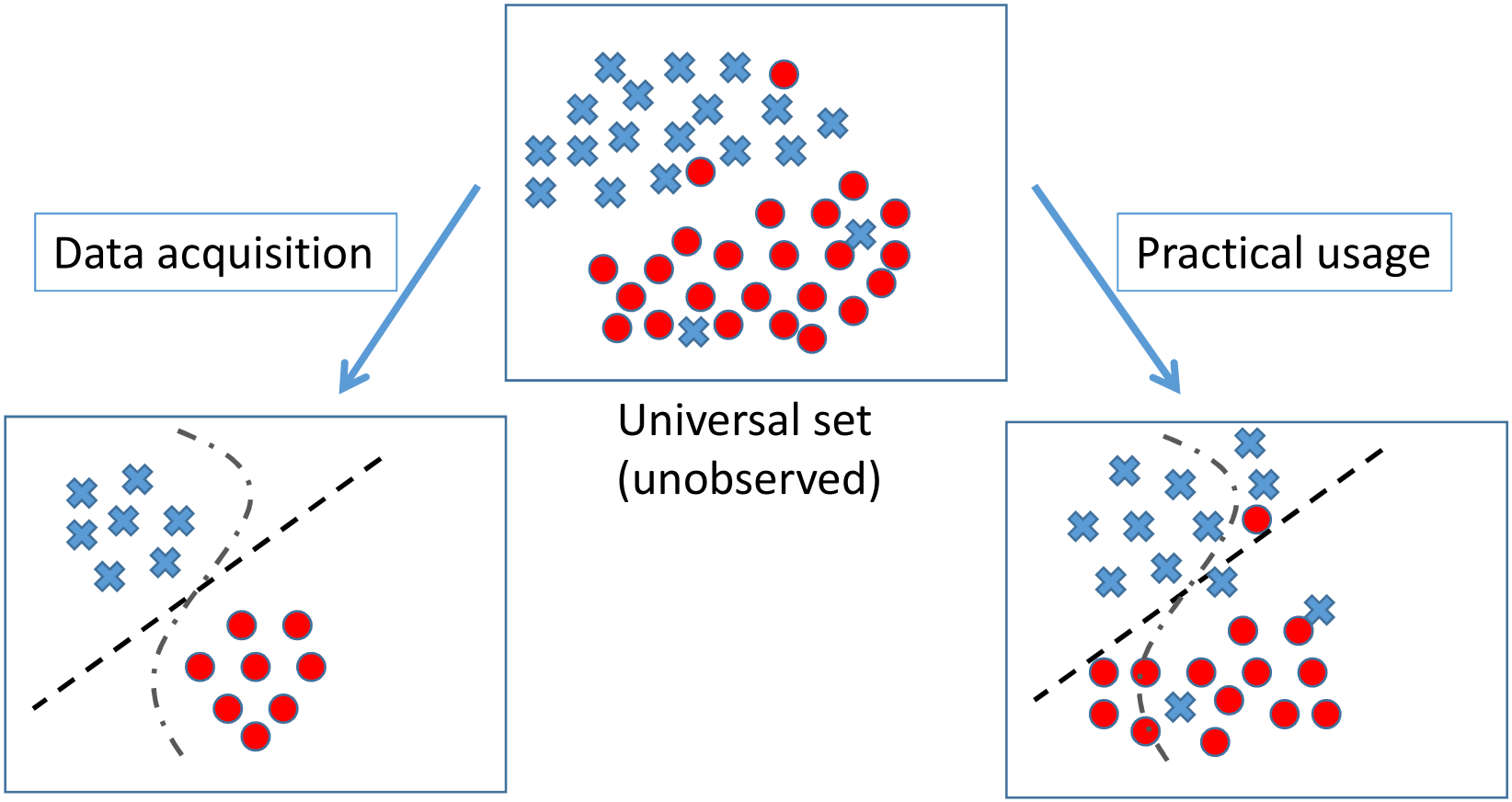
Unsupervised Learning

- Learning “what normally happens”
- No output
- Clustering: Grouping similar instances
- Other applications: Summarization, Association Analysis
- Example applications
  - Customer segmentation in CRM
  - 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)
- No supervised output but delayed reward
- Credit assignment problem (what was responsible for the outcome)
- Applications:
  - Game playing
  - Robot in a maze
  - Multiple agents, partial observability, ...

Training and testing



Why Python

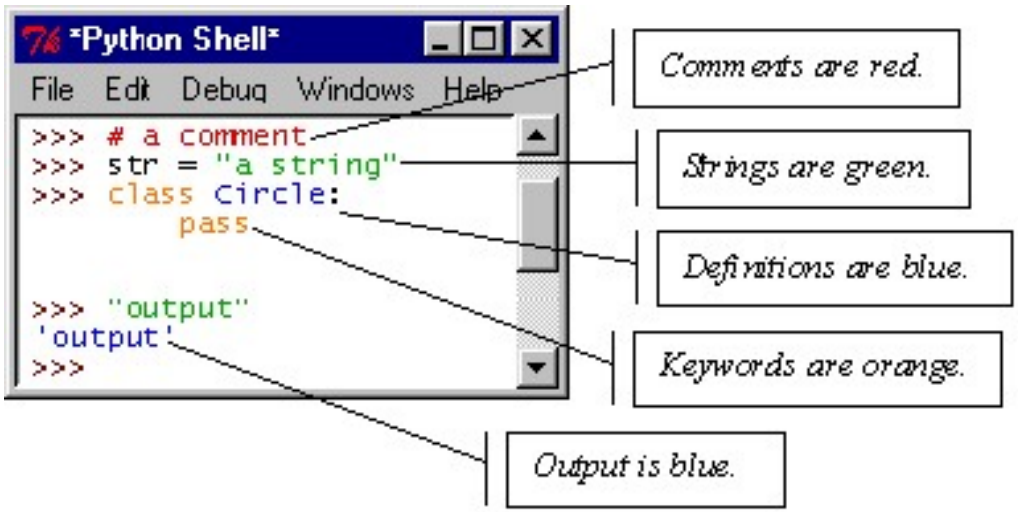
- Python is a high-level programming language
- Open source and community driven
- “Batteries Included”
  - a standard distribution includes many modules
- Dynamic typed
- Source can be compiled or run just-in-time
- Similar to perl, tcl, ruby

