DATA SCIENCE TECH TALK

Diego Rodrigues

AGENDA

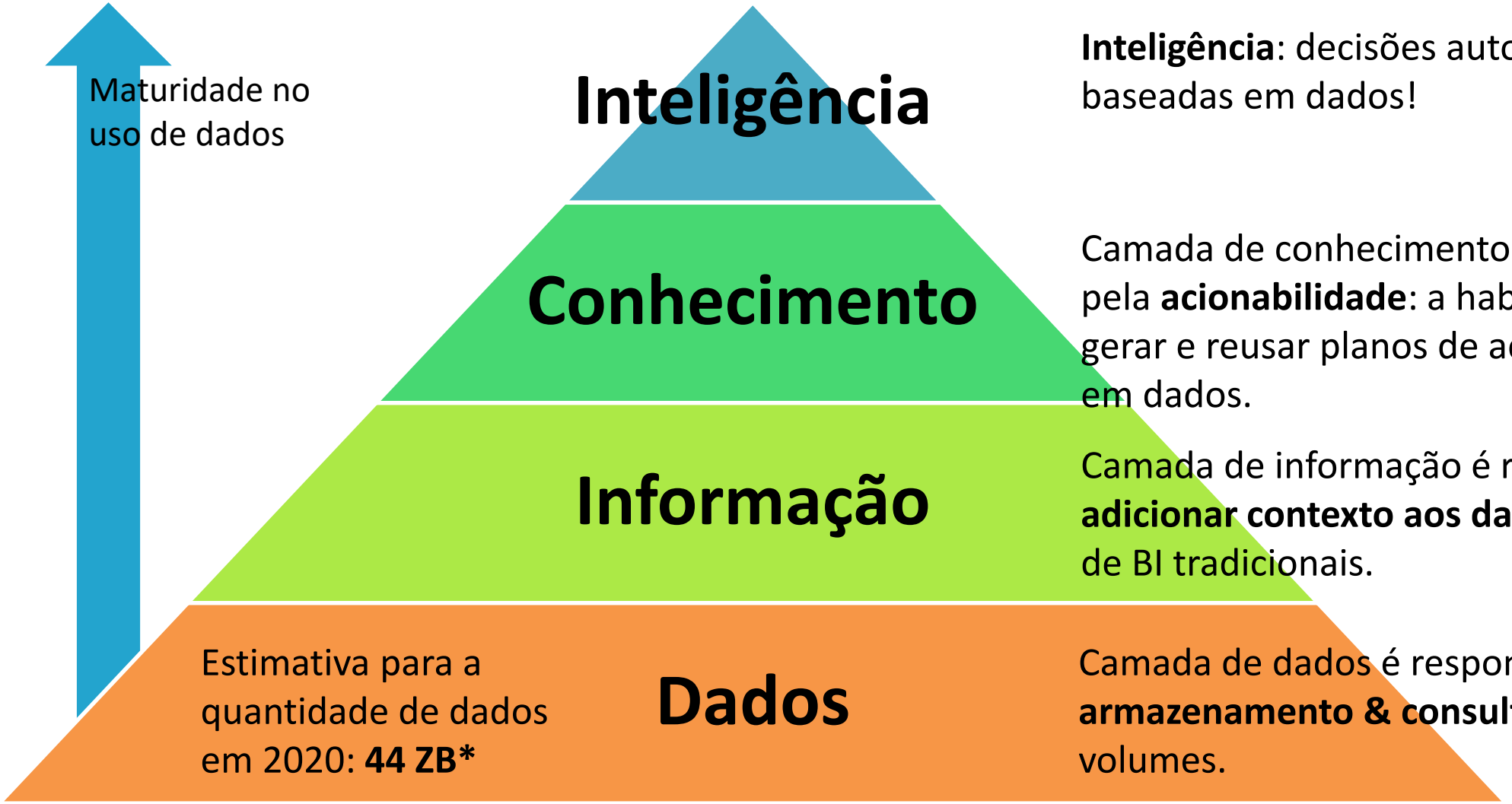
Introdução

Modelos de Machine Learning

Ciclo de Vida de Desenvolvimento

INTRODUÇÃO

Big Data, Business Intelligence, Analytics, Data Science...



Inteligência: decisões automáticas baseadas em dados!

Camada de conhecimento é responsável pela **acionabilidade**: a habilidade de gerar e reusar planos de ação baseados em dados.

Camada de informação é responsável por **adicionar contexto aos dados**: sistemas de BI tradicionais.

Camada de dados é responsável pelo **armazenamento & consulta** em grandes volumes.

*ZB = 10^21 Bytes

CICLO DE DESENVOLVIMENTO DE SOLUÇÕES DE “DATA MINING”

Cross Industry Standard Process for Data Mining - IBM

1) Requerimentos e Análise de Negócio

Entendimento do problema decisório, dados relacionados & revisão bibliográfica.

2) Preparação dos Dados

Cargas de ETL & consolidação de diferentes fontes.

3) Modelagem

Análise Exploratória, Seleção de atributos e treinamento.

4) Avaliação

Seleção do melhor modelo.

5) Liberação

Liberação do modelo no ambiente de produção.

Data Scientist Core Skills

Apresentação

- Relatórios
- Visualização
- Diversão

Programação

- Programação procedural, OO, funcional...
- Algoritmos & Estruturas de Dados
- Linguagens de Acesso à Dados

Matemática

- Cálculo
- Probabilidade
- Álgebra Linear
- Otimização

MODELOS DE MACHINE LEARNING

Que tipo de problema pode ser resolvido com modelos de machine learning?

1) Classificação

Um bebê consegue separar e ordenar blocos com diferentes tamanhos, formas e cores. Ele também consegue **identificar os tipos diferentes de objetos**.

Data Scientist da Nova Geração

Os diferentes tipos de objetos são chamados de **classes**. As características dos objetos são chamadas de **variáveis** ou **atributos**.

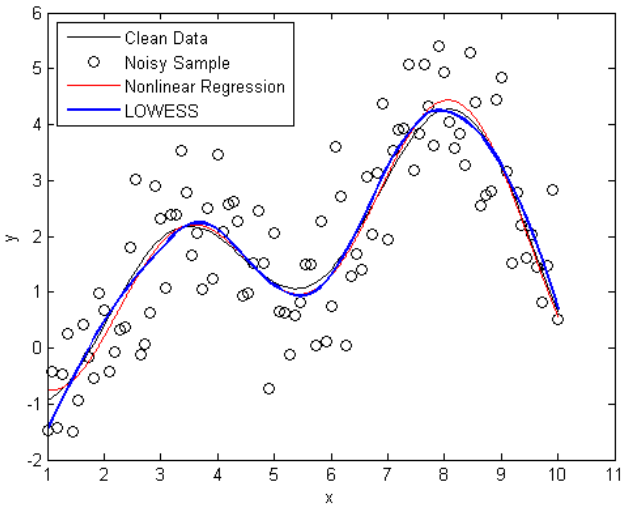
Então, um classificador é um modelo **treinado para discriminar objetos** pertencentes a duas ou mais classes, baseado em seus atributos.

Que tipo de problema pode ser resolvido com modelos de machine learning?

2) Regressão

O objetivo da regressão é **modelar as relações funcionais** entre dois conjuntos de variáveis.

As variáveis que representam as causas são chamadas de **variáveis independentes**, e as variáveis cujo objetivo é prever, são chamadas **variáveis dependentes**.



As vezes quando o mundo não é linear & gaussiano...

Então, uma **regressão** é um modelo utilizado para prever **uma ou mais variáveis dependentes**, baseado em causas, ou variáveis independentes.

Que tipo de problema pode ser resolvido com modelos de machine learning?

3) Agrupamento (Clustering)

Um bebê consegue **agrupar objetos por cor, tamanho, formato** e muitos outros atributos que ele pode observar nos objetos.

Diferentes maneiras de organizar os objetos são diferentes **estruturas de agrupamentos** existentes em uma amostra de dados.

De quantas maneiras estes blocos podem ser organizados em grupos?

Um **modelo de agrupamento** é usado para **identificar grupos**, ou estruturas de agrupamentos, nos dados.

Que tipo de problema pode ser resolvido com modelos de machine learning?

4) Modelagem Probabilística

Em uma cesta de supermercado existe uma variedade grande de itens comprados juntos. A presença de um item específico afeta a **chance de outro produto estar presente na cesta?**

Será que algum desses itens são comprados juntos?

O objetivo de modelos probabilísticos é **identificar padrões** dentro de comportamentos supostamente **aleatórios**.

CICLO DE VIDA DE DESENVOLVIMENTO

COMO UM SISTEMA AUTOMÁTICO PODE TOMAR DECISÕES?

Seleção de Atributos

- O problema de decisão deve ser levado do mundo “físico” para o mundo dos “dados”, através da **mensuração dos atributos relevantes** ao processo decisório.
- **Análise exploratória** provê feedback para o Data Scientist sobre a qualidade, relevância e relacionamento entre os atributos.

Treinamento

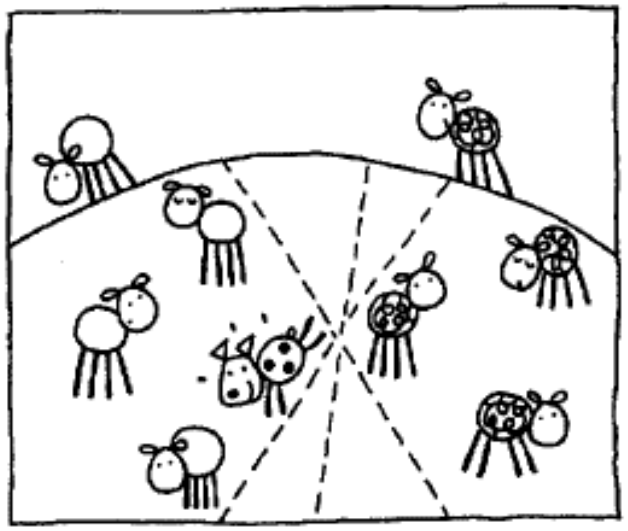
- Baseado em características específicas do problema decisório, um **modelo apropriado** deve ser utilizado para se ajustar aos dados.
- A preocupação mais importante desta etapa é relacionada a **capacidade de generalização** do modelo: o poder de operar além da amostra de treinamento.

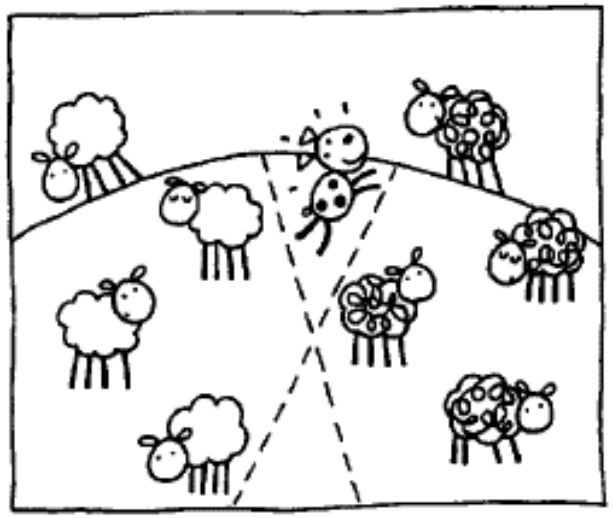
Avaliação

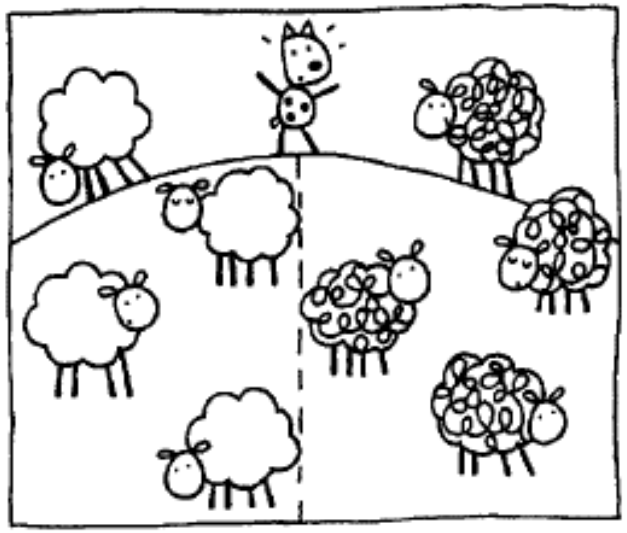
- Normalmente diversos modelos são testados contra os dados.
- Esses modelos devem ser **comparados para decidir qual é o melhor**.
- Após o melhor modelo ser selecionado, a teoria da decisão deve ser utilizada para definir como o sistema irá **selecionar ações** baseado na resposta do modelo.

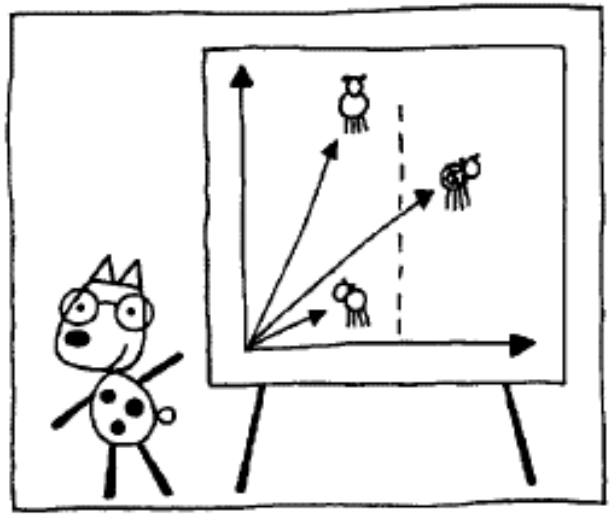
SELEÇÃO DE ATRIBUTOS

Representação: como atribuir números às características de uma ovelha?









Exercício (1): qual representação o cachorro deve escolher para diferenciar ovelhas pretas de brancas?

Exercício (2): qual seria uma boa representação para diferenciar ~~homens e mulheres~~ ratos e elefantes?

Exemplo: espaço de atributos para o dataset Iris

Iris Setosa

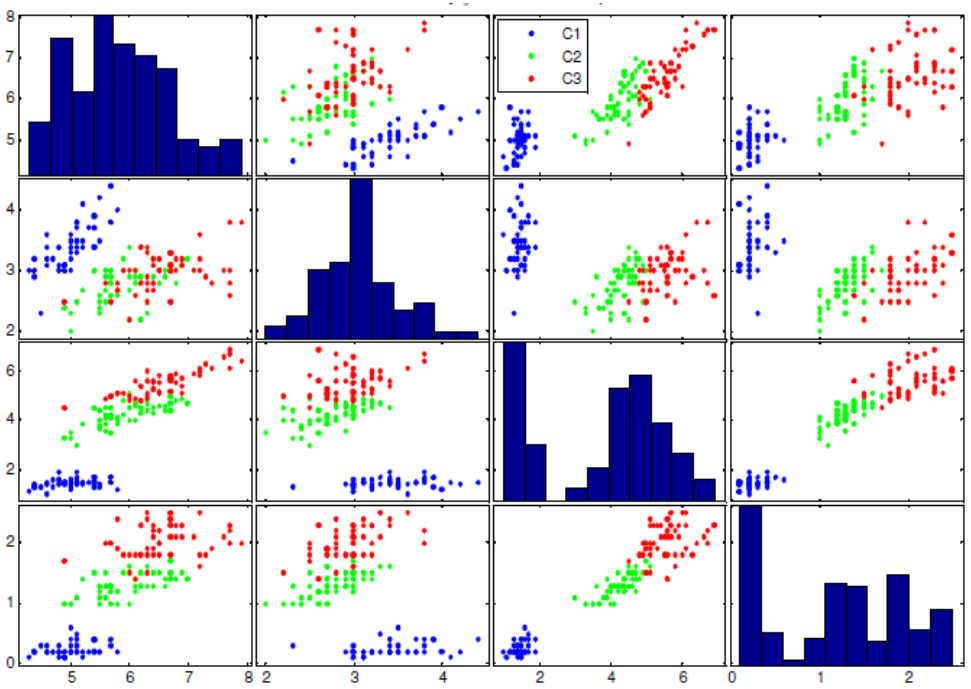
Iris Versicolor

Iris Virginica

Características das flores

Largura & comprimento da pétala

Largura & comprimento da sépala



Espaço de Atributos com **4 dimensões!**

http://archive.ics.uci.edu/ml/datasets/Iris

QUAIS SÃO OS TIPOS MAIS COMUNS DE ATRIBUTOS?

Nominal ou Categórica

- Conjunto de diferentes valores não ordenados.
- Exemplo: Sexo, cor, palavras, tipo de coisas.

Ordinal

- Conjunto ordenado, mas a diferença entre os valores não tem significado.
- Exemplo: scores quantitativos como “excelente”, “bom”, “regular”, “ruim”.

Intervalo

- Conjunto ordenado, a diferença tem significado mas não as proporções.
- Exemplo: Datas.

Ratio

- Conjunto ordenado onde diferenças & proporções tem significado.
- Exemplo: Idade, peso, altura, dinheiro, massa, etc.

PORQUE É NECESSÁRIO MANTER O NÚMERO DE ATRIBUTOS O MENOR POSSÍVEL?

1) “Maldição” da Dimensionalidade

Suponha que 10.000 observações são distribuídas aleatoriamente no intervalo [0, 1]. Qual é a distância média entre os pontos? E se as observações são distribuídas no cubo $[0, 1]^3$? Ou em um hipercubo de 100 dimensões?

Esparsidade do espaço de atributos aumenta com o **número de dimensões!**

2) Multi-colinearidade

Dois atributos com uma relação significativa pode sugerir causalidade entre ambos ou relação com uma variável latente desconhecida. De uma maneira ou outra, o atributo “independente” vai ter mais importância para o modelo, já que está representado por mais de um atributo.

TÉCNICAS DE SELEÇÃO DE ATRIBUTOS

Filtragem – mede a relação entre atributos ou atributos e classes, utilizando estatísticas, sem depender do modelo.

- **Coeficiente de Correlação de Pearson** – Estatística que mede a relação linear entre duas variáveis aleatórias.
- **Teste T de diferença de médias** – Informa se a média de um determinado atributo muda de acordo com a classe.
- **ANOVA** – O mesmo que o teste T, mas serve para múltiplas classes.
- **Informação Mútua** – Estatística que mede relação não-linear entre duas variáveis aleatórias.

Wrapper – mede a relação entre atributos e classes, