Curso 2 – CD, AM e DM
MBA EM IA e BIGDATA

COMITÊS DE CLASSIFICADORES
RANDOM FOREST
BOOSTING (XGB)

PROFA. ROSELI AP. FRANCELIN ROMERO

SCC – ICMC - USP

COMITÊS DE CLASSIFICADORES

- Comites de Classificadores
- Random Forest
- Boosting
- XGB

COMITÊS DE CLASSIFICADORES

- Procuram melhorar acurácia combinando predições de múltiplos estimadores
- Classificação

  Constroem conjunto de classificadores a partir de dados de treinamento

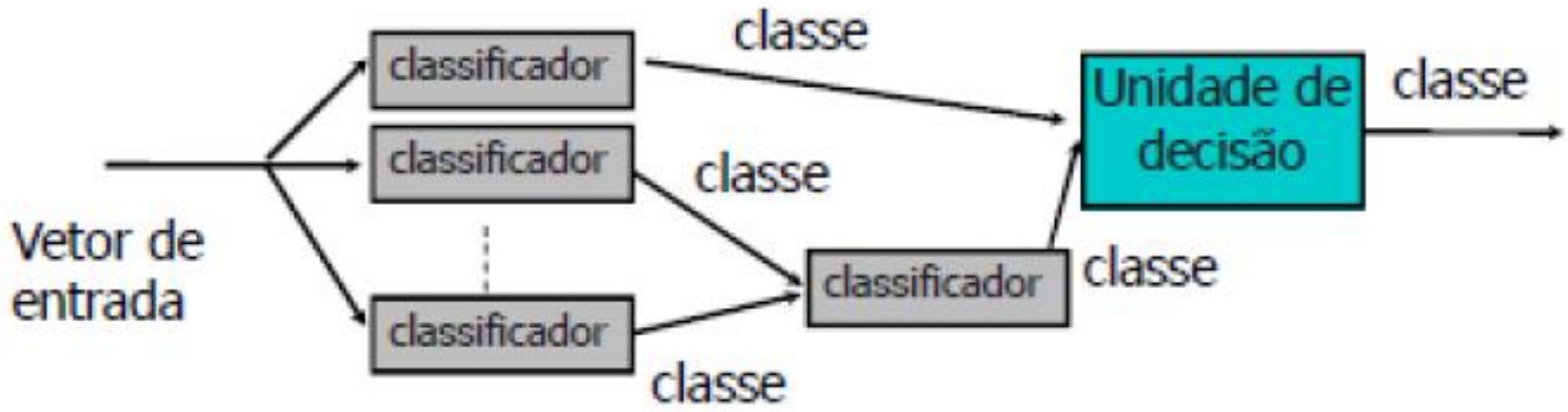
  - Classificadores (base)
  - Classe do novo exemplo é definida pela agregação da predição dos múltiplos classificadores (base)
- Também podem ser usados em tarefas de regressão e de agrupamento de dados

COMITÊS DE CLASSIFICADORES

- Treinamento independente
- Algoritmos aplicados a:
  - Mesmo conjunto de dados
  - Conjuntos de dados formados por diferentes amostras do conjunto de dados original
  - Conjuntos de dados com diferentes atributos preditivos do conjunto de dados originais
- Explora semelhanças e diferenças