Python Interfaces

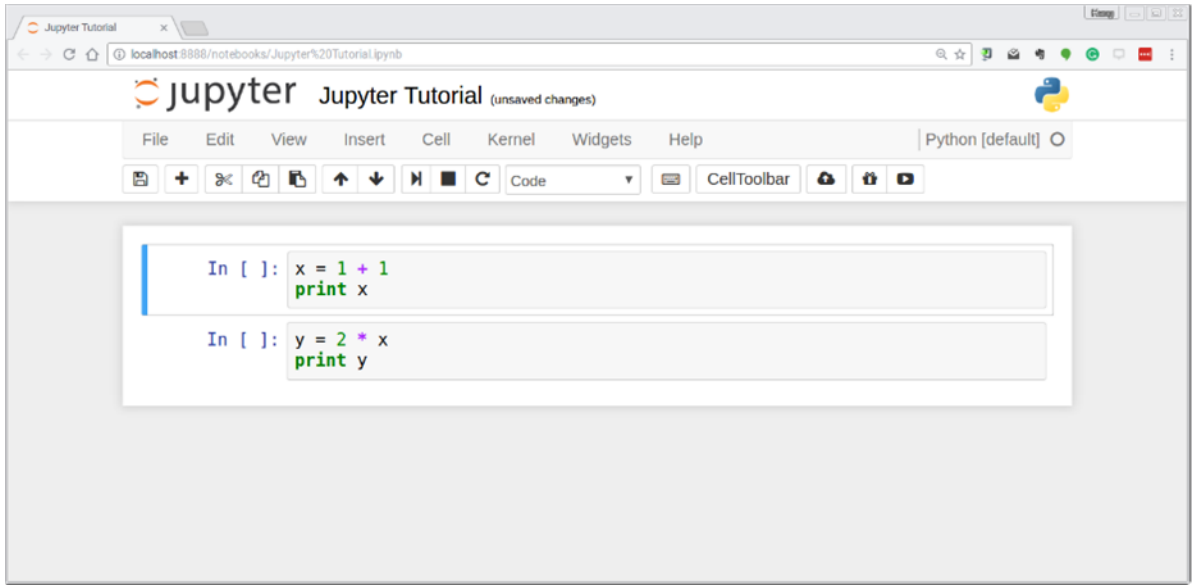
- IDLE – a cross-platform Python development environment
- PythonWin – a Windows only interface to Python
- Python Shell – running 'python' from the Command Line opens this interactive shell
- Jupyter Notebook - The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text.

IDLE – Development Environment

- IDLE helps you program in Python by:
  - color-coding your program code
  - debugging
  - auto-indent
  - interactive shell



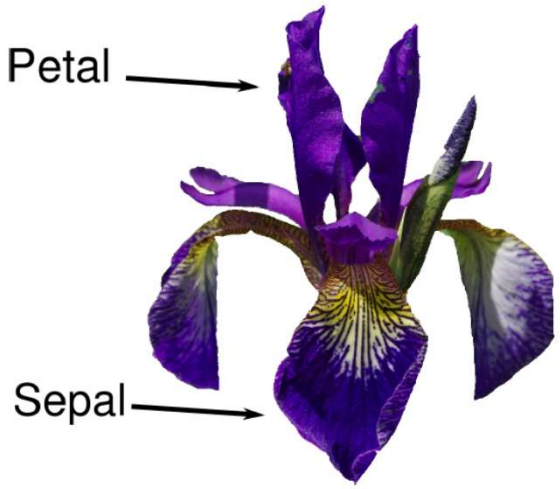
Jupyter Notebook



Python scikit-learn

- Popular machine learning toolkit in Python http://scikit-learn.org/stable/
- Requirements
  - Anaconda
  - Available from https://www.anaconda.com/products/individual
  - Includes numpy, scipy, and scikit-learn (former two are necessary for scikit-learn)
  - In Anaconda prompt "conda install -c anaconda scikit-learn"

A First Application: Classifying Iris Species



- The Iris dataset is a classic dataset for classification, machine learning, and data visualization.

- The dataset contains: 3 classes (different Iris species) with 50 samples each, and then four numeric properties about those classes: Sepal Length, Sepal Width, Petal Length, and Petal Width.
- One species, Iris Setosa, is "linearly separable" from the other two. This means that we can draw a line (or a hyperplane in higher-dimensional spaces) between Iris Setosa samples and samples corresponding to the other two species.
- Predicted Attribute: Different Species of Iris plant.

Open Google Colab

Task -1

https://colab.research.google.com

https://github.com/arunpandianj/Introduction-to-Machine-Learning-with-Python

References

1. Andreas C. Müller and Sarah Guido, Introduction to Machine Learning with Python: A Guide for Data Scientists, O'Reilly, 2016
2. Jake VanderPlas, Python Data Science Handbook: Essential Tools for Working with Data, O'Reilly, 2016
3. Cognitiveclass.ai
4. Deeplearning.ai