GEAM
Grupo de Estudos de Aprendizado de Máquina

Introdução ao Algoritmo de Boosting

- Bolsista: Ítalo Lima Dantas
- Curso: Engenharia de Software
- Orientador: Regis Pires Magalhães

Sumário

Algoritmo Boosting

- No encontro de hoje, estudaremos o conceito de Algoritmo de Boosting, com introdução ao AdaBoost, também conhecido como Adaptive Boosting.

- O introdução ao estudo do Algoritmo será contextualizada, explicada e exemplificada na resolução de um problema na ferramenta Orange Data Mining.

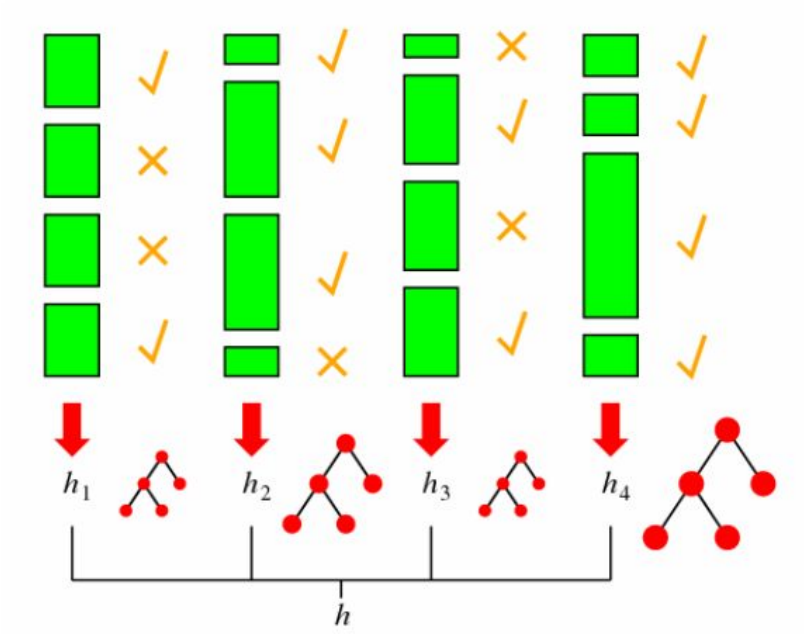
- A ideia do algoritmo de agrupamento Boosting é construir hipóteses sucessivas, de modo que as hipóteses seguintes devem ser melhoradas, de acordo com os erros das hipóteses anteriores. Esse processo é feito através de pesos que são atribuídos à conjunto de dados.
    - Quanto mais alto é um exemplo, maior é a importância dada pelo algoritmo.

Funcionamento do Algoritmo

- O Boosting começa com um peso P = 1 para todos exemplos base.
  - A partir do conjunto inicial é gerada a primeira hipótese H1.
  - Ocorrerá então a primeira classificação. Haverá erros e acertos nessa classificação.
  - Os erros devem receber maiores pesos, para que a próxima hipótese se encarregue de melhorar a classificação. Os acertos devem receber menores pesos.
  - A partir desse novo conjunto, é gerada a hipótese H2, e assim por diante.

- O processo continua até que sejam geradas H hipóteses.

A idéia geral de boosting



O AdaBoost

- O AdaBoost é um algoritmo de aprendizado de máquina, criado por Yoav Freund e Robert Schapire, em 1996. É um algoritmo meta-heurístico. Pode ser utilizado com o intuito de aumentar a performance de outros algoritmos de aprendizagem. É uma variação do algoritmo de boosting.

- De forma bem simples, o algoritmo funciona da seguinte maneira: A cada iteração, há a adaptação, baseada nas classificações feitas anteriormente, o ajuste acontece em relação às instâncias atribuídas com viés negativo.

- O AdaBoost chama um “aprendiz fraco” em n iterações, para cada chamada a distribuição de pesos Dn é atualizada, indicando a importância do exemplo no conjunto de dados.

Dificuldades do AdaBoost

- O algoritmo é sensível a ruído dos dados e a outliers.
- Suscetível ao Overfitting, que é a perda da capacidade de generalização, após o aprendizado de muitos padrões de treino.

