Advanced Machine Learning

Bogdan Alexe,

bogdan.alexe@fmi.unibuc.ro

University of Bucharest, 2nd semester, 2020-2021

Practical Machine Learning (sem 1) …

Lectures

Lecture 1

- Introduction to Machine Learning
- Basic Concepts
- Learning Paradigms

Lecture 2

- Basic Concepts
- Naive Bayes
- Performance Metrics

Lecture 3

- Nearest Neighbors
- Local Learning
- Curse of Dimensionality

Lecture 4

- Decision Trees
- Random Forests

Lecture 5

- Kernel Methods
- Ridge Regression

Lecture 6

- Support Vector Machines
- Logistic Regression

Lecture 7

- Loss Functions and Optimization
- Gradient Descent
- Code

Lecture 8

- Neural Networks
- Introduction to Deep Learning
- Dropout
- Code

Lecture 9

- Convolutional Neural Networks
- Convolutional Layer
- Pooling Layer

Lecture 10

- Bag-of-Words
- Term Frequency - Inverse Document Frequency
- Bag-of-Visual-Words
- Histogram of Oriented Gradients

Lecture 11

- K-means
- Clustering Goodness
- Soft k-means
- Kernel k-means

Lecture 12

- DBSCAN
- Clustering by unmasking
- Hierarchical Clustering

Practical Machine Learning (sem 1) …

Labs

Installing Anaconda – Windows

Installing Anaconda – Linux

Lab 1

- Introduction to Python
- Introduction to Numpy
- Introduction to Matplotlib
- Solution

Lab 2

- K Nearest Neighbors
- Naive Bayes
- Solution

Lab 3

- Kernel Ridge Regression
- Support Vector Machines
- Solution

Lab 4

- Neural Networks
- Convolutional Neural Networks
- Solution

Lab 5

- K-means
- DBSCAN
- Solution

Lab 6

- PCA
- Solution

Lab 7

- Hierachical Clustering
- Solution

Projects

Supervised Task on Kaggle

- Link to challenge

- Lectures and lab classes oriented towards written ML programs in Python

… vs. Theoretical Machine Learning (sem 2)

- Lectures and seminars oriented towards understanding the theory behind the ML algorithms

Example of a slide from a lecture

Theorem 5.1. (No-Free-Lunch) *Let A be any learning algorithm for the task of binary classification with respect to the 0–1 loss over a domain $\mathcal{X}$. Let m be any number smaller than $|\mathcal{X}|/2$, representing a training set size. Then, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that:*

1. *There exists a function $f : \mathcal{X} \to \{0,1\}$ with $L_{\mathcal{D}}(f) = 0$.*
2. *With probability of at least $1/7$ over the choice of $S \sim \mathcal{D}^m$ we have that $L_{\mathcal{D}}(A(S)) \geq 1/8$.*

… vs. Theoretical Machine Learning (sem 2)

- Lectures and seminars oriented towards understanding the theory behind the ML algorithms

Example of an exercise from seminar

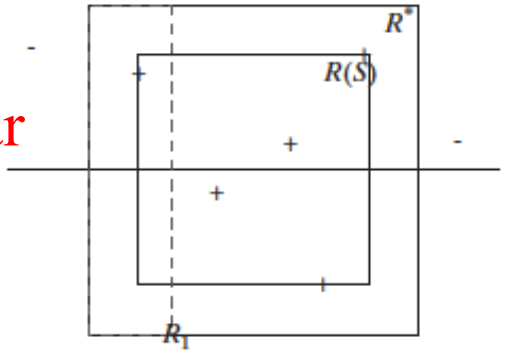


Figure 2.2. Axis aligned rectangles.

Formally, given real numbers $a_1 \leq b_1, a_2 \leq b_2$, define the classifier $h_{(a_1,b_1,a_2,b_2)}$ by

$$h_{(a_1,b_1,a_2,b_2)}(x_1, x_2) = \begin{cases} 1 & \text{if } a_1 \leq x_1 \leq b_1 \text{ and } a_2 \leq x_2 \leq b_2 \\ 0 & \text{otherwise} \end{cases}. \qquad (2.10)$$

The class of all axis aligned rectangles in the plane is defined as

$$\mathcal{H}^2_{\text{rec}} = \{h_{(a_1,b_1,a_2,b_2)} : a_1 \leq b_1, \text{ and } a_2 \leq b_2\}.$$

Note that this is an infinite size hypothesis class. Throughout this exercise we rely on the realizability assumption.

