Introdução a Boosting

Ramon Durães

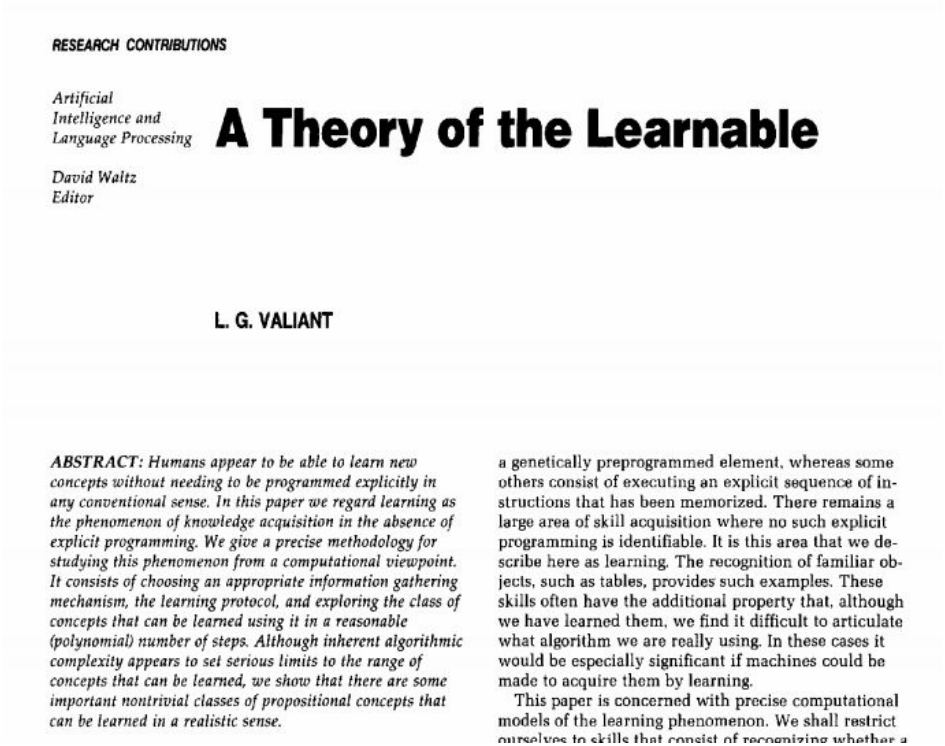
Apresentação

1. PAC Learning
2. Weak Learners
3. Boosting
4. ADABoost
    a. Implementação
5. Gradient Boosting
6. XGBoost

PAC Learning

- Trabalho por L. G. Valiant, 1984

- Uma forma quantitativa de definir se um problema é “resolvível”



RESEARCH CONTRIBUTIONS

Artificial Intelligence and Language Processing

David Waltz Editor

A Theory of the Learnable

L. G. VALIANT

ABSTRACT: *Humans appear to be able to learn new concepts without needing to be programmed explicitly in any conventional sense. In this paper we regard learning as the phenomenon of knowledge acquisition in the absence of explicit programming. We give a precise methodology for studying this phenomenon from a computational viewpoint. It consists of choosing an appropriate information gathering mechanism, the learning protocol, and exploring the class of concepts that can be learned using it in a reasonable (polynomial) number of steps. Although inherent algorithmic complexity appears to set serious limits to the range of concepts that can be learned, we show that there are some important nontrivial classes of propositional concepts that can be learned in a realistic sense.*

a genetically preprogrammed element, whereas some others consist of executing an explicit sequence of instructions that has been memorized. There remains a large area of skill acquisition where no such explicit programming is identifiable. It is this area that we describe here as learning. The recognition of familiar objects, such as tables, provides such examples. These skills often have the additional property that, although we have learned them, we find it difficult to articulate what algorithm we are really using. In these cases it would be especially significant if machines could be made to acquire them by learning.

This paper is concerned with precise computational models of the learning phenomenon. We shall restrict ourselves to skills that consist of recognizing whether a

PAC Learning

- **P**robably **A**pproximately **C**orrect Model:

PAC Learning

- **P**robably **A**pproximately **C**orrect Model:
  - **C**orrect Model:
    - Um modelo que sempre acerta as predições em dados novos.
    - Só dá pra saber isso vendo todas as observações, o que vira só “memorização”, não aprendizado.

PAC Learning

- **P**robably **A**pproximately **C**orrect Model:
  - **C**orrect Model:
    - Um modelo que sempre acerta as predições em dados novos.
    - Só dá pra saber isso vendo todas as observações, o que vira só “memorização”, não aprendizado.
  - **A**pproximately **C**orrect Model:
    - Dá pra treinar num conjunto de treinamento e estimar a acurácia do modelo!
    - Mas não dá pra garantir que o erro é sempre tolerável... Os resultados só se aplicam pro conjunto de treinamento.