Vantagens do AdaBoost

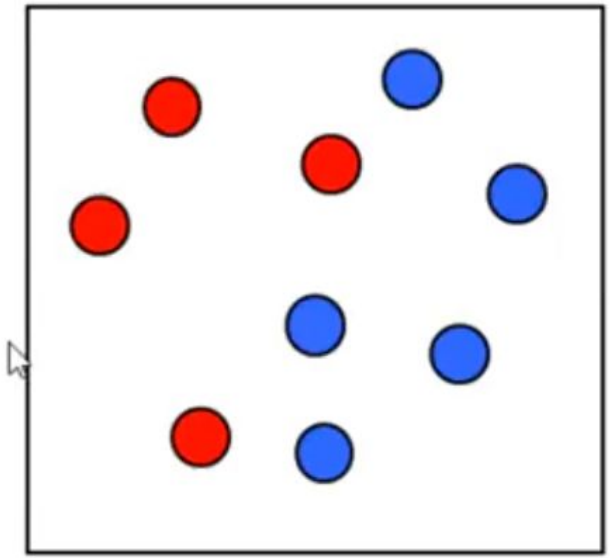
- Flexibilidade
- Facilidade na implementação, em diversas áreas
- Em relação a maioria dos outros algoritmos, o AdaBoost é menos suscetível ao overfitting.

Caso de Estudo

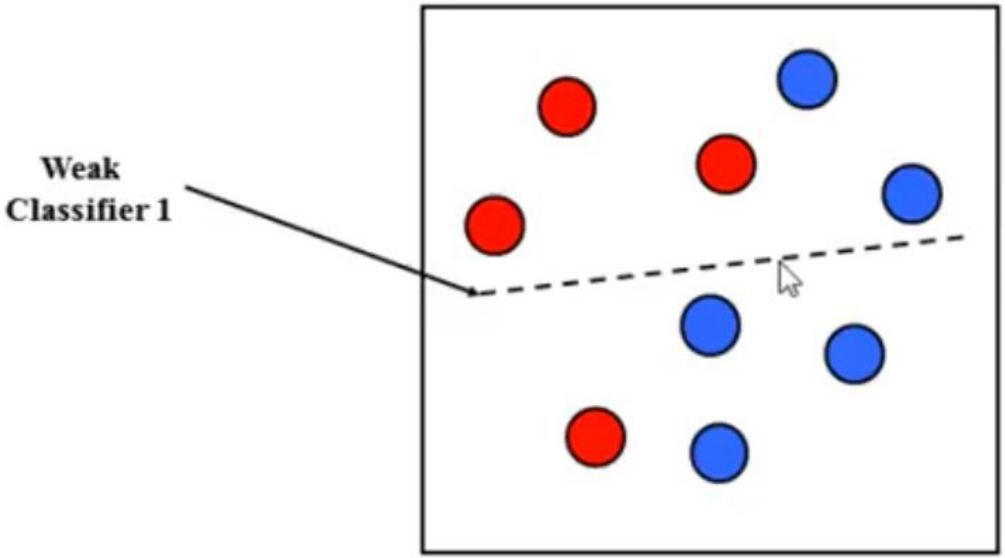
- Dissertação de Mestrado de Bruno Butilhão Chaves, aluno da USP.

- “O objetivo desta dissertação é estudar e desenvolver o conhecimento do algoritmo AdaBoost para aplicação em sensores, de forma a aprimorar a sensibilidade e precisão das medições, tanto de sensores isolados como de sistemas complexos com vários sensores, sem que seja necessário realizar modificações no próprio sensor. Para demonstrar a utilidade da técnica, foi realizado um estudo de caso utilizando um sistema composto de sensores capacitivos inter digitalizados e micro fabricados, sensores de temperatura e sensor a fibra óptica, para verificar adulterações em combustíveis automotivos, em especial, do etanol combustível. Sete experimentos são apresentados no trabalho. Índices acima de 90% de classificações corretas foram obtidos, indicando a viabilidade da utilização do algoritmo para calibração de sensores ou rede de sensores”.