Administrative

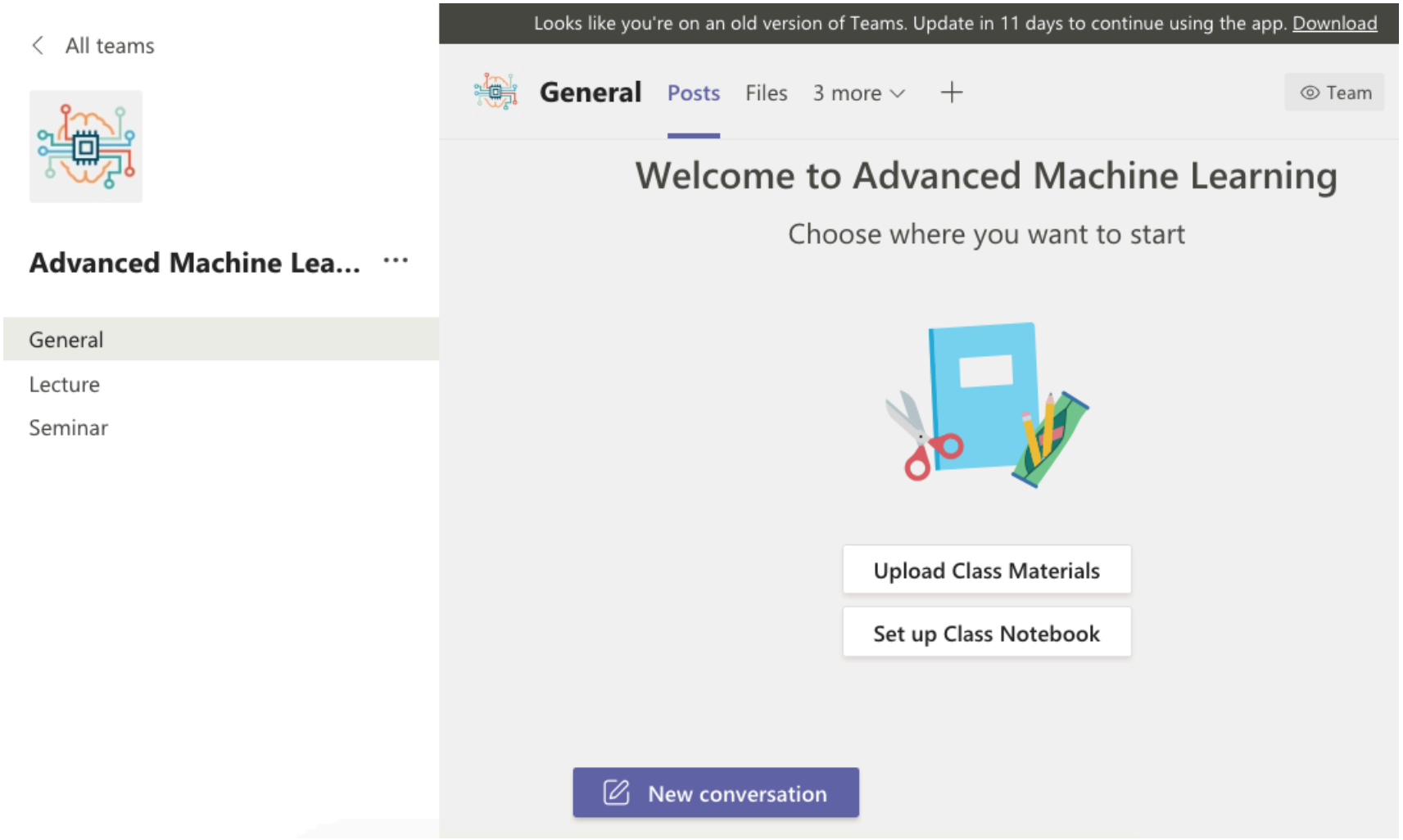
Schedule

- Lecture : Thursday 8 -10 am (weekly)

- Seminar:
  - Thursday 10-12 am (once every two weeks with small groups, today is off)
  - Tuesday 8-10 am (once every two weeks with small groups, next Tuesday, 23th of February, is off)
  - in total there will have 6 seminars (one seminar every two weeks)