PAC Learning

- O objetivo é limitar o erro real utilizando o erro de treinamento

  m pontos gerados pela distribuição real **D**,

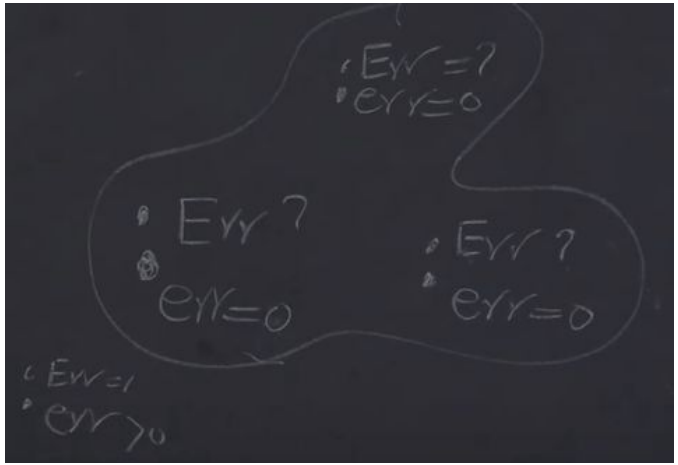
  D é a distribuição em mãos (amostrada)

  y = c(x) : função real, constante

  yhat = h(x) : função estimada, hipótese. h ϵ H

  err: erro estimado **Err** : erro real

PAC Learning



Qual a probabilidade de um classificador que resolveu bem o problema pros dados em mãos ser, na verdade, ruim? (Erro real maior que um threshold)

$$P(Err > \varepsilon) < |H|e^{-\varepsilon m}$$

PAC Learning

Theorem:
If the hypothesis space H is finite, and D is a sequence of $m \geq 1$ independent random examples of some target concept c, then for any $0 \leq \epsilon \leq 1$, the probability that $VS_{H,D}$ contains a hypothesis with error greater than ϵ is less than

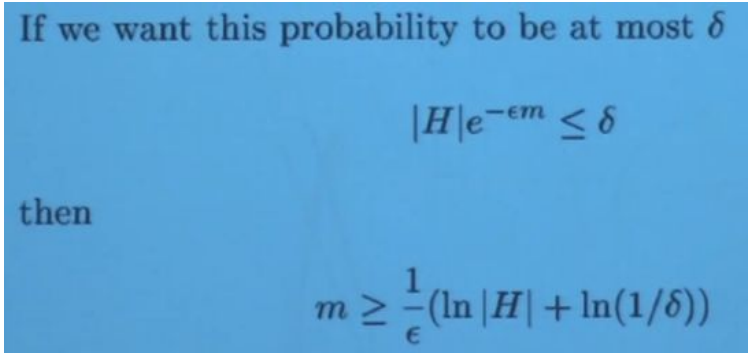
$$|H|e^{-\epsilon m}$$

PAC Learning

- **P**robably **A**pproximately **C**orrect Model:
  - **C**orrect Model:
    - Um modelo que sempre acerta as predições em dados novos.
    - Só dá pra saber isso vendo todas as observações, o que vira só “memorização”, não aprendizado.
  - **A**pproximately **C**orrect Model:
    - Dá pra treinar num conjunto de treinamento e estimar a acurácia do modelo!
    - Mas não dá pra garantir que o erro é sempre tolerável... Os resultados só se aplicam pro conjunto de treinamento.
  - **P**robably **A**pproximately **C**orrect Model:
    - Certifica que a probabilidade de obter um modelo ruim devido a dados de treinamento ruins é baixa (menor que um threshold **1 -** γ).

PAC Learning

- Dá um limite ao tamanho amostral: quantos pontos são necessários para garantir que eu estou “Approximately Correct”



PAC Learning

- Um problema é “PAC Learnable” se um algoritmo de aprendizado consegue encontrar uma solução com erro menor ϵ com uma probabilidade maior que γ.

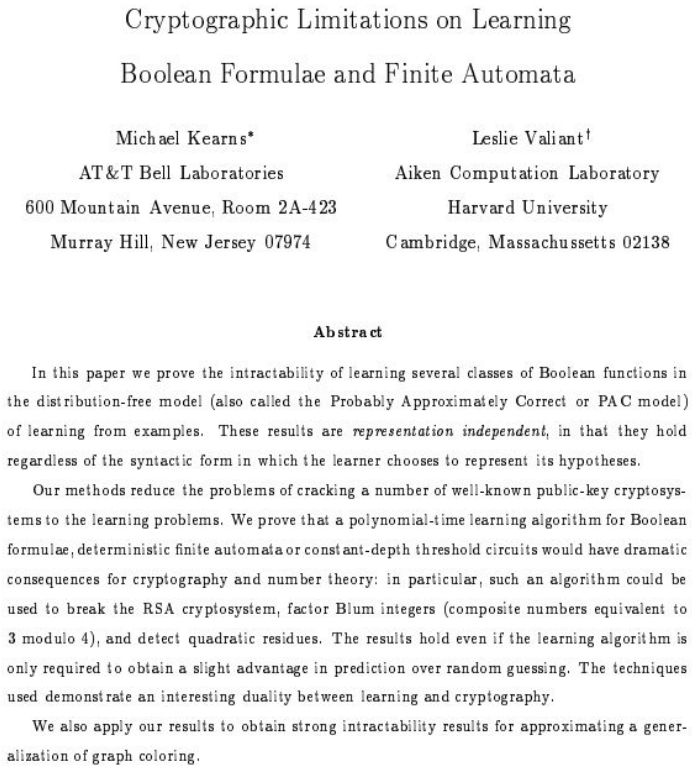
- Um modelo com essas características é chamado de “**strong learner**”.

Weak Learners

- Trabalho de 1988 por Michael Kearns e Leslie Valiant

- Weak learners: algoritmos cuja performance é ligeiramente melhor que o “chute aleatório”

- Ex: árvore de decisão com apenas um nível (“stump”)



Cryptographic Limitations on Learning

Boolean Formulae and Finite Automata

Michael Kearns*
AT&T Bell Laboratories
600 Mountain Avenue, Room 2A-423
Murray Hill, New Jersey 07974

Leslie Valiant†
Aiken Computation Laboratory
Harvard University
Cambridge, Massachussetts 02138

Abstract

In this paper we prove the intractability of learning several classes of Boolean functions in the distribution-free model (also called the Probably Approximately Correct or PAC model) of learning from examples. These results are *representation independent*, in that they hold regardless of the syntactic form in which the learner chooses to represent its hypotheses.

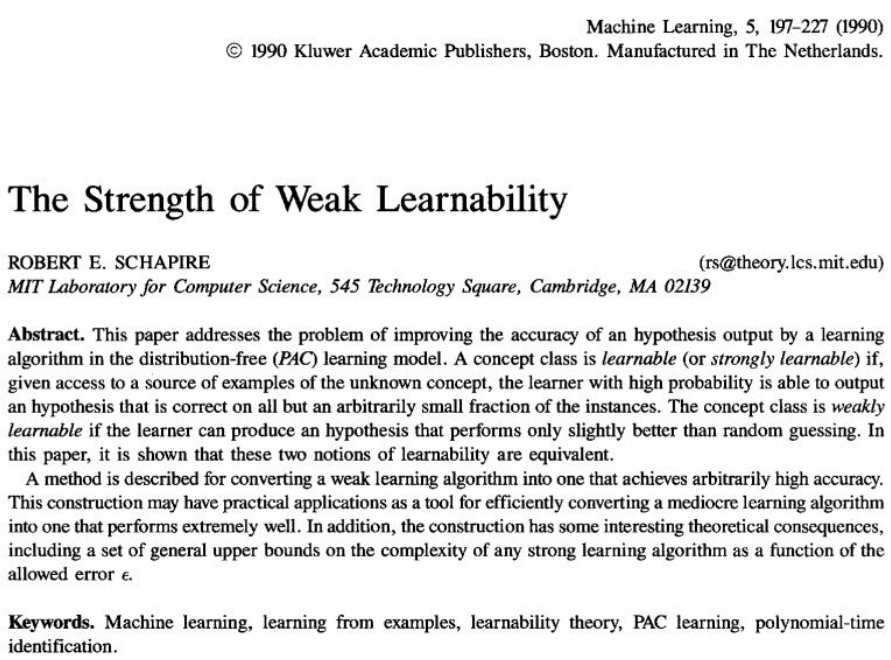
Our methods reduce the problems of cracking a number of well-known public-key cryptosystems to the learning problems. We prove that a polynomial-time learning algorithm for Boolean formulae, deterministic finite automata or constant-depth threshold circuits would have dramatic consequences for cryptography and number theory: in particular, such an algorithm could be used to break the RSA cryptosystem, factor Blum integers (composite numbers equivalent to 3 modulo 4), and detect quadratic residues. The results hold even if the learning algorithm is only required to obtain a slight advantage in prediction over random guessing. The techniques used demonstrate an interesting duality between learning and cryptography.

We also apply our results to obtain strong intractability results for approximating a generalization of graph coloring.

Weak Learners

- Trabalho de 1990 por Robert E Schapire

- Se um problema consegue ser resolvido por um “strong learner”, então os “weak learners” também conseguem, mas de outra forma



Machine Learning, 5, 197–227 (1990)

The Strength of Weak Learnability

ROBERT E. SCHAPIRE (rs@theory.lcs.mit.edu)
MIT Laboratory for Computer Science, 545 Technology Square, Cambridge, MA 02139

Abstract. This paper addresses the problem of improving the accuracy of an hypothesis output by a learning algorithm in the distribution-free (*PAC*) learning model. A concept class is *learnable* (or *strongly learnable*) if, given access to a source of examples of the unknown concept, the learner with high probability is able to output an hypothesis that is correct on all but an arbitrarily small fraction of the instances. The concept class is *weakly learnable* if the learner can produce an hypothesis that performs only slightly better than random guessing. In this paper, it is shown that these two notions of learnability are equivalent.

A method is described for converting a weak learning algorithm into one that achieves arbitrarily high accuracy. This construction may have practical applications as a tool for efficiently converting a mediocre learning algorithm into one that performs extremely well. In addition, the construction has some interesting theoretical consequences, including a set of general upper bounds on the complexity of any strong learning algorithm as a function of the allowed error ϵ.

Keywords. Machine learning, learning from examples, learnability theory, PAC learning, polynomial-time identification.