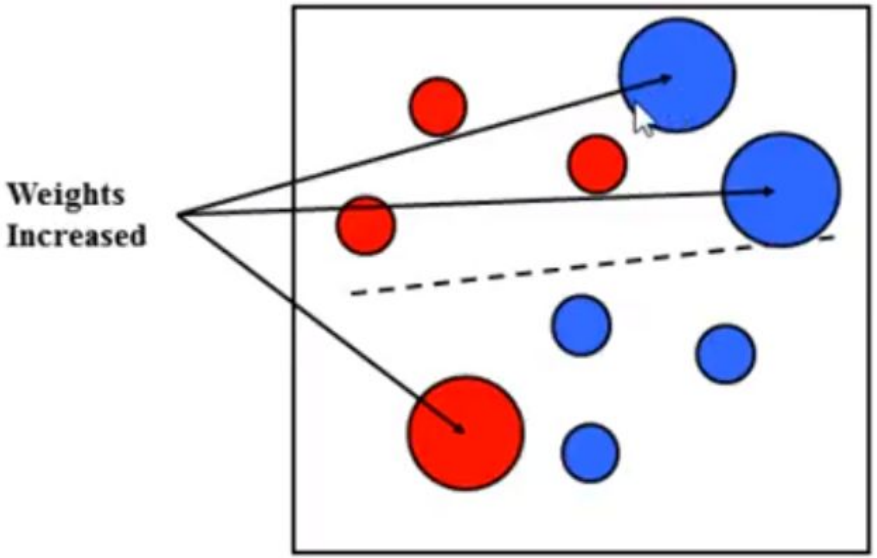
Boosting illustration (perceptron as weak learner)



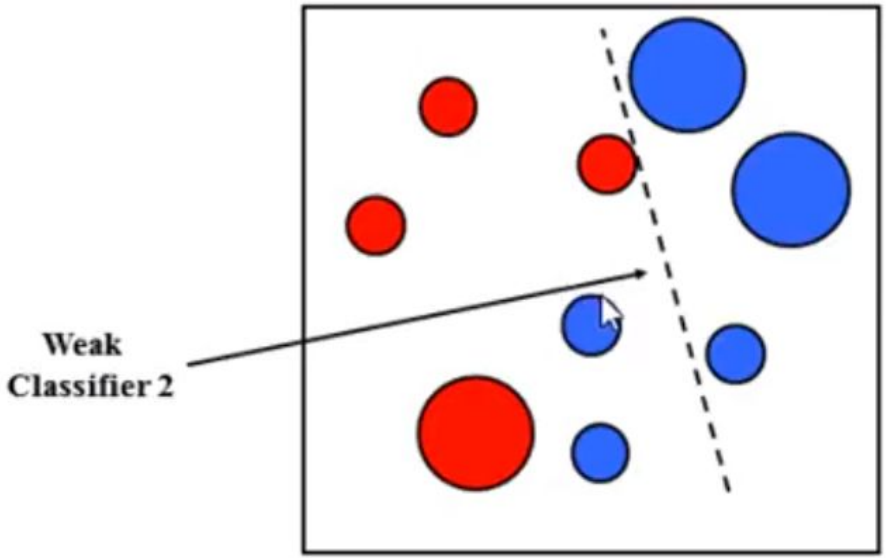
Boosting illustration (perceptron as weak learner)



