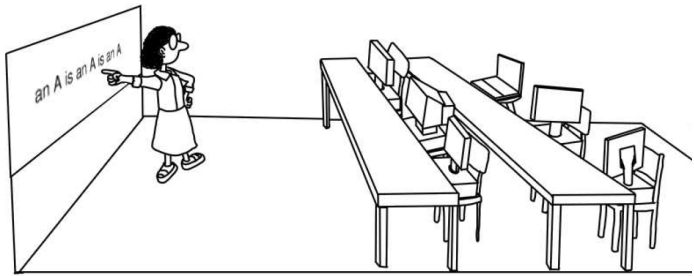
EE3-23: Machine Learning

Krystian Mikolajczyk & Deniz Gunduz

Department of Electrical and Electronic Engineering
Imperial College London



Goal

- To introduce to the fundamental theory and practice of modern machine learning methods
- Part of a course on machine learning and related topics:
  - EE3-10 (Autumn): Maths for Signals and Systems
  - EE3-23 (Autumn): Machine Learning
  - EE3-25 (Spring): Deep Learning
  - EE3-08 (Spring): Advance Signal Processing
  - EE4-68 (Autumn): Pattern Recognition
  - EE4-62 (Spring): Selected Topics in Computer Vision
  - EE4 (Spring): Final Year Project

Course Information

- Objectives
  - become familiar with standard machine learning scenarios and the general methodology to learn from data under various conditions
  - have understanding of basic concepts and ideas, as well as the theory underlying machine learning problems and algorithms
  - learn how to approach real world ML problems, how to formulate them, pre-process data, choose appropriate algorithm and its parameters
  - to employ the ML theory in in arbitrary projects where learning from data is essential

Course Information

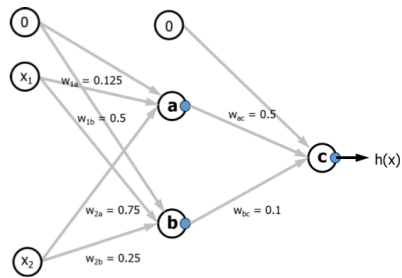
- Dr Krystian Mikolajczyk
  - Room 1015
  - Office hour: Friday 2:00pm-3:00pm
  - Email: k.mikolajczyk@imperial.ac.uk
- Dr Deniz Gunduz
  - Room 1014
  - Office hour:
  - Email: d.gunduz@imperial.ac.uk
- GTAs
  - Dylan Auty
  - Tony Ng
  - Roy Miles
  - Office hour: Friday 2:00pm-3:00pm

Course information

- Lectures (20h): Thursday 10:00am-11:00am (403A) and Friday 3:00-5:00pm (407AB) lectures
- Assessment:
  - 20% coursework assignment returned on week 8 (21 Nov) – a set of exam style questions
  - 80% closed book exam on 12 Dec, 9am

Problem 3 (MLP)**.** A network of three neurons with ReLU activations has been trained with stochastic gradient descent, L2 loss, and learning rate of 0.1. The current weights are given in the figure. Forward propagate training example $(x_1, x_2) = (16, 8)$ with label $y = 5$, calculate outputs of all neurons, then calculate the update of weight w_{bc} using backpropagation. Explain your calculations with appropriate formulas.

[5 points]



Solving past exam questions during lectures

Course information

- Prerequisites
  - mathematical tools: maths, linear algebra, (vector) calculus, probability
- Co-requisite
  - Start learning programming (matlab, python) - many online tutorials
- Reading/watching
  - Lecture slides on Blackboard and videos
  - Learning From Data, Y. S. Abu-Mostafa et al. 2012
- Blackboard FAQ forum

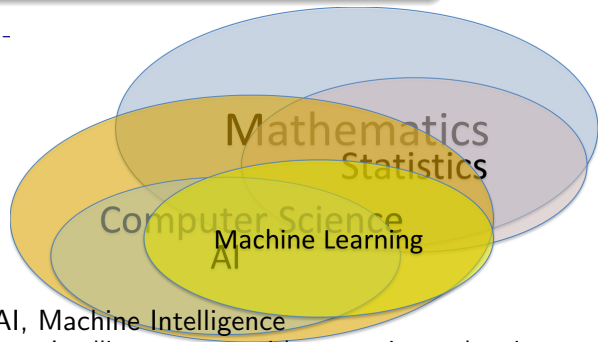
Today's lecture

- Components of learning
- Tasks, types of learning
- Types of data
- Example applications
- ML problem formulation
- A simple hypothesis class

What is Machine Learning?

Extract a description (approximate) of the pattern!