Course Materials

- TEAMS



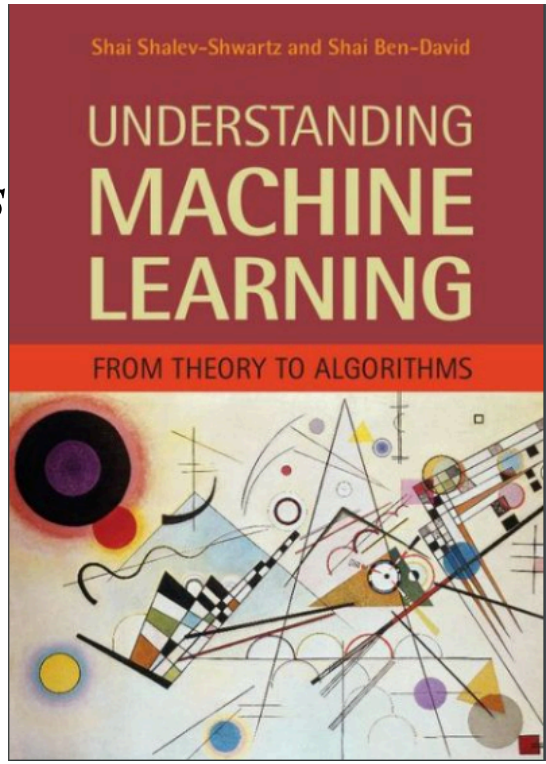
Course Materials

- TEAMS

- Course Book
  - ***Understanding ML from theory to algorithms***

    Shai Shalev-Shwartz, Shai Ben-David,

    Cambridge University Press, 2014
  - **available online**



https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf

Course Materials

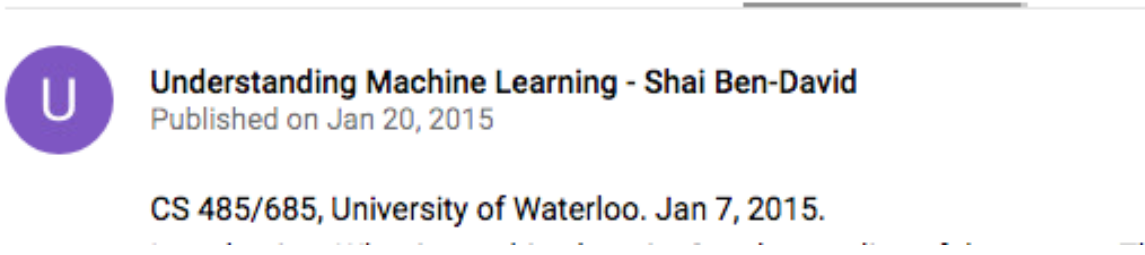
- TEAMS

- Course Book
  - ***Understanding ML from theory to algorithms***

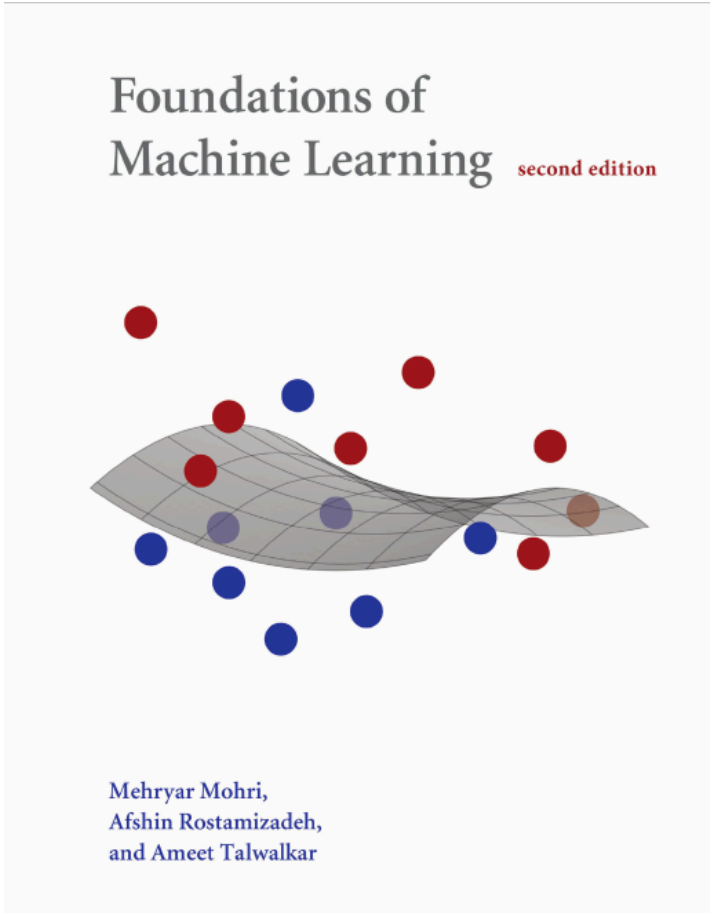
  Shai Shalev-Shwartz, Shai Ben-David,

- Online Lectures
  - Shai Ben-David

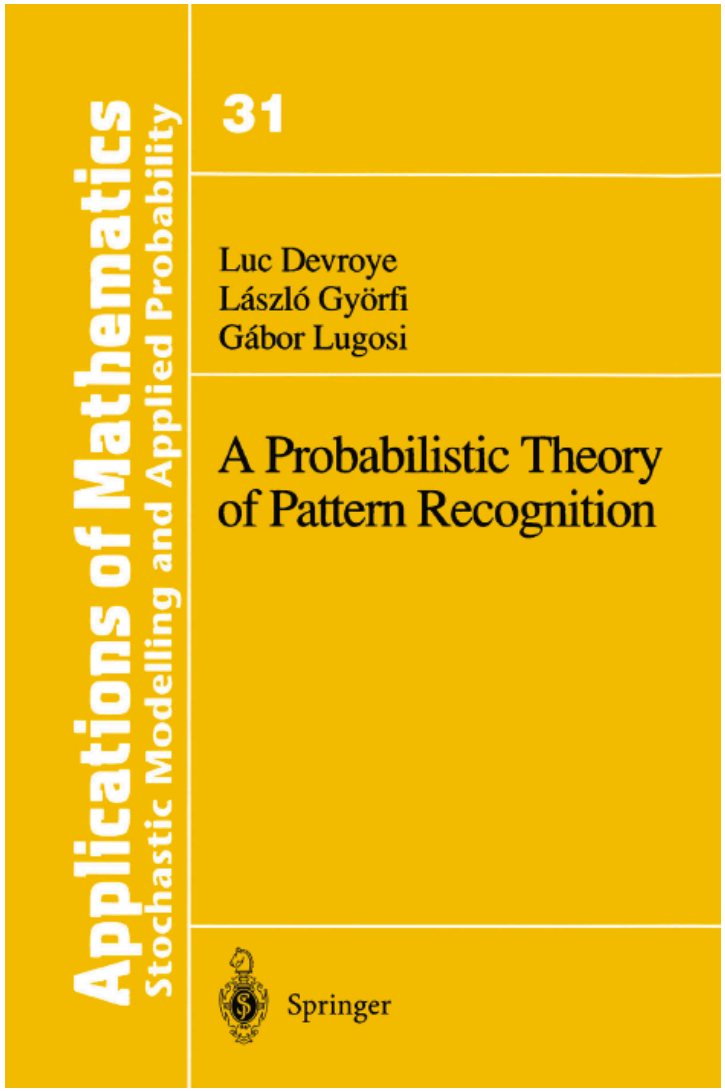
  Youtube videos
  - Shai Shalev-Shwartz: http://www.cs.huji.ac.il/~shais/IML2014.html



Other Books



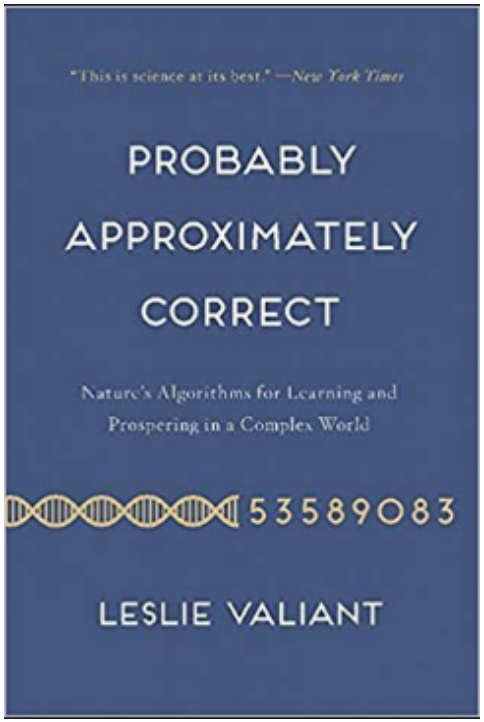
Free online:
https://cs.nyu.edu/~mohri/mlbook/



Other Books

Leslie Valiant, Turing award 2010

For transformative contributions to the theory of computation, including the theory of probably approximately correct (PAC) learning, the complexity of enumeration and of algebraic computation, and the theory of parallel and distributed computing.



Exam – evaluation in June

Grade = min(10, A1 + A2)

- A1 = assignment 1 (written) = 5 points (+ some bonus?)
- A2 = assignment 2 (written) = 5 points (+ some bonus?)
- no constraints, with 4.99 you fail, with 5 you pass

Exam – evaluation in July (restanță)

$$\text{Grade} = \max(2 \times A3, A3 + A1, A3 + A2)$$

- A3 = assignment 3 (written) = 5 points

- no constraints, with 4.99 you fail, with 5 you pass

Exam – evaluation in September (reexaminare)

Grade = max(2 × A4, A4 + A1, A4 + A2, A4 + A3)

- A4 = assignment 4 (written) = 5 points

- no constraints, with 4.99 you fail, with 5 you pass

Exam – evaluation in September (mărire)

Grade = 2 × A4

- A4 = assignment 4 (written) = 5 points

About assignments

- 3-5 exercises, similar with the ones we solve during the seminar class
- handouts in weeks 6/7 and 13/14, you will have about 2/3 weeks to submit your solution
- late submission policy: maximum 3 days allowed, -10% (= 0.5 points) for each day
- submit your solutions as a pdf written with a scientific text software (Word, Latex, LyX) - this is mandatory, your solution will not be accepted if you submit a handwritten solution
- do not share/copy the solution with/from your colleagues: you + your colleague/s will get 0 points

This is acceptable assignment submission

Solution: The choice between the two approaches is highly dependent on the available data. Since the problem addressed by this algorithm is very sensible, any chosen algorithm may target a very high precision for its predictions.

For analyzing the strengths and weaknesses of each approach we refer to the fact that for each hypothesis h_S we can decompose its error into the sum of an *approximation error* and an *estimation error* [Shai & Shai '14],

$$L_{\mathcal{D}}(h_S) = \epsilon_{\text{app}} + \epsilon_{\text{est}}, \text{ where: } \epsilon_{\text{app}} = \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) \text{ and } \epsilon_{\text{est}} = L_{\mathcal{D}}(h_S) - \epsilon_{\text{app}}.$$

The advantage of an algorithm that picks axis aligned rectangle in the two dimensional space spanned by the features BP and BMI is that it has a much smaller search space for the hypothesis that reaches the minimum approximation error, ϵ_{app}. We know that the time needed for computing an ERM hypothesis in the class of axis aligned rectangles is quadratic w.r.t. the dimension of search space, therefore this algorithm would run faster than its proposed counterpart. Nevertheless, because the hypothesis space is smaller there are chances that the computed hypothesis will not be exhaustive enough for the general problem. Thus, the approximation error of this algorithm will be higher. Because the estimation error ϵ_{est} will be smaller, this approach can be harder prone to overfitting.