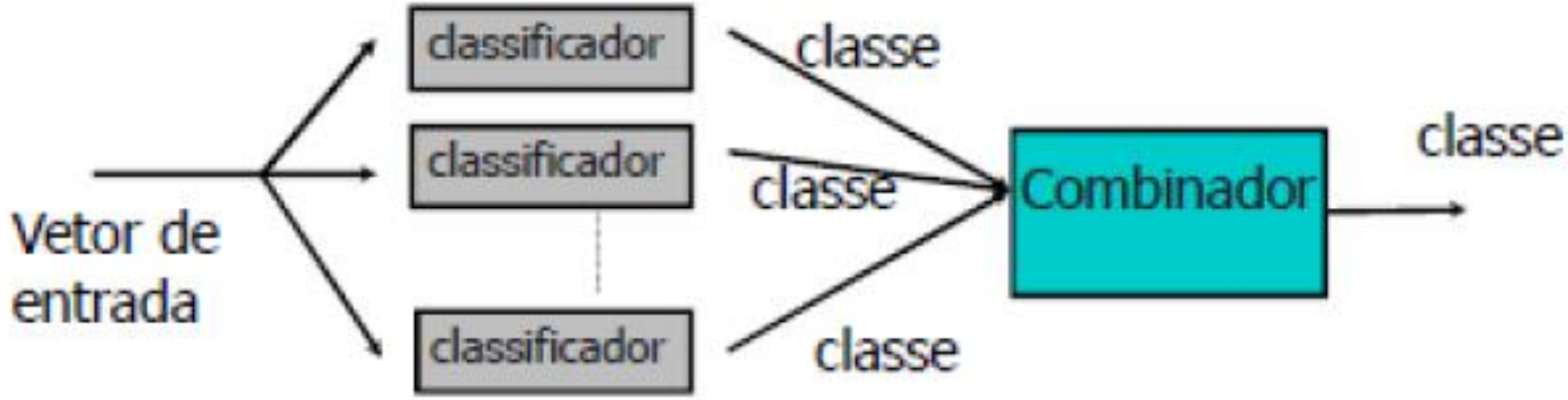
COMITÊS DE CLASSIFICADORES

- Combinação hierárquica
- Mistura das combinações anteriores



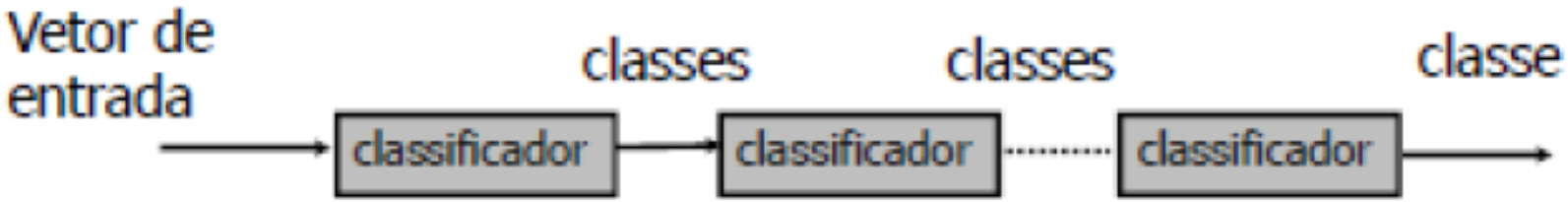
COMITÊS DE CLASSIFICADORES

- Combinação Paralela



COMITÊS DE CLASSIFICADORES

- Combinação sequencial
- Saída de um classificador é utilizada como entrada para o próximo classificador
- Não precisa combinar saídas
- Problema: propagação de erro



COMITÊS DE CLASSIFICADORES

- Combinação de predições
  - classe majoritária
  - Voto (média)
  - Voto (média) ponderado
  - Algoritmo combinador

RANDOM FOREST (RF)

- Combinar ADs, mas pode usar modelos gerados por qualquer

algoritmo de AM

- Combina k ADs
- Cada arvore é induzida usando um subconjunto aleatório dos atributos

Preditivos usado na escolha do atributo para cada no

- Hiper-parâmetros definem número de ADs e número de atributos preditivos para cada AD
- Classificação ocorre por voto majoritário.

RANDOM FOREST (RF)

- N é numero de atributos do conjunto de dados
- RFs usa bootstrap para selecionar exemplos de treinamento
- Várias alternativas para escolher aleatoriamente os atributos preditivos:
  - Forest-RI (Random Input Selection)
  - Forest-RC (Random Combination)

Random Forest (Random Input Selection)

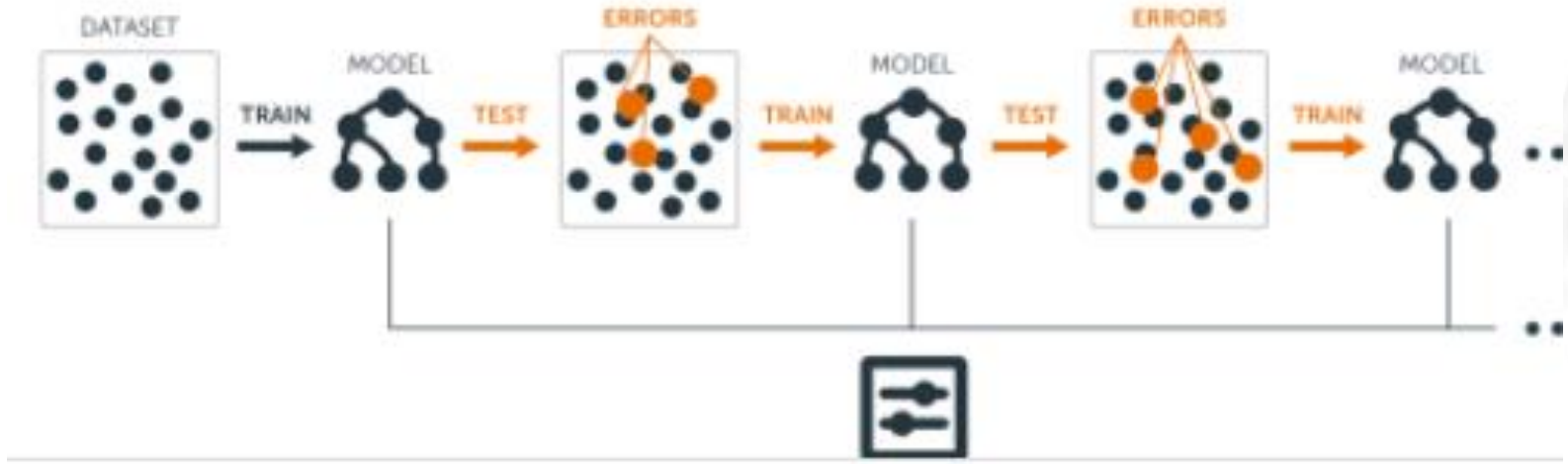
- Seleciona aleatoriamente, para cada nó, um subconjunto de F atributos preditivos
- Algoritmo CART e usado para crescer as arvores sem poda (serie de divisões binárias, cujos nós terminais são descritos por regras.
- Problema: conjunto de dados com poucos atributos preditivos