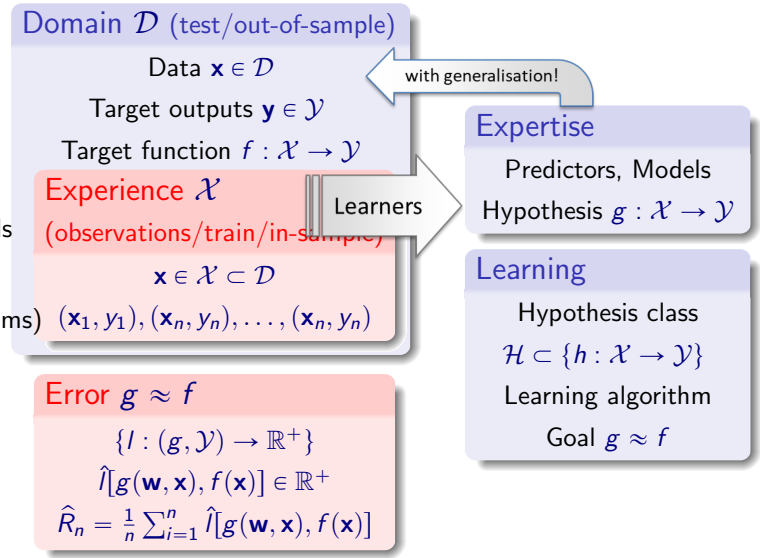
- A pattern exists in a problem
- We cannot solve it analytically $f(\mathbf{x}) \neq a\mathbf{x}$ -
- We have data on it
- Mathematics
  - Study of quantity, structure, space and change
- Statistics
  - Data analysis: from hypothesis to validation by data
- Computer Science
  - Theory, experimentation and engineering for design and use of computers



- AI, Machine Intelligence
  - Intelligent agents with perception and actions to achieve goals
- Machine Learning
  - Ability to learn: from data to hypothesis

Components of learning

- Theory
  - ML problem formulation
  - ML types and tasks
  - Data representations
  - Metrics, error, loss and bounds
  - Empirical Risk Minimization
- Predictors and Learners (algorithms)
  - Linear and non linear
  - SVM
  - Neural Networks
- Learning frameworks
  - Supervised
  - Reinforcement learning

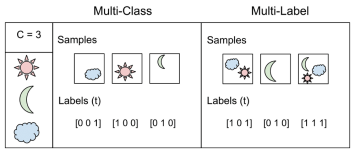


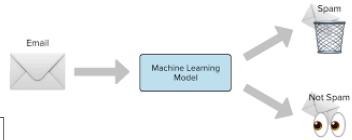
Course Information

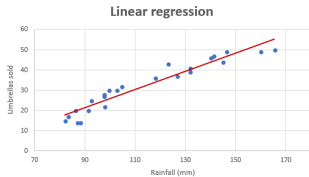
- Content
  - Part 1, KM: Components of learning, tasks, types of learning, ML problem formulation, simple predictors
  - Part 2, KM: Feasibility of learning, error function, Empirical Risk Minimization, generalisation bounds, performance vs complexity, bias/variance trade off, Hoeffding/VC inequalities
  - Part 3, KM: Feature transformations, noisy data, overfitting, regularisation
  - Part 4, DG: Hyperplane, separation with hard margin, soft margin, Support Vector Machines,
  - Part 5, DG: Nearest neighbour classification, introduction to reinforcement learning
  - Part 6, DG: Multi-armed bandits, Markov decision processes (MDPs), Dynamic programming (DP), Value/ policy iteration, RL algorithms
  - Part 7, KM: Logistic regression, gradient descent, Perceptron, Multi Layer Perceptron, Neural Network, backpropagation, revision

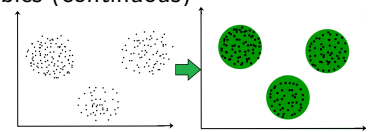
Learning tasks

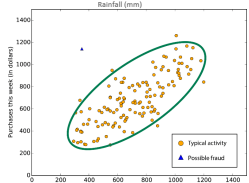
- Classification
  - Binary
  - Multiclass
  - Multilabel
- Regression
  - Univariate (continuous)
  - Multivariate, group of variables (continuous)
- Clustering
  - Hierarchical, Flat
- Anomaly detection
- Dimensionality reduction
- Feature selection