Hypothesis **Boosting** Mechanism

- Hypothesis: representa “o que o modelo aprendeu”. Exemplo da regressão linear:

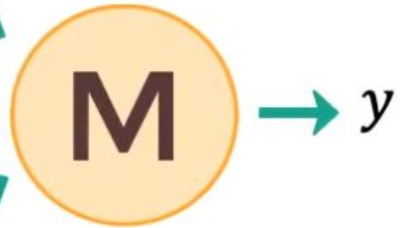
$$y = b + w_1 x_1 + w_2 x_2 + \cdots + w_p x_p$$

- Combinação de várias hipóteses diferentes!

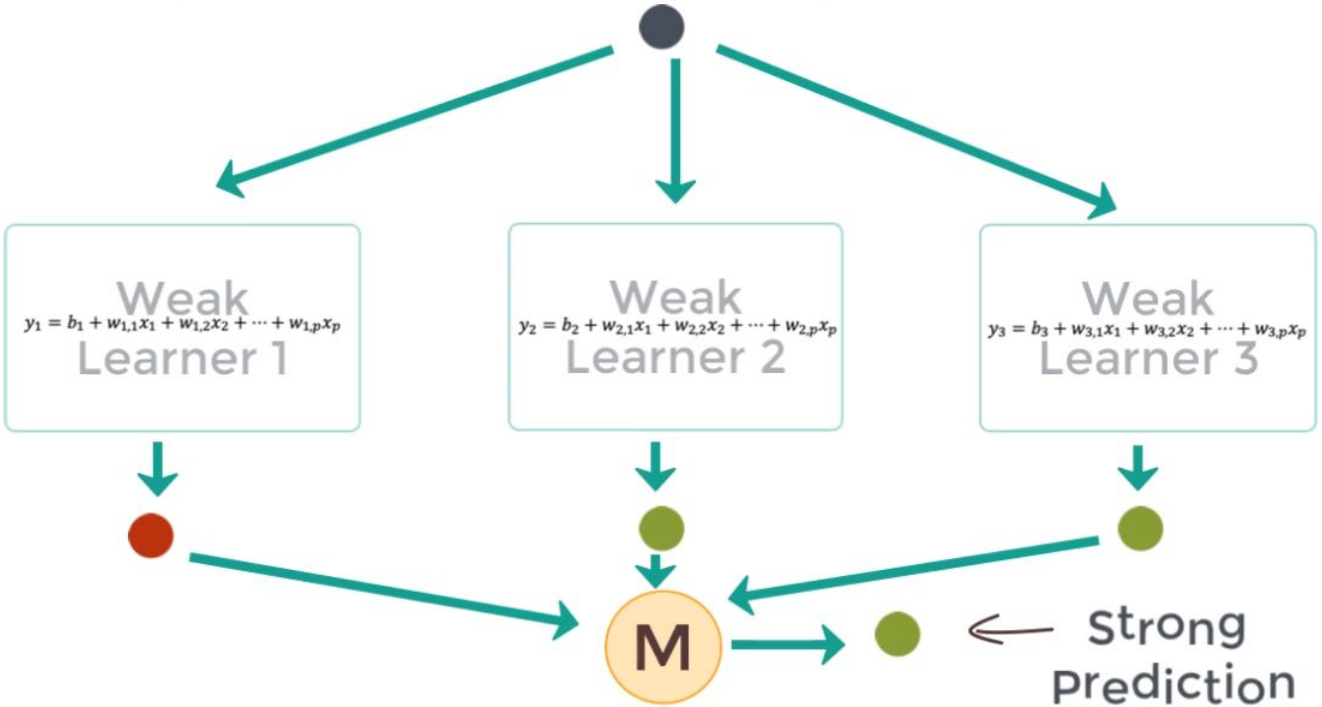
$$y_1 = b_1 + w_{1,1} x_1 + w_{1,2} x_2 + \cdots + w_{1,p} x_p$$