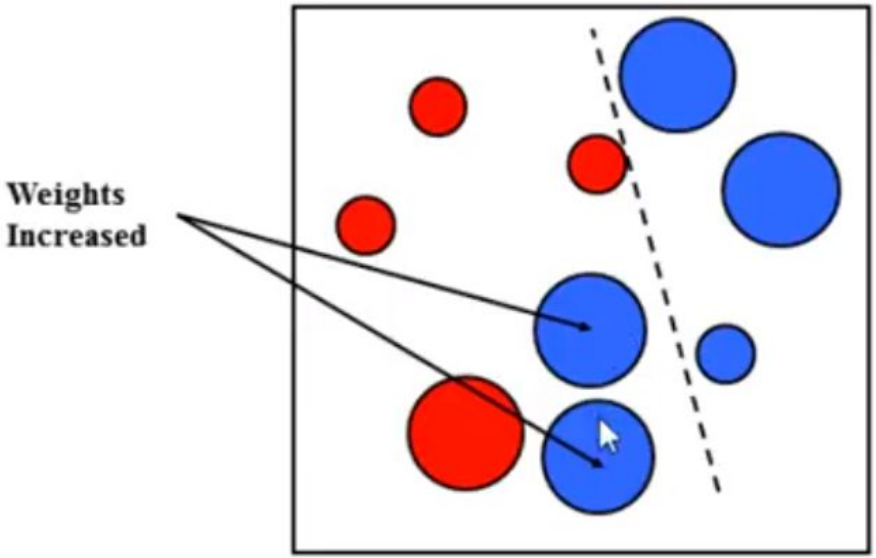
Boosting illustration



Boosting illustration

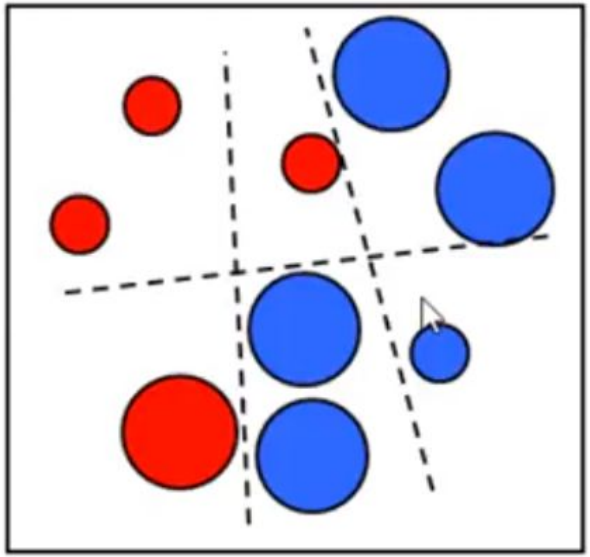


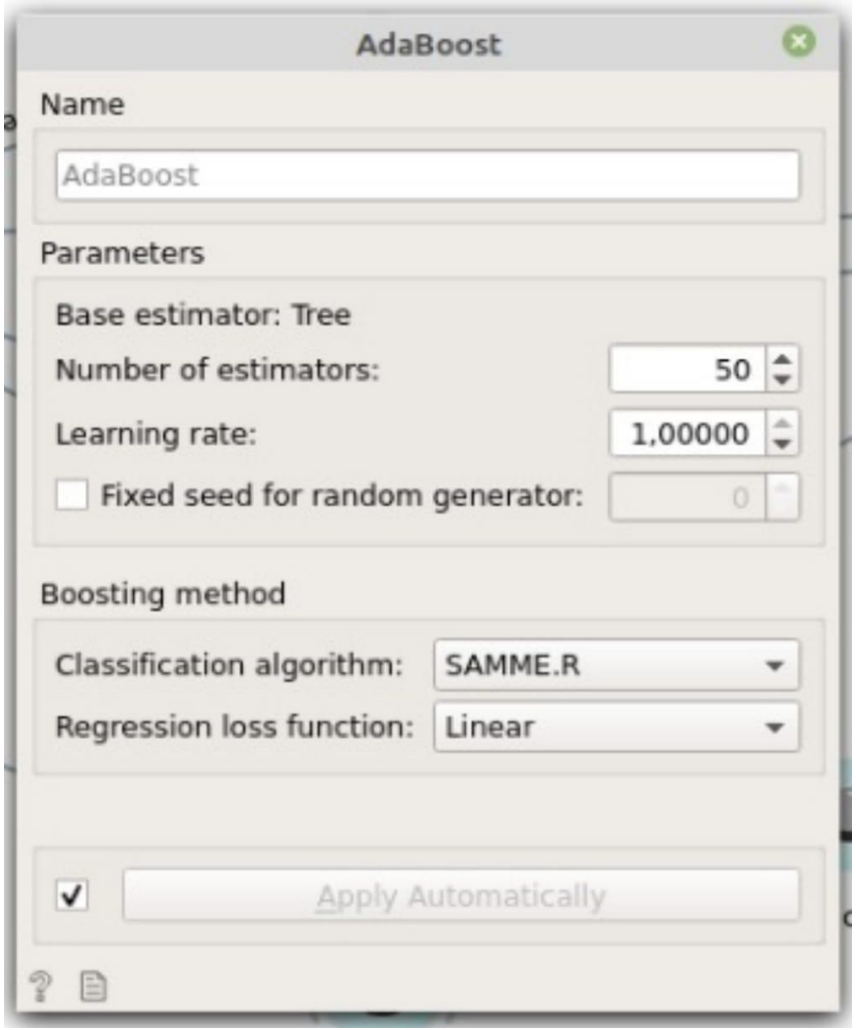
Boosting illustration



Boosting illustration

Final classifier is a combination of weak classifiers




AdaBoost
Name
AdaBoost
Parameters
Base estimator: Tree
Number of estimators:
50
Learning rate:
1,00000
Fixed seed for random generator:
0
Boosting method
Classification algorithm:
SAMME.R
Regression loss function:
Linear
Apply Automatically