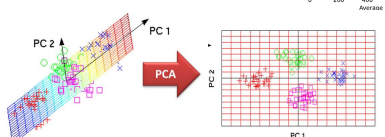




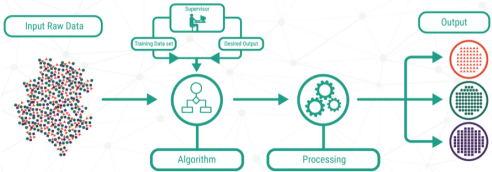


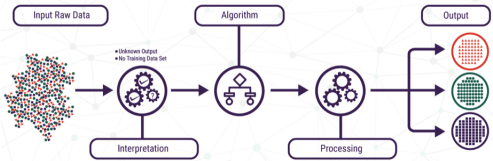






Types of learning





- Supervised learning
  - Learns explicitly, data with clearly defined input–output pairs
  - Predicts outcome, future
  - Direct feedback is given, prediction error
  - Classification, regression, ranking
- Unsupervised learning
  - Attempts to understand the data, find patterns/structures
  - Only input is given, no specific predictions made
  - Evaluation is qualitative, indirect
  - Clustering, association mining, dimensionality reduction
- Reinforcement learning
  - Learns how to act in a given environment to maximize rewards
  - Input, some output, and some valuation is given
  - Decision process, reward or recommendation systems
- Weakly-, semi-supervised learning
  - Output given for part of data, or partially/auxiliary/intermediate outputs are available

Types of learning

Passive vs. active learners

- Data is given vs. learner actively decides which data to use during training

Oblivious vs. adversarial teacher

- Randomly sampled data vs. selected hard samples

Batch vs. online learning

- All training data available vs. training and testing at the same time

Applications

Supervised

- Classification
  - Spam detection
  - Diagnostics
  - Fraud detection
  - Image/audio/text classification
- Regression
  - Risk assessment
  - Score prediction
  - Energy consumption
- Ranking
  - Search engines
  - Information retrieval
  - Machine translation

Unsupervised

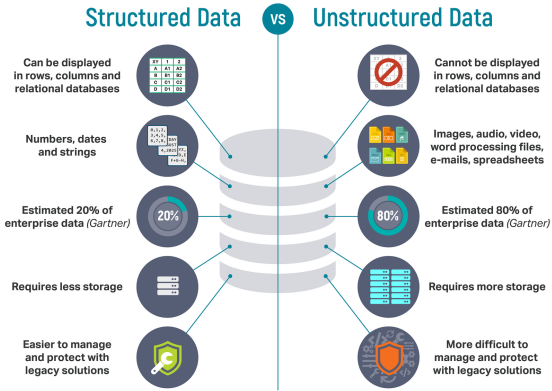
- Dimensionality reduction
  - Big data analytics
- Clustering
  - Biology
  - City planning
  - Marketing

Reinforcement

- Gaming
- Finance
- Manufacturing
- Inventory management
- Robot navigation

Types of data

- Structured
  - Organized into specific types
  - Can be used as direct inputs into ML algorithms
- Unstructured
  - Raw data
  - Diffeternt types and modalities of data are mixed
  - Needs preprocessing for ML algorithms



ML problem formulation

- Input: $\mathbf{x} \in \mathcal{X} \subset \mathcal{D}$
  - text, image, audio clip, sensor output
- Output $\mathbf{y} \in \mathcal{Y}$
  - label, output value
- Target function $f : \mathcal{X} \to \mathcal{Y}$
  - Actual true function - never known in practice
- Data: $(\mathbf{x}_1, y_1), (\mathbf{x}_n, y_n), \ldots, (\mathbf{x}_n, y_n)$
  - available for training
- Error function $\widehat{R}_n = \frac{1}{n} \sum_n \hat{l}(g(\mathbf{w}, \mathbf{x}), y)$
  - Defining $\hat{l}$ is crucial in supervised learning

$\Rightarrow$

Hypothesis: $g : \mathcal{X} \to \mathcal{Y}$.

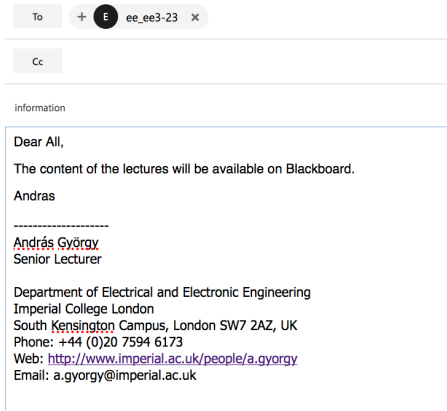
learnt model, one of many possible

Learning

- Hypothesis class: $\mathcal{H} \subset \{h : \mathcal{X} \to \mathcal{Y}\}$
- Learning algorithm
- Goal: $g \approx f$

Applications

- Spam Filtering
  - Goal: Identify spam emails.