$$y_2 = b_2 + w_{2,1} x_1 + w_{2,2} x_2 + \cdots + w_{2,p} x_p$$

$$y_3 = b_3 + w_{3,1} x_1 + w_{3,2} x_2 + \cdots + w_{3,p} x_p$$

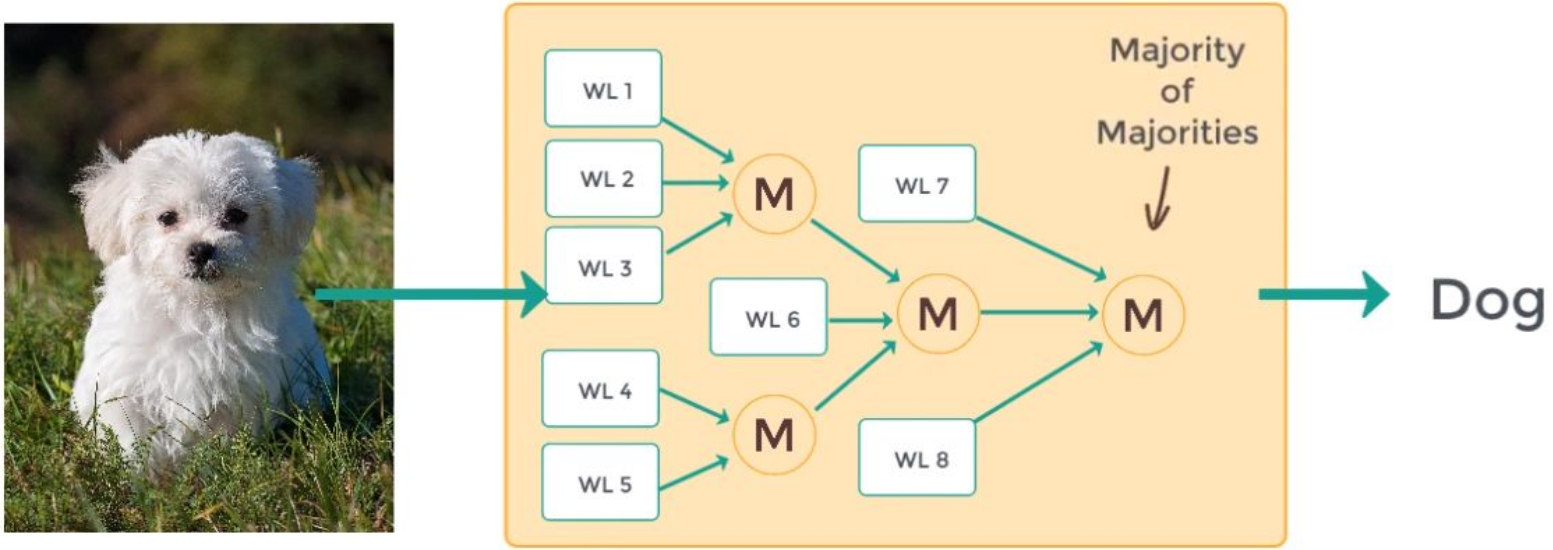


Hypothesis **Boosting** Mechanism



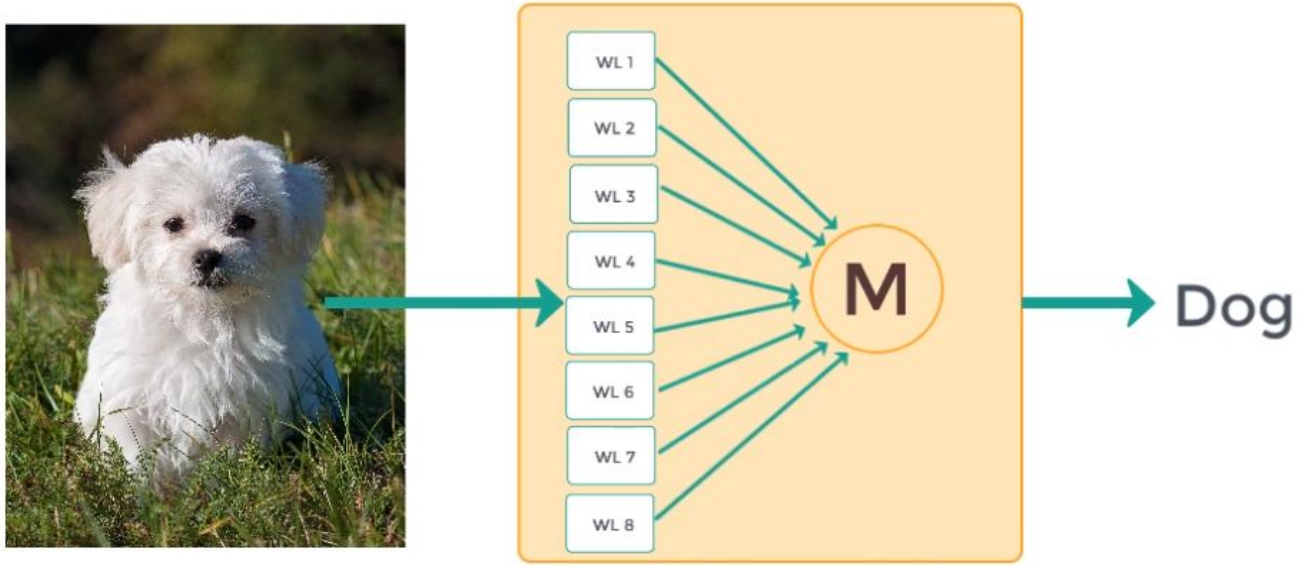
Hypothesis **Boosting** Mechanism

- Para problemas muito complexos, essas hipóteses seriam combinadas 3 a 3 por voto majoritário:



Hypothesis **Boosting** Mechanism

- Modificação proposta por pesquisadores do AT&T em 1995:



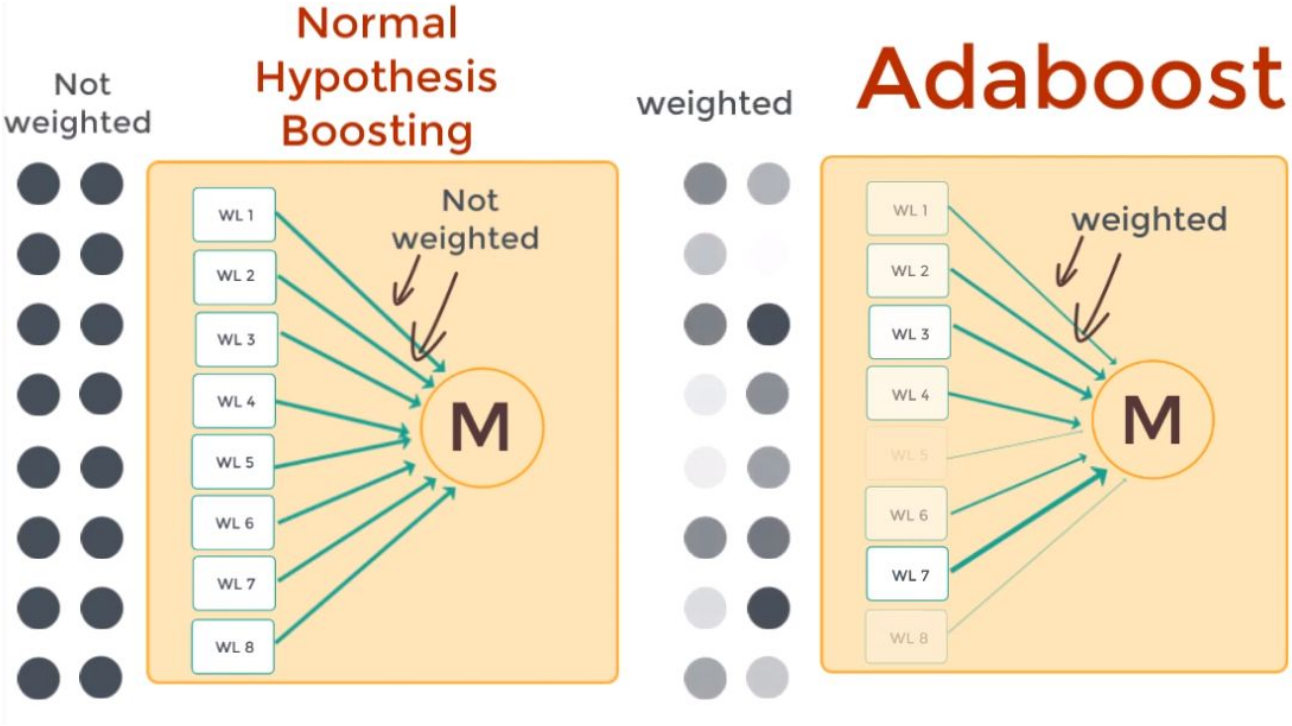
Hypothesis **Boosting** Mechanism

- “Meta algorithm”: uma forma de combinar modelos

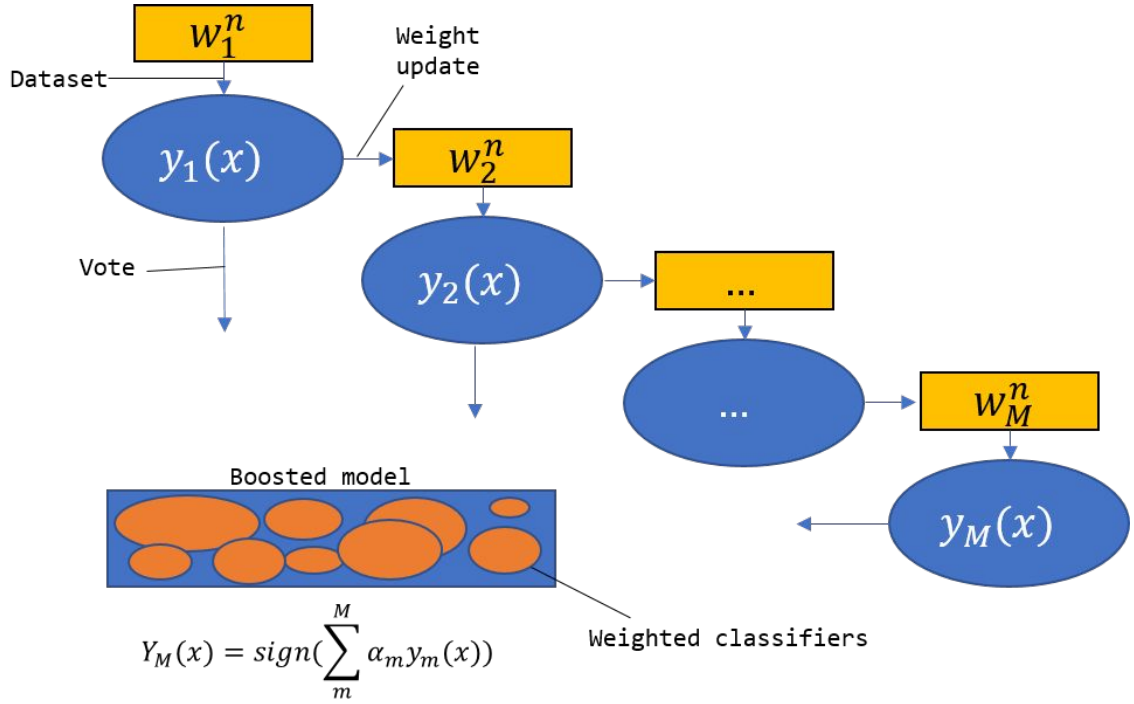
- “Ensemble model”

- Modelos treinados de forma sequencial: a criação de um modelo depende dos resultados do modelo anterior

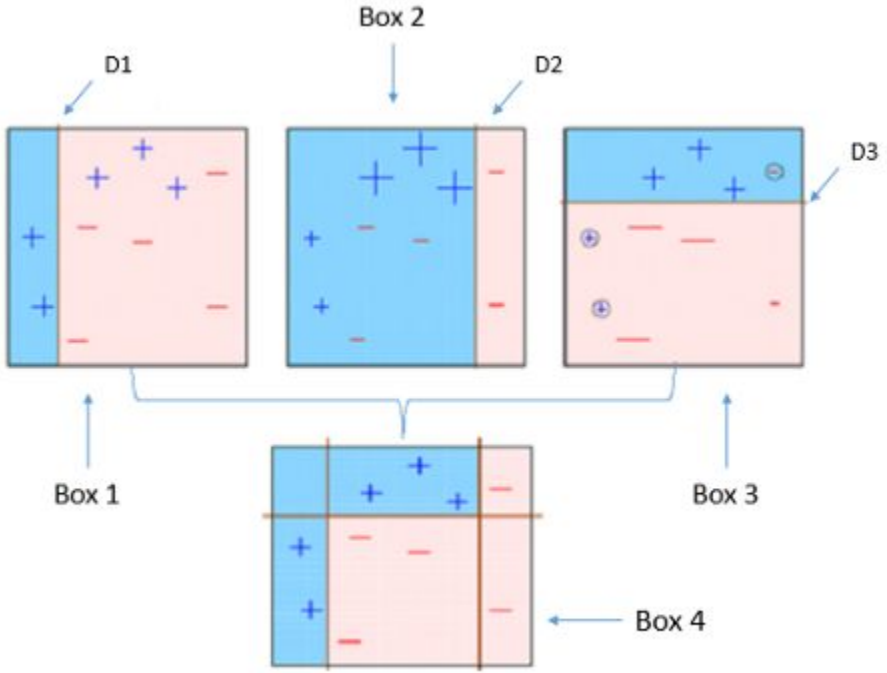
Adaptive Boosting: ADABOOST



ADABOOST


w_1^n
Dataset
Weight update
$y_1(x)$
w_2^n
Vote
$y_2(x)$
...
...
w_M^n
$y_M(x)$
Boosted model
Weighted classifiers
$Y_M(x) = sign(\sum_m^M \alpha_m y_m(x))$

ADABOOST



ADABOOST

1. Combina vários “weak learners” para fazer a classificação;
2. Alguns “learners” têm mais peso na classificação final que outros;
    a. O peso dos modelos é proporcional à acurácia
3. Os “learners” são treinados de forma sequencial, de forma que os erros do modelo anterior influenciam o treinamento do modelo atual.

ADABOOST: Pseudo-código

1. Sendo **n** o número de observações, inicializar todos os pesos **w** como: $w = 1/n$
2. Sendo **T** o número total de modelos: enquanto **t < T**:
   a. Criar modelo e obter hipótese $h_t(x_n) \forall x_n$
   b. Calcular o erro **ϵ** do conjunto de treinamento com a fórmula (soma dos pesos das observações classificadas incorretamente):

$$\epsilon_t = \frac{\sum_{n=1}^{N} w_n^{(t)} * I(y_n \neq h_t(x_n))}{\sum_{n=1}^{N} w_n^{(t)}}$$

   c. Computar ***α*** (“amount of say”: peso do modelo **t**) como:

$$\alpha_t = \frac{1}{2} * log(\frac{1-\epsilon_t}{\epsilon_t})$$

   d. Atualizar os pesos para as **N** observações de treinamento no próximo modelo (**t+1**):

$$w_n^{(t+1)} = w_n^t * exp(\alpha_t * I(y_n \neq h_t(x_n))))$$

3. Após **T** iterações, calcular a saída com: $f(x) = sign(\sum_t^T \alpha_t * h_t(x))$

ADABOOST: Exemplo Computacional

Gradient Boosting

- Usa os gradientes pra tudo
- Observações cuja predição precisa melhorar: gradiente alto!
- Técnica muito conhecida e que funciona muito bem
- Problemas:
  - em casos complexos, com muitos dados, cálculo dos gradientes pode se tornar muito lento;
  - por ter tantos parâmetros, o algoritmo é suscetível a overfitting