  Pode selecionar atributos fortemente correlacionados

Forest-RC (Random Combination)

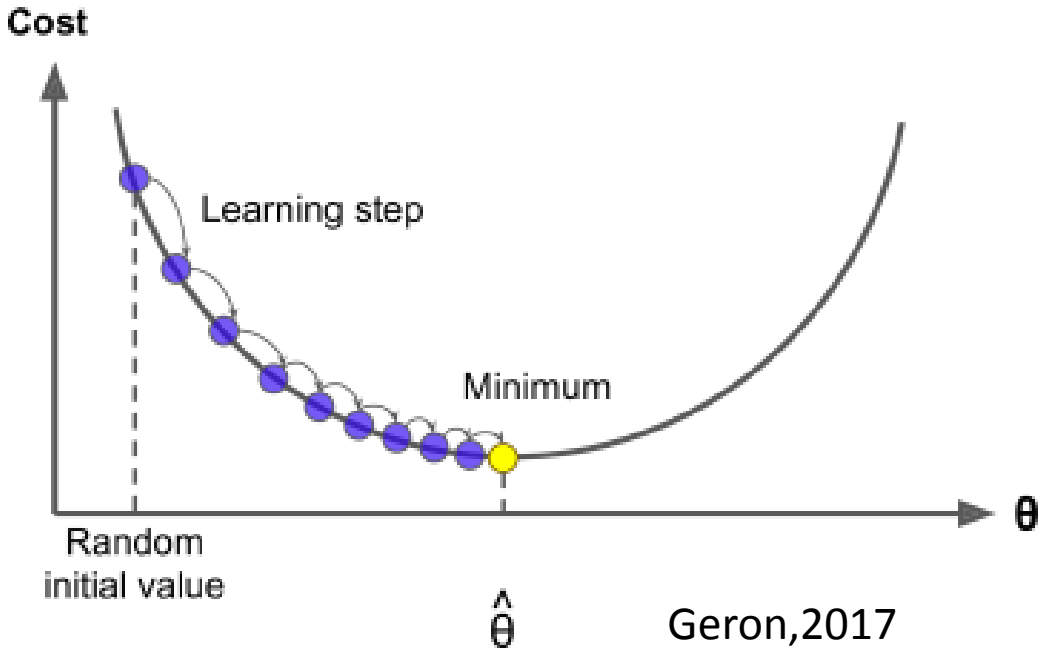
- Expande número de atributos criando combinações lineares aleatórias de atributos
- A cada nó, F combinações de L atributos são aleatoriamente geradas
- Combina atributos utilizando pesos aleatoriamente gerados entre -1 e +1
- Cada combinação é um novo atributo
- Usada quando conjunto de dados tem poucos atributos preditivos

Boosting



Boosting

- Gradiente Descendente



Geron,2017

XGBOOST

- Combina Árvores geradas pelo algoritmo CART
- Treinamento aditivo
  - Induz uma arvore
  - Inclui no ensemble
  - Induz próxima arvore
  - …
- Pondera a resposta de cada Árvore para reduzir complexidade do modelo

COMITÊS DE CLASSIFICADORES

- Combinação de estimadores em geral aumenta desempenho preditivo
- E reduz variância
- As vezes é chamado de meta-aprendizado

Referências

- **LIVRO. Mitchell, T. M., & Learning, M. (1997). Mcgraw-hill.**
- LIVRO. Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.
- LIVRO. Von Luxburg, U., & Schölkopf, B. (2011). Statistical learning theory: Models, concepts, and results. In Handbook of the History of Logic (Vol. 10, pp. 651-706). North-Holland.
- Trevor Hastie - Gradient Boosting Random Forests at H2O World 2014 (YouTube)
- Trevor Hastie - Data Science of GBM (2013) (slides)
- Mark Landry - Gradient Boosting Method and Random Forest at H2O World 2015 (YouTube)
- Peter Prettenhofer - Gradient Boosted Regression Trees in scikit-learn at PyData London 2014 (YouTube)