To
ee_ee3-23
Cc
information
Dear All,
The content of the lectures will be available on Blackboard.
Andras

András György
Senior Lecturer
Department of Electrical and Electronic Engineering
Imperial College London
South Kensington Campus, London SW7 2AZ, UK
Phone: +44 (0)20 7594 6173
Web: http://www.imperial.ac.uk/people/a.gyorgy
Email: a.gyorgy@imperial.ac.uk

An email is represented by the vector of word counts (bag of words model):

information	dear	all	content	blackboard	Friday	business	proposal	late	husband	...
1	1	1	1	1	0	0	0	0	0	...

Determine if an email is spam based on its word-count vector!

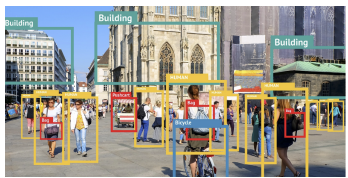
Applications

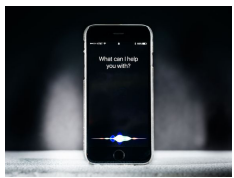
- Character recognition
- Face recognition
- Object recognition

MNIST

Baba Dash (mathworks.com)

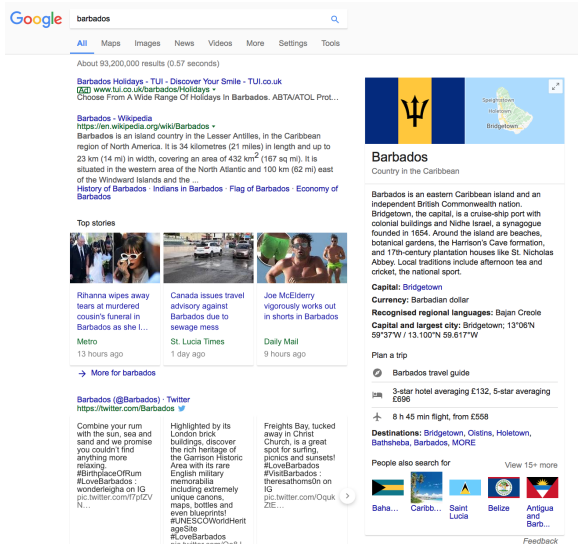
- Intelligent assistants
  - Alexa, Google Now, Siri, Cortana





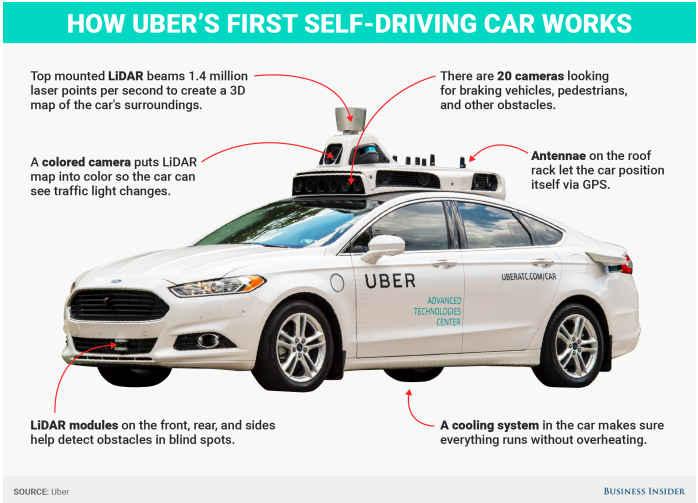
Applications

- Ad placement
  - Google, Bing, Facebook, Yahoo
- Recommendations
  - Amazon, Ebay, Netflix
- Assembly of web pages
  - Yahoo, CNN



Applications

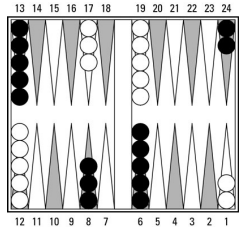
- Self-driving car
  - Google, Uber, several car manufacturers



Applications

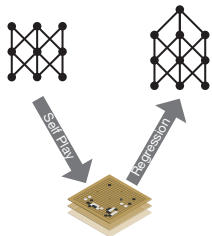
Computer Games:

- Backgammon (1992)



- Poker (Texas hold'em) (2015, 2017)

- Go (2016, 2017)



Silver et al., Nature 2016

A Simple Hypothesis Class – Linear predictor

For input $\mathbf{x} = (x_1, \ldots, x_d)$ (numerical representation of data),

and hypothesis $\mathbf{w} = (w_1, \ldots, w_d)$ (model parameters),

the linear predictor is $h(\mathbf{x}, \mathbf{w}) \to y$, with $\sum_{i=1}^{d} w_i x_i = \mathbf{w}^T \mathbf{x}$

Classification (binary)

Label: $y \in \mathcal{Y} = \{-1, +1\}$

$$h(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T \mathbf{x} < t : \quad h(\mathbf{x}, \mathbf{w}) = -1$$

$$h(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T \mathbf{x} \geqslant t : \quad h(\mathbf{x}, \mathbf{w}) = +1$$

$$\sum_{i=1}^{d} w_i x_i - t \geqslant 0$$

$$\sum_{i=1}^{d} w_i x_i - w_0 x_0 \geqslant 0, \text{ with } x_0 = 1$$

$$\sum_{i=0}^{d} w_i x_i \geqslant 0 \to \text{sign}(\sum_{i=0}^{d} w_i x_i)$$

Predictor: $h(\mathbf{x}, \mathbf{w}) = \text{sign}(\mathbf{w}^T \mathbf{x})$

Regression

Label: $y \in \mathcal{Y} \subset \mathbb{R}$

Predictor: $h(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T \mathbf{x}$

Perceptron Learning Algorithm

Perceptron

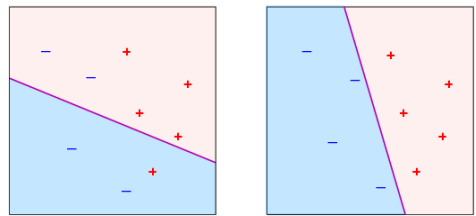
While there exists a missclassified data point with

$$\text{sign}(\mathbf{w}^\top \mathbf{x}_i) \neq y_i \qquad y \in \{-1, +1\}$$

update $\mathbf{w} \leftarrow \mathbf{w} + y_i \mathbf{x}_i$

Intuition: $y_i h(\mathbf{x}_i) > 0$ if $\mathbf{x}_i$ is correctly classified and $y_i h(\mathbf{x}_i) < 0$ if incorrectly.

$$\begin{aligned} y_i \cdot \mathbf{w}'^\top \mathbf{x}_i &= y_i \cdot (\mathbf{w} + y_i \mathbf{x}_i)^\top \mathbf{x}_i \\ &= y_i \cdot \mathbf{w}^\top \mathbf{x}_i + y_i^2 \cdot \mathbf{x}_i^\top \mathbf{x}_i \\ &= y_i \cdot \mathbf{w}^\top \mathbf{x}_i + \|\mathbf{x}_i\|^2 \end{aligned}$$



linearly separable data

Remark: Algorithm stops after a finite number of steps if the data is separable.

Simple Regressor – Closed form solution

Regression error $\widehat{R}_n(h)$

$\widehat{R}_n(h) = \frac{1}{n}\sum_{i=0}^{n}(h(\mathbf{w}, \mathbf{x}_i) - y_i)^2$

$\widehat{R}_n(\mathbf{w}) = \frac{1}{n}\sum_{i=0}^{n}(\mathbf{w}^T\mathbf{x}_i - y_i)^2$

$\widehat{R}_n(\mathbf{w}) = \frac{1}{n}\|X\mathbf{w} - \mathbf{y}\|^2$

with data matrix X and label vector $\mathbf{y}$

$$X = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Solution

$\hat{R}_n(w) = \frac{1}{n}(w^T X^T X w - 2y^T X w + y^T y)$

$\nabla \hat{R}_n(w) = \frac{2}{n}(X^T X w - X^T y) \Rightarrow X^T X w = X^T y \Rightarrow w = (X^T X)^{-1} X^T y$

find $\mathbf{w}$ that minimizes error $\widehat{R}(\mathbf{w})$ i.e. $\dot{\mathbf{w}} = \text{argmin}_{\mathbf{w}}\, \widehat{R}(\mathbf{w})$

$$\nabla \widehat{R}_n(\mathbf{w}) = 0, \qquad \tfrac{2}{n}X^T(X\mathbf{w} - y) = 0$$

$$X^T X\mathbf{w} = X^T\mathbf{y}$$

$$\mathbf{w} = (X^T X)^{-1}X^T\mathbf{y} = \dot{X}\mathbf{y}$$

where $\dot{X}$ is Moore-Penrose pseudoinverse

Classification/regression example

$x_1 = 1, x_2 = 2, x_3 = 7, x_4 = 8$

$y_1 = -1, y_2 = -1, y_3 = +1, y_4 = +1$

$\mathbf{w} = ?$

$\mathbf{x}_1 = (1, 1), \mathbf{x}_2 = (1, 2), \mathbf{x}_3 = (1, 7), \mathbf{x}_4 = (1, 8)$