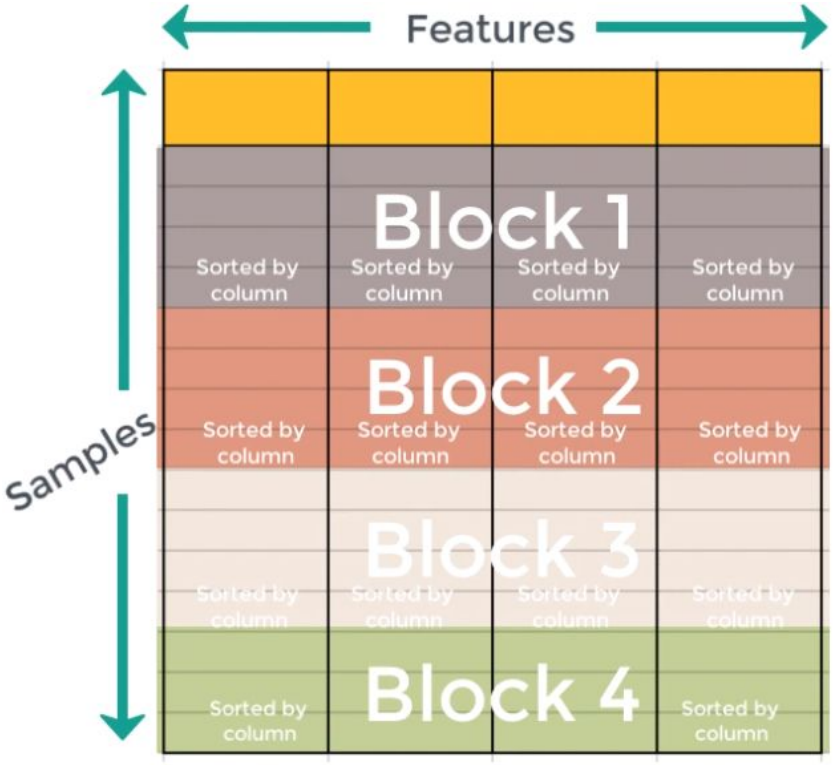
$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]$$

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) - \gamma \nabla_{F_{m-1}} L(y_i, F_{m-1}(x_i))\right)$$

XGBoost (EXtreme Gradient Boosting)

- É “Gradient Boosting” mas implementado de uma forma muito eficiente.

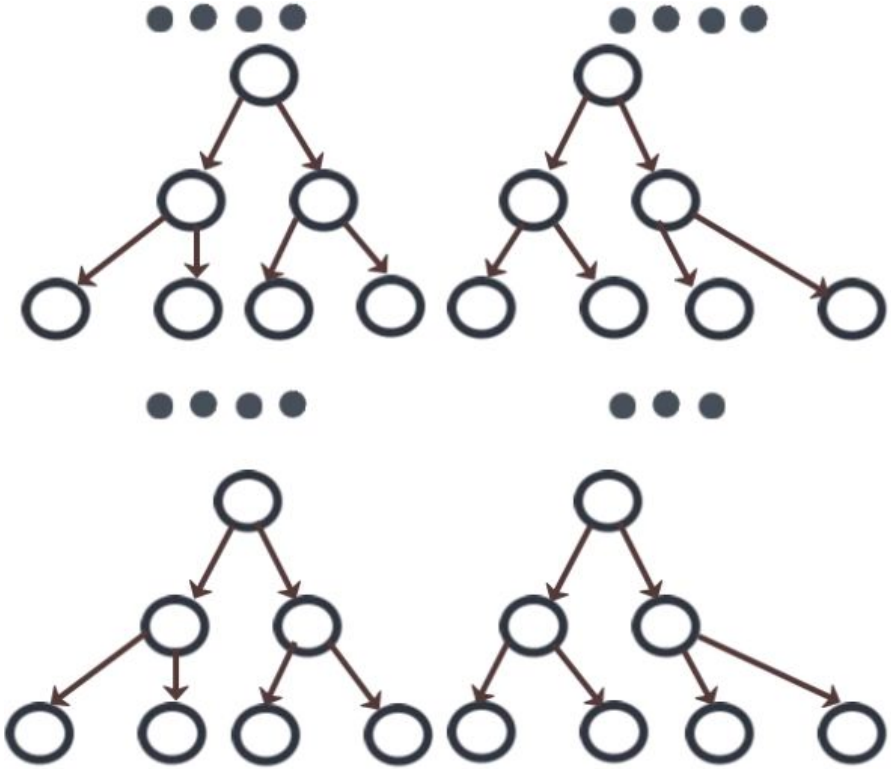
- Os dados são guardados de maneira inteligente:
  - formato comprimido e esparso (.csc)
  - ordenados por valor, por bloco
  - podem ser divididos em blocos.



XGBoost

- Os blocos de dados podem ser processados paralelamente;

- Os nós das árvores de decisão são computados de forma “breadth-first”, o que permite que esse cálculo seja também seja paralelizável.



XGBoost

Em 2015, 17 das 29 soluções vencedoras de competições do Kaggle utilizaram alguma forma de Gradient Boosting.

Conclusões

- Boosting não é um modelo, mas uma metodologia de combinação de modelos.
- É muito importante para essa técnica que os modelos utilizados sejam “weak learners”. Boosting em modelos complexos leva a overfitting.
- Adaboost é uma forma de boosting que adapta os pesos das observações de acordo com o erro do modelo anterior. O voto majoritário também é ponderado de acordo com a performance do modelo.
- Técnicas de Gradient Boosting representam o estado da arte de diversos problemas de machine learning. XGBoost é uma implementação extremamente eficiente desta técnica.