By comparison an algorithm picking axis aligned rectangle in 5 dimensions is more likely to have a smaller approximation error, since there are much more expressive hypothesis available in this class. Still, this algorithm will run slower then the one considering only 2 features and will be also more exposed to overfitting.

This is not acceptable assignment submission

a) Din punct de vedere al Bias complexity trade-off avem
erroare foarte pentru pro și contra referitor la cele 2 modele. avem
urmărească tipuri de erori.
Eroare de estimare crește odată cu sample size.
Avand mai multe features va duce la scăderea erorii de
aproximare dar va crește eroare de estimare și va duce la
overfitting.
Folosind mai puține features va reduce eroarea de estimare
dar va crește eroare de aproximare ducând la underfitting

b). Un model ce conține mai multe ~~labels~~ ~~training~~ ~~sau~~ features
va avea o scădere a biasului dar este mereu de un set de
date destul de mare.
Chiar dacă dispun de multe date acestea nu pot oferi informații
relevante de unde se folosește feature selection.
Astfel și va reduce overfitting-ul deoarece nu mai există date
redundante.

What is Learning?

What is Learning?

Using Experience to gain Expertise

“Learning” (in nature): using past experience to make future decisions or guide future actions

Applied Animal Behaviour Science
Volume 24, Issue 2, September 1989, Pages 89-99

"Poison-shyness" and "bait-shyness" developed by wild rats (*Rattus rattus* L.). I. Methods for eliminating "shyness" caused by barium carbonate poisoning

Ghazala Naheed, Jamil Ahmad Khan

⊞ Show more

https://doi.org/10.1016/0168-1591(89)90037-3

Abstract

Colonies of wild rats, *Rattus rattus* L., were offered the choice between two baits — cereal grains, flours, mixtures, oily and sweet cereals, and also gram flour. The rats were poisoned in the preferred baits with barium carbonate (100 mg per 10 g food; 200 mg per 10 g food in oily baits) and then presented with the same choice of unpoisoned foods as before.

Poisoning caused a change in the feeding patterns of rats. Foods mixed with barium carbonate were avoided ("poison-shyness"), the same foods then offered without poison were also rejected ("bait-shyness"). Intermittent poisoning also caused aversion to the eating of both poison and bait. Apparently, both the quality and the strength of tastes perceived in the poisonous mixtures influenced the development of "bait-shy" behaviour in the rats.

Bait shyness – Rats Learning to avoid Poisonous Baits

- learning mechanism for rats: they use past experience with some food to acquire expertise in detecting the safety of the food

- a successful learner should be able to progress from individual examples to broader *generalization* . This is also referred to as ***inductive reasoning*** or ***inductive inference***

- rats apply their attitude on new, unseen examples of food of similar smell and taste

Pigeon superstition

https://www.youtube.com/watch?v=TtfQlkGwE2U

"Superstition" in the pigeon.

Database: PsycARTICLES Journal Article

Skinner, B. F.

Citation

Skinner, B. F. (1992). "Superstition" in the pigeon. *Journal of Experimental Psychology: General, 121*(3), 273-274. http://dx.doi.org/10.1037/0096-3445.121.3.273

Abstract

(This reprinted article originally appeared in the *Journal of Experimental Psychology,* 1948, Vol 38, 168–272. The following abstract of the original article appeared in PA, Vol 22:4299.) A pigeon is brought to a stable state of hunger by reducing it to 75% of its weight when well fed. It is put into an experimental cage for a few minutes each day. A food hopper attached to the cage may be swung into place so that the pigeon can eat from it. A solenoid and a timing relay hold the hopper in place for 5 sec at each reinforcement. If a clock is now arranged to present the food hopper at regular intervals with no reference whatsoever to the bird's behavior, operant conditioning usually takes place. The bird tends to learn whatever response it is making when the hopper appears. The response may be extinguished and reconditioned. The experiment might be said to demonstrate a sort of superstition. The bird behaves as if there were a causal relation between its behavior and the presentation of food, although such a relation is lacking. (PsycINFO Database Record (c) 2016 APA, all rights reserved)

Pigeon superstition

Pigeon Superstition: In an experiment performed by the psychologist B. F. Skinner, he placed a bunch of hungry pigeons in a cage. An automatic mechanism had been attached to the cage, delivering food to the pigeons at regular intervals with no reference whatsoever to the birds' behavior. The hungry pigeons went around the cage, and when food was first delivered, it found each pigeon engaged in some activity (pecking, turning the head, etc.). The arrival of food reinforced each bird's specific action, and consequently, each bird tended to spend some more time doing that very same action. That, in turn, increased the chance that the next random food delivery would find each bird engaged in that activity again. What results is a chain of events that reinforces the pigeons' association of the delivery of the food with whatever chance actions they had been performing when it was first delivered. They subsequently continue to perform these same actions diligently.[1]

Bait shyness revisited

Relation of cue to consequence in avoidance learning.

EXPORT ★ Add To My List

Database: PsycINFO Journal Article

Garcia, John Koelling, Robert A.

Citation

Garcia, J., & Koelling, R. A. (1966). Relation of cue to consequence in avoidance learning. *Psychonomic Science, 4*(3), 123-124.
http://dx.doi.org/10.3758/BF03342209

Abstract

An audiovisual stimulus was made contingent upon the rat's licking at the water spout, thus making it analogous with a gustatory stimulus. When the audiovisual stimulus and the gustatory stimulus were paired with electric shock the avoidance reactions transferred to the audiovisual stimulus, but not the gustatory stimulus. Conversely, when both stimuli were paired with toxin or X-ray the avoidance reactions transferred to the gustatory stimulus, but not the audiovisual stimulus. Apparently stimuli are selected as cues dependent upon the nature of the subsequent reinforcer.

Bait shyness revisited

- repeated trials in which the consumption of some food is followed by the administration of unpleasant electrical shock the rats do not tend to avoid that food

- similar failure of conditioning occurs when the characteristic (taste, smell) of the food that implies nausea is replaced by a vocal sign

https://psychology110hc.wordpress.com/2015/04/16/relation-of-cue-to-consequence-in-avoidance-learning/

Prior knowledge

- one distinguishing feature between the bait shyness learning and the pigeon superstition is the incorporation of prior knowledge that biases the learning mechanism = inductive bias.

- the pigeons in the experiment are willing to adopt any explanation for the occurrence of food.

- the rats "know" that food cannot cause an electric shock and that the co-occurrence of noise with some food is not likely to affect the nutritional value of that food. The rats' learning process is biased toward detecting some kind of patterns while ignoring other temporal correlations between events.

Prior knowledge

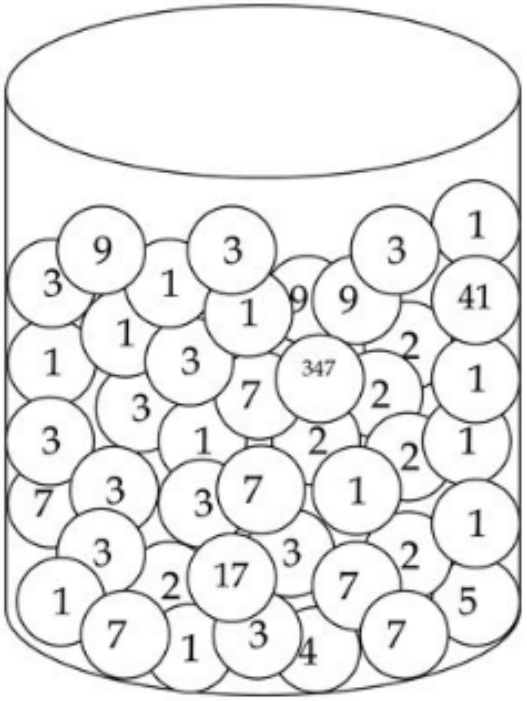
- the incorporation of prior knowledge, biasing the learning process, is inevitable for the success of learning algorithms - “No-Free-Lunch theorem”

- the stronger the prior knowledge (or prior assumptions) that one starts the learning process with, the easier it is to learn from further examples.

- the stronger these prior assumptions are, the less flexible the learning is

Mathematical Analysis of Learning

Induction in an urn (Valiant, 1984)



- consider an urn containing a very large number (millions) of marbles, possibly of different types. You are allowed to draw 100 marbles and asked what kind of marbles the urn contains.

L. G. Valiant, *A theory of the Learnable*, Communications ACM, 27(11):1134-1142, 1984
L. G. Valiant, *Probably Approximately Correct. Nature's Algorithms for Learning and Prospering in a Complex World*, Basic Books, 2013

Induction in an urn (Valiant, 1984)

- consider an urn containing a very large number (millions) of marbles, possibly of different types. You are allowed to draw 100 marbles and asked what kind of marbles the urn contains.

- no assumptions
  - impossible task!

- assumption 1: all the marbles are of different types
  - impossible task!

- assumption 2: all the marbles are identical
  - one single draw is sufficient to solve the task!

Induction in an urn (Valiant, 1984)

- assumption 3: 50% of all marbles are of one type
  - probability to miss that type is $(1/2)^{100} = 7.8 * 10^{-31}$
  - predict that after 100 draws you will have seen representatives for 50% of the urn content

- assumption 4: there are at most 5 different marble types
  - **don't know** the distribution of the marbles
  - could be any distribution: (20%,20%,20%,20%,20%) ,(92%,2%,2%,2%,2%), (49.85%,49.85%,0.1%,0.1%,0.1%), etc
  - predict with 97% confidence that after 100 draws you will have seen representatives for more than 80% of the urn content
  - reasoning: (A) if any of the 5 types occurs with frequency > 5%, the probability to miss that type is < $(1\text{-}0.05)^{100}$ = 0.6%. The probability to miss one type is < 5*0.6% = 3%; (not A) There exists types that occurs with frequency < 5%. There can be at most 4 types with frequency < 5% so the rare marble types are < 20%. Probability to miss the common marble types (which account > 80% of the urn) is < 3%.

PAC learning (Valiant, 1984)

- induction with minimal assumption is very powerful, achieve a useful level of generalization knowing that there a fixed small number of marbles in the urn

- two sources of errors:

  (1) rarity: rare types of marbles are unlikely to be drawn in any small samples

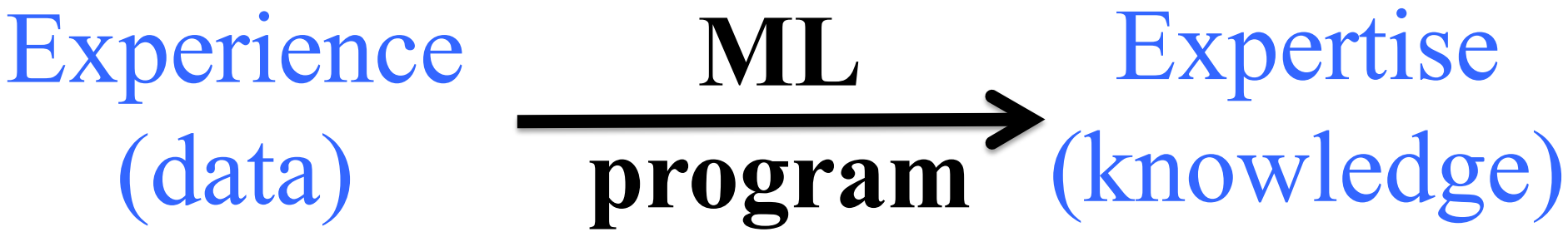
  (2) misfortune: with small probability the sample drawn will be unrepresentative of the contents of the urn because it missed some common marble types

  – neither of these two sources of errors can be totally eliminated BUT we can controlled them by increasing the number of marbles drawn.

- PAC learning: probably approximately correct learning
  – "probably" – misfortune errors
  – "approximately" – rarity errors

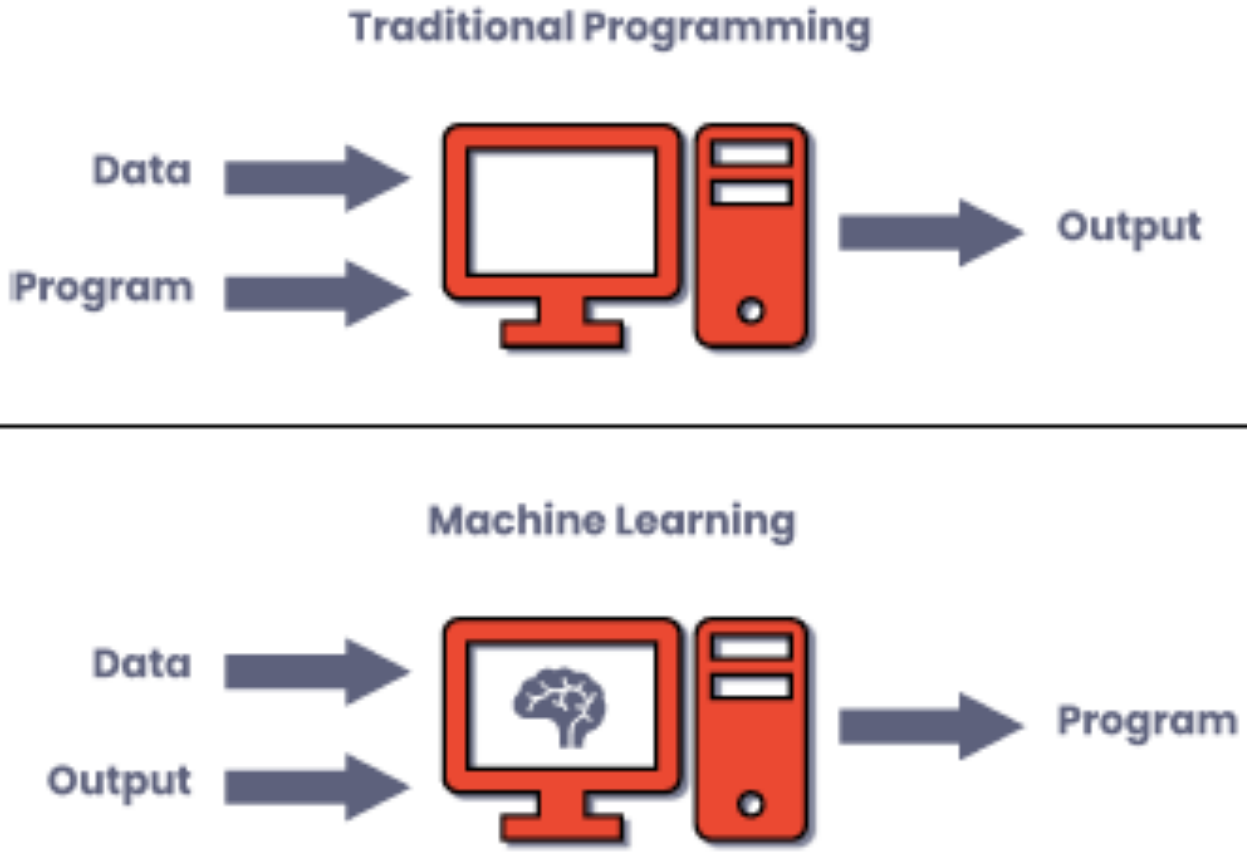
What is Machine Learning?

What is Machine Learning?



"Machine Learning" as an Engineering Paradigm: Use data and examples, instead of expert knowledge, to automatically create systems that perform complex tasks

What is Machine Learning?



Slide Credit: Sparktech

Machine Learning Everywhere

Spam filtering

Machine translation

Speech recognition

Advertising and ad placement

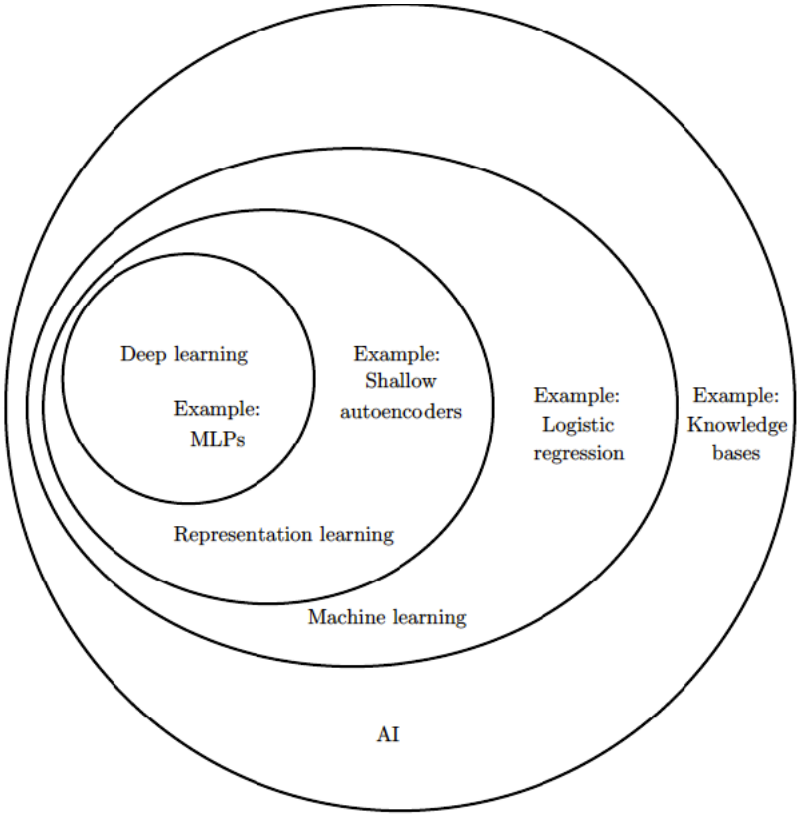
Recommendation systems

Driving assistance systems

Why do we need machine learning?

- Tasks that are too complex to program!
- Computer vision: we know to detect objects but have no idea how we do it!
- Search engines: a human can’t read the entire internet!
- Adaptivity and speed of development

Machine Learning in AI



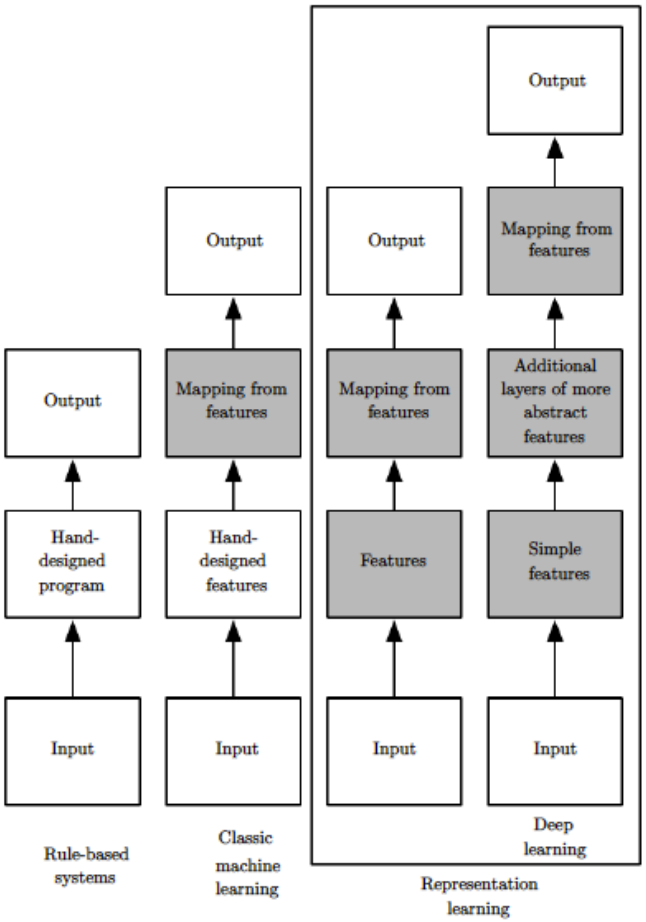


Figure 1.5: Flowcharts showing how the different parts of an AI system relate to each other within different AI disciplines. Shaded boxes indicate components that are able to learn from data.

I. Goodfellow, Y. Bengio and A. Courville, Deep Learning, MIT, 2016, http://www.deeplearningbook.org

Machine Learning vs. Statistics

Machine learning is essentially a form of applied statistics with increased emphasis on the use of computers to statistically estimate complicated functions and a decreased emphasis on proving confidence intervals around these functions

Statistics	Computer Science	Meaning
estimation	learning	using data to estimate an unknown quantity
classification	supervised learning	predicting a discrete Y from $X \in \mathcal{X}$
clustering	unsupervised learning	putting data into groups
data	training sample	$(X_1, Y_1), \ldots, (X_n, Y_n)$
covariates	features	the X_i's
classifier	hypothesis	a map from covariates to outcomes
hypothesis	—	subset of a parameter space Θ
confidence interval	—	interval that contains unknown quantity with a prescribed frequency
directed acyclic graph	Bayes net	multivariate distribution with specified conditional independence relations
Bayesian inference	Bayesian inference	statistical methods for using data to update subjective beliefs
frequentist inference	—	statistical methods for producing point estimates and confidence intervals with guarantees on frequency behavior
large deviation bounds	PAC learning	uniform bounds on probability of errors

Machine Learning vs. Statistics

Differences:

- algorithmic consideration play a major role in ML as we develop algorithms to perform the learning tasks and are concerned with their computational efficiency (*computational complexity)*

- in statistics we are interested in asymptotic behavior (number of training examples grows to infinity) while in ML we work with finite sample sizes (*sample complexity*)

- in ML we consider working under a "distribution-free" setting, where the learner assumes as little as possible about the nature of the data distribution

Course goals

- First goal: to provide a rigorous, yet easy to follow, introduction to the main concepts underlying machine learning:
  - *What is learning?*
  - *How can a machine learn?*
  - *How do we quantify the resources needed to learn a given concept?*
  - *Is learning always possible?*
  - *Can we know if the learning process succeeded or failed?*

- Second goal: present several key machine learning algorithms with strong theoretical foundations

Computational resources of learning

For learning we need 2 type of resources:

1. *Information = training data*
   - analyze in the first part of the course how much training data (sample size) we need in order to learn
   - *sample complexity*

2. *Computation = runtime*
   - for how much time an algorithm (that implements learning) will run, once we have sufficiently many training examples
   - *computational complexity*
   - crucial when we need fast ML applications (driver surveillance, stock exchange trading, etc)
   - runtime = number of elementary instructions executed - arithmetic operations over real numbers - in an asymptotic sense (with respect to input size) of the algorithm, e.g. *O(n) – where n is the size of the input size*

Course Structure – Part 1

- What is learning?
  - Probably Approximately Correct (PAC) model - Vaillant 1984
- How can a machine learn?
  - Empirical Risk Minimization (ERM)
- Resources needed to learn a given concept?
  - sample complexity, time complexity
- Is learning always possible? Did the learning process succeeded?
  - “no free lunch” theorem

Course Structure – Part 2

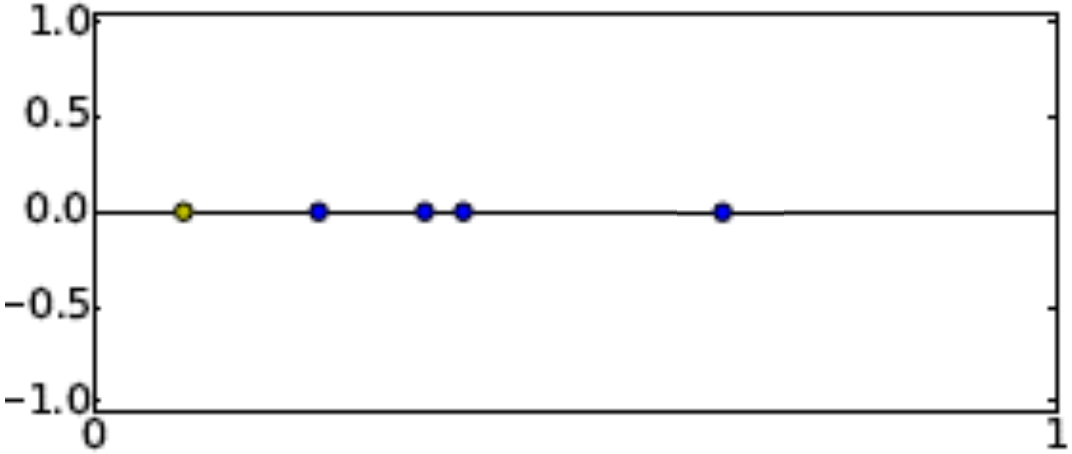
- Survey of prominent methods and approaches with strong theoretical foundations such as:
  - Boosting
  - SVMs
  - neural networks? (loose bounds, work in progress)
  - etc

Usefulness of Theoretical Machine Learning

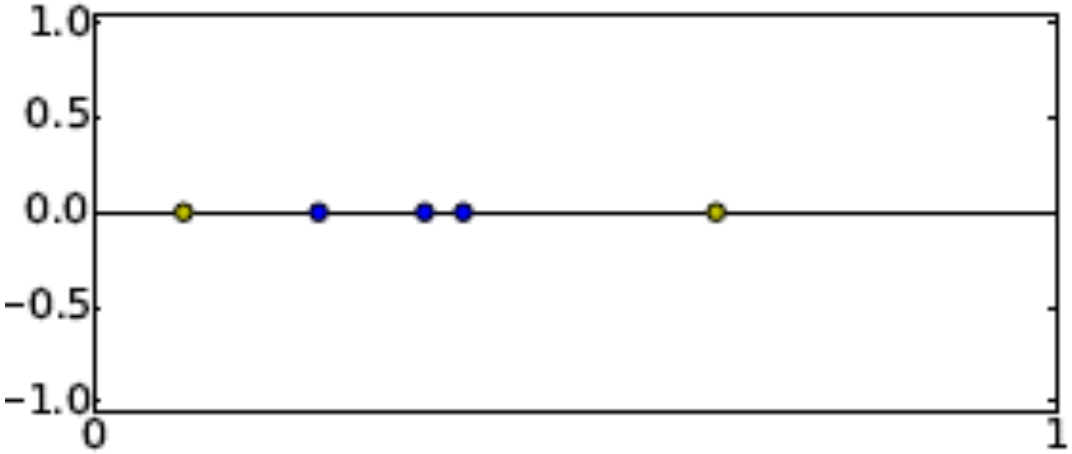
Perceptron in 1D

Class +1

Class -1



Linear separable

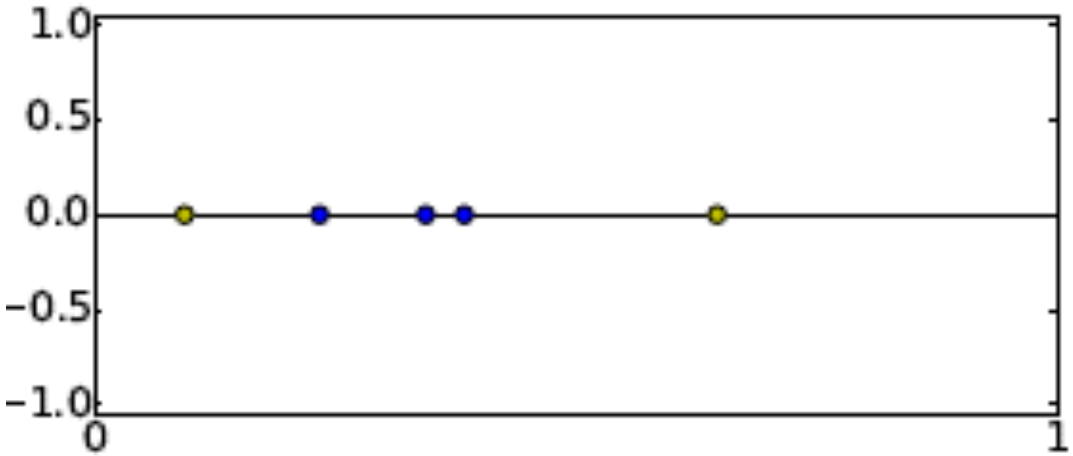


Non-linear separable

Learning sin(λx) in 1D

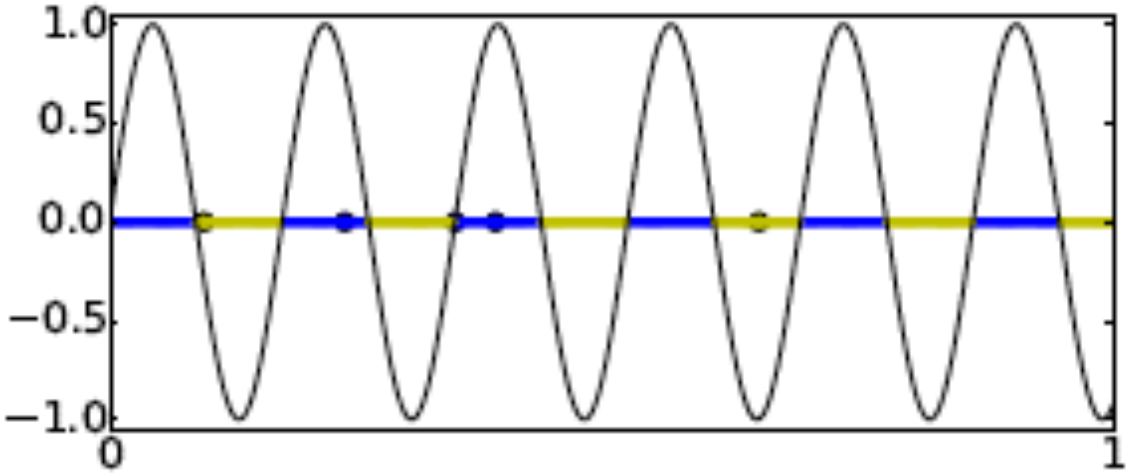
Class +1

Class -1



Non-linear separable

Can you think of an algorithm learning λ for solving the problem?
Will it generalize (small generalization error)?



Theoretical result: cannot learn λ with small generalization error

No Free Lunch theorem

- averaged over all possible data generating distributions, every classification algorithm has the same error rate when classifying previously unobserved points.

- if we make assumptions about the kinds of probability distributions we encounter in real-world applications, then we can design learning algorithms that perform well on these distributions.

- the goal of machine learning research is not to seek a universal learning algorithm or the absolute best learning algorithm. Instead, the goal is to understand what kinds of distributions are relevant to the "real world" that an AI agent experiences and what kinds of machine learning algorithms perform well on data drawn from the kinds of data generating distributions we care about.

Application of Machine Learning: AdaBoost for face detection

ACCEPTED CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION 2001

Rapid Object Detection using a Boosted Cascade of Simple Features

Paul Viola
viola@merl.com
Mitsubishi Electric Research Labs
201 Broadway, 8th FL
Cambridge, MA 02139

Michael Jones
mjones@crl.dec.com
Compaq CRL
One Cambridge Center
Cambridge, MA 02142

Abstract

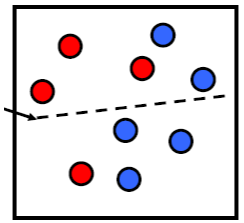
This paper describes a machine learning approach for visual object detection which is capable of processing images

tected at 15 frames per second on a conventional 700 MHz Intel Pentium III. In other face detection systems, auxiliary information, such as image differences in video sequences, or pixel color in color images, have been used to achieve

AdaBoost Algorithm

- Given example images $(x_1, y_1), \ldots, (x_n, y_n)$ where $y_i = 0, 1$ for negative and positive examples respectively.
- Initialize weights $w_{1,i} = \frac{1}{2m}, \frac{1}{2l}$ for $y_i = 0, 1$ respectively, where m and l are the number of negatives and positives respectively.
- For $t = 1, \ldots, T$:
  1. Normalize the weights,
  $$w_{t,i} \leftarrow \frac{w_{t,i}}{\sum_{j=1}^{n} w_{t,j}}$$
  so that w_t is a probability distribution.
  2. For each feature, j, train a classifier h_j which is restricted to using a single feature. The error is evaluated with respect to w_t, $\epsilon_j = \sum_i w_i \left| h_j(x_i) - y_i \right|$.
  3. Choose the classifier, h_t, with the lowest error ϵ_t.
  4. Update the weights:
  $$w_{t+1,i} = w_{t,i} \beta_t^{1-e_i}$$
  where $e_i = 0$ if example x_i is classified correctly, $e_i = 1$ otherwise, and $\beta_t = \frac{\epsilon_t}{1-\epsilon_t}$.
- The final strong classifier is:
$$h(x) = \begin{cases} 1 & \sum_{t=1}^{T} \alpha_t h_t(x) \geq \frac{1}{2} \sum_{t=1}^{T} \alpha_t \\ 0 & \text{otherwise} \end{cases}$$
where $\alpha_t = \log \frac{1}{\beta_t}$

Start with uniform weights on training examples



$\{x_1, \ldots x_n\}$

For T rounds

Evaluate *weighted* error for each feature, pick best.

Re-weight the examples:
Incorrectly classified -> more weight
Correctly classified -> less weight

Final classifier is combination of the weak ones, weighted according to error they had.

Freund & Schapire 1995

Theoretical Machine Learning

- understand the main concepts underlying machine learning through basic theory
  - the goal is not detailed theory and theorem-proving;
  - emphasis on concepts, less on specific algorithms.

- know the prominent methods used in contemporary machine learning

- learn how to use machine learning *correctly*

- no programming, do ML problems in seminar

Next time

- will cover Chapters 2 and 3 from the book
- do exercises in seminar (from the ones proposed)