$\dot{\mathbf{w}} = (w_0, w_1) = ?$

$$\text{find } \text{sign}(\mathbf{w}^\top \mathbf{x}_i) \neq y_i \qquad \text{update } \mathbf{w} \leftarrow \mathbf{w} + y_i \mathbf{x}_i$$

$$\text{initial } \mathbf{w}_{(1)} = (0, 0) \qquad \text{iteration 1: } \mathbf{w}_{(1)}^T \mathbf{x}_1 = 0 \Rightarrow +1 \neq y_1, \Rightarrow \mathbf{w} = y_1 \mathbf{x}_1$$

$$\mathbf{w}_{(2)} = (-1, -1) \qquad \text{iteration 2: } \mathbf{w}_{(2)}^T \mathbf{x}_1 = -2 \Rightarrow -1 = y_1$$

$$\mathbf{w}_{(3)} = (-1, -1) \qquad \mathbf{w}_{(3)}^T \mathbf{x}_2 = -3 \Rightarrow -1 = y_2$$

$$\mathbf{w}_{(4)} = (-1, -1) \qquad \mathbf{w}_{(4)}^T \mathbf{x}_3 = -8 \Rightarrow -1 \neq y_3 \Rightarrow \mathbf{w}_{(5)} = \mathbf{w}_{(4)} + y_3 \mathbf{x}_3$$

$$\mathbf{w}_{(5)} = (0, 6) \qquad \mathbf{w}_{(5)}^T \mathbf{x}_1 = 6 \Rightarrow +1 \neq y_1 \Rightarrow \mathbf{w}_{(6)} = \mathbf{w}_{(5)} + y_1 \mathbf{x}_1$$

$$\mathbf{w}_{(6)} = (-1, 5) \qquad \mathbf{w}_{(6)}^T \mathbf{x}_1 = 4 \Rightarrow +1 \neq y_1 \Rightarrow \mathbf{w}_{(7)} = \mathbf{w}_{(6)} + y_1 \mathbf{x}_1$$

Classification/regression example

$\mathbf{w}_{(7)} = (-2, 4)$ $\quad \mathbf{w}_{(7)}^T \mathbf{x}_1 = 2 \Rightarrow +1 \neq y_1 \Rightarrow \mathbf{w}_{(8)} = \mathbf{w}_{(7)} + y_1 \mathbf{x}_1$

$\mathbf{w}_{(8)} = (-3, 3)$ $\quad \mathbf{w}_{(8)}^T \mathbf{x}_1 = 0 \Rightarrow +1 \neq y_1 \Rightarrow \mathbf{w}_{(9)} = \mathbf{w}_{(8)} + y_1 \mathbf{x}_1$

$\mathbf{w}_{(9)} = (-4, 2)$ $\quad \mathbf{w}_{(9)}^T \mathbf{x}_1 = -2 \Rightarrow -1 = y_1 \Rightarrow \mathbf{w}_{(9)}$

$\mathbf{w}_{(9)} = (-4, 2)$ $\quad \mathbf{w}_{(9)}^T \mathbf{x}_2 = 0 \Rightarrow +1 \neq y_2 \Rightarrow \mathbf{w}_{(10)} = \mathbf{w}_{(9)} + y_2 \mathbf{x}_2$

$\mathbf{w}_{(10)} = (-5, 0)$ $\quad \mathbf{w}_{(10)}^T \mathbf{x}_3 = -5 \Rightarrow -1 \neq y_3 \Rightarrow \mathbf{w}_{(11)} = \mathbf{w}_{(10)} + y_3 \mathbf{x}_3$

$\mathbf{w}_{(11)} = (-4, 7)$ $\quad \mathbf{w}_{(10)}^T \mathbf{x}_1 = 3 \Rightarrow +1 \neq y_1 \Rightarrow \mathbf{w}_{(12)} = \mathbf{w}_{(11)} + y_1 \mathbf{x}_1$

$\mathbf{w}_{(12)} = (-5, 6)$ $\quad \mathbf{w}_{(12)}^T \mathbf{x}_1 = 1 \Rightarrow +1 \neq y_1 \Rightarrow \mathbf{w}_{(13)} = \mathbf{w}_{(12)} + y_1 \mathbf{x}_1$

$\mathbf{w}_{(13)} = (-6, 5)$ $\quad \mathbf{w}_{(13)}^T \mathbf{x}_2 = 4 \Rightarrow +1 \neq y_2 \Rightarrow \mathbf{w}_{(14)} = \mathbf{w}_{(13)} + y_2 \mathbf{x}_2$

$\mathbf{w}_{(14)} = (-7, 3)$ $\quad \mathbf{w}_{(14)}^T \mathbf{x}_4 = 17 \Rightarrow +1 = y_4 \Rightarrow \mathbf{w}_{(14)} = \mathbf{w}_{(14)}$

SOLVED! $\dot{\mathbf{w}} = (-7, 3)$

Classification/regression example

$$\widehat{R}_n(\mathbf{w}) = \tfrac{1}{n}\|X\mathbf{w} - \mathbf{y}\|^2$$

$$\nabla \widehat{R}_n(\mathbf{w}) = 0, \qquad \tfrac{2}{n}X^T(X\mathbf{w} - y) = 0$$

$$X^T X\mathbf{w} = X^T\mathbf{y}$$

$$\mathbf{w} = (X^T X)^{-1}X^T\mathbf{y} = \dot{X}\mathbf{y}$$

$$X = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} = \begin{bmatrix} 1,1 \\ 1,2 \\ 1,7 \\ 1,8 \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \\ +1 \\ +1 \end{bmatrix}$$

Classification/regression example

$\mathbf{w} = (X^T X)^{-1} X^T \mathbf{y} = \dot{X}\mathbf{y}$

$$X^T X = \begin{bmatrix} 1,1,1,1 \\ 1,2,7,8 \end{bmatrix} \begin{bmatrix} 1,1 \\ 1,2 \\ 1,7 \\ 1,8 \end{bmatrix} = \begin{bmatrix} 4,18 \\ 18,118 \end{bmatrix}$$

$$\dot{X} = (X^T X)^{-1} X^T = \begin{bmatrix} 0.7973 - 0.1216 \\ -0.12160.0270 \end{bmatrix} \begin{bmatrix} 1,1,1,1 \\ 1,2,7,8 \end{bmatrix} = \begin{bmatrix} 0.676, 0.554, -0.054, -0.176 \\ -0.095, -0.068, 0.068, 0.095 \end{bmatrix}$$

$$\mathbf{w} = \dot{X}\mathbf{y} = \begin{bmatrix} 0.676, 0.554, -0.054, -0.176 \\ -0.095, -0.068, 0.068, 0.095 \end{bmatrix} \begin{bmatrix} -1, -1, +1, +1 \end{bmatrix}^T = \begin{bmatrix} -1.5 \\ 0.32 \end{bmatrix} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

Classification/regression example

Regression solution used for classification:

$\mathbf{w}_{(1)} = (-1.5, 0.32)$ $\quad \mathbf{w}_{(1)}^T \mathbf{x}_1 = -1.18 \Rightarrow -1 = y_1$

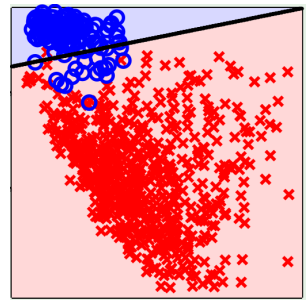
$\mathbf{w}_{(1)} = (-1.5, 0.32)$ $\quad \mathbf{w}_{(1)}^T \mathbf{x}_2 = -0.86 \Rightarrow -1 = y_2$

$\mathbf{w}_{(1)} = (-1.5, 0.32)$ $\quad \mathbf{w}_{(1)}^T \mathbf{x}_3 = 0.74 \Rightarrow +1 = y_3$

$\mathbf{w}_{(1)} = (-1.5, 0.32)$ $\quad \mathbf{w}_{(1)}^T \mathbf{x}_4 = 1.06 \Rightarrow +1 = y_4$

Linear Regression for Classification

- Linear regression learns a real-valued function $y = f(\mathbf{x}) \in \mathbb{R}$.
- Binary valued functions are also real valued: $\pm 1 \in \mathbb{R}$.
- Use linear regression to get $\dot{\mathbf{w}}$ such that $\mathbf{w}^\top \mathbf{x}_i \approx y_i \in \{+1, -1\}$
- Then it is likely that $\text{sign}(\mathbf{w}^\top \mathbf{x}_i) = y_i$.
- Good initial weights for classification
- Error function suboptimal



$$\widehat{R}_{regression}(\mathbf{w}) = \frac{1}{n}\sum_i (\mathbf{w}^T \mathbf{x}_i - y_i)^2 \qquad \widehat{R}_{classification}(\mathbf{w}) = \frac{1}{n}\sum_i \mathbb{I}[\text{sign}(\mathbf{w}^T \mathbf{x}_i) \neq y_i] \qquad (1)$$

ML datasets – example

- UCI Machine Learning Repository
  - https://archive.ics.uci.edu/ml
  - Wine Quality Data Set
  - Human Activity Recognition Using Smartphones

Wine Quality Dataset – example

- Two datasets were created, using red and white wine samples
- The inputs include objective tests (e.g. PH values)
- The output is based on sensory data
  - Median of at least 3 evaluations made by wine experts
  - Each expert graded the wine quality between 0 (very bad) and 10 (very excellent)

Input variables (based on physicochemical tests):
1 - fixed acidity
2 - volatile acidity
3 - citric acid
4 - residual sugar
5 - chlorides
6 - free sulfur dioxide
7 - total sulfur dioxide
8 - density
9 - pH
10 - sulphates
11 - alcohol
Output variable (based on sensory data):
12 - quality (score between 0 and 10)

Wine Quality Dataset – example

Data Set Characteristics:	Multivariate	**Number of Instances:**	4898
Attribute Characteristics:	Real	**Number of Attributes:**	12
Associated Tasks:	Classification, Regression	**Missing Values?**	N/A

- Domain
  - red and white variants of wine
  - $\mathcal{D} \subset \{\mathbf{x} : \mathbf{x} \in \mathbb{R}^{11}\}$
- Data samples
  - physicochemical (inputs) and sensory (the output) variables
  - $\mathcal{X} \subset \mathcal{D}$, $\mathbf{x}_n \in \mathbb{R}^{11}$, $n = 1, \ldots N = 4898$
- Target function
  - Assign quality score, unbalanced
  - Regression
  - $f : \mathcal{X} \to \mathcal{Y}$, $f(\mathbf{x}) = y$, $\mathbf{x} \in \mathcal{X}$, $y \in \mathbb{R}$, $y \in [1, 10]$
- Hypothesis class $\mathcal{H} \subset \{h : \mathcal{X} \to \mathcal{Y}\}$
- Error function $\widehat{R} = \frac{1}{n} \sum_i \ldots$
- Learning algorithm
- Other possible tasks
  - Binary classification (supervised)
  - Multiclass classification (supervised)
  - Feature selection
  - Anomaly detection (unsupervised)
  - Clustering (unsupervised)

Human Activity Recognition Using Smartphones – example

- Recorded from 30 volunteers within an age bracket of 19-48 years wearing a smartphone (Samsung Galaxy S II) on the waist
  - Each person performed six activities (WALKING, WALKING UPSTAIRS, WALKING DOWNSTAIRS, SITTING, STANDING, LAYING) .
  - Accelerometer and gyroscope captured 3-axial linear acceleration and 3-axial angular velocity at a constant rate of 50Hz
  - The experiments have been video-recorded to label the data manually
  - The obtained dataset has been randomly partitioned into two sets, where 70% of the subjects was selected for the training data and 30% for the test data
  - The signals were pre-processed by applying noise filters and then sampled in fixed-width sliding windows of 2.56 sec and 50% overlap (128 readings/window). From each window, a vector of features was obtained by calculating variables from the time and frequency domain
  - The acceleration was separated by a Butterworth low-pass filter into body acceleration and gravity
  - The gravitational force is assumed to have only low frequency components, therefore a filter with 0.3 Hz cutoff frequency was used

Human Activity Recognition Using Smartphones – example

Data Set Characteristics:	Multivariate, Time-Series	Number of Instances:	10299
Attribute Characteristics:	N/A	Number of Attributes:	561
Associated Tasks:	Classification, Clustering	Missing Values?	N/A

- Domain
  - Human activity represented by
  - $\mathcal{D} \subset \{\mathbf{x} : \mathbf{x} \in \mathbb{R}^{561}\}$
- Data samples
  - sensory (the output) variables
  - $\mathcal{X} \subset \mathcal{D}$, $\mathbf{x}_n \in \mathbb{R}^{561}$, $n = 1, \ldots N = 10299$
- Target function
  - Assign activity label
  - Multiclass classification (supervised)
  - $f : \mathcal{X} \to \mathcal{Y}$, $f(\mathbf{x}) = y$, $\mathbf{x} \in \mathcal{X}$, $y \in \mathbb{N}$, $y \in [1, 6]$, $y \in \{1, \ldots, 6\}$
- Hypothesis class $\mathcal{H} \subset \{h : \mathcal{X} \to \mathcal{Y}\}$
- Error function $\widehat{R} = \frac{1}{n} \sum_i \ldots$
- Learning algorithm
- Other possible tasks
  - Binary classification (supervised)
  - Feature selection
  - Clustering (unsupervised)

Some simple terms

- A is impossible: $P(A) = 0$
- A is certain: $P(A) = 1$
- A is almost certain $P(A) \approx 1$
- A is probable $P(A) > 0 \ \wedge P(A) < 1$
- A is correct if error $\varepsilon_A = 0$
- A is approximately correct $\varepsilon_A \approx 0$, very small
- A is probably (how certain?) correct $P(\varepsilon_A = 0) \approx 1$
- A is probably approximately correct P.A.C. $P(\varepsilon_A) > B$ or $P(\varepsilon_A) < B$, with $B \approx 1$

Part 1 summary

- Components of learning
- Tasks, types of learning
- Types of data
- Example applications
- ML problem formulation
- A simple hypothesis class