Prática no Orange

- O algoritmo de AdaBoost é utilizado no Orange através do Widget AdaBoost.
- Os parâmetros do mesmo são:
  - Estimador Base
  - Número de Estimadores
  - Taxa de Aprendizado (Entre 0 e 1)
  - Semente Fixa para Gerador Aleatório
- No método Boosting, há a seguinte subdivisão:
  - Algoritmo de Classificação
    - SAMME
    - SAMME.R
- Função da Regressão (Se houver Regressão) :
  - Linear
  - Quadrada
  - Exponencial

Exemplo no Orange

- Exemplo do Dataset Irís, já utilizado no grupo de estudos.
  - 4 Feactures:
    - Comprimento da sépala em cm
    - Largura da sépala em cm
    - Comprimento da pétala em cm
    - Largura da pétala em cm

- 3 Classes possíveis:
  - Iris Setosa
  - Íris Versicolour
  - Iris Virginica

- As imagens a seguir exemplificam todo o fluxo seguido no Orange, e exibem o Test and Score realizado dos 3 algoritmos utilizados: Tree, Logistic Regression e AdaBoost

File Edit View Widget Options Help

Visualize

Model

Constant | CN2 Rule Induction | Calibrated Learner | kNN

Tree | Random Forest | SVM | Linear Regress...

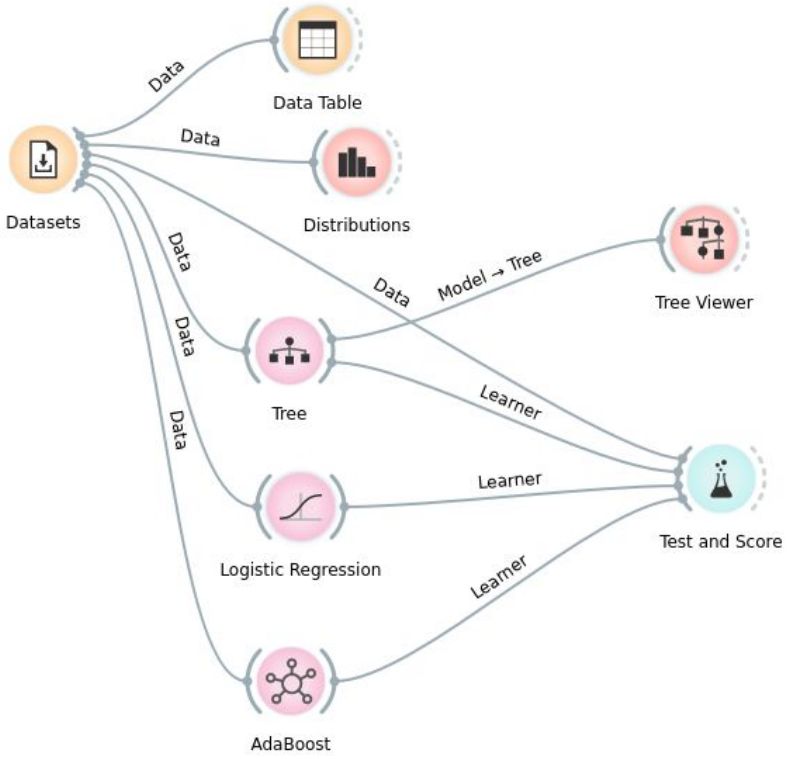
Logistic Regress... | Naive Bayes | AdaBoost | Neural Network

Stochastic Gradien... | Stacking | Save Model | Load Model

Evaluate

Select a widget to show its description.

See workflow examples, YouTube tutorials, or open the welcome screen.



16:39

Sampling

- Cross validation
  - Number of folds: 10
  - Stratified
- Cross validation by feature
- Random sampling
  - Repeat train/test: 10
  - Training set size: 66 %
  - Stratified
- Leave one out
- Test on train data
- Test on test data

Target Class

(Average over classes)

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Tree	0.959	0.935	0.936	0.936	0.935
Logistic Regression	0.985	0.943	0.943	0.948	0.943
AdaBoost	0.962	0.949	0.949	0.950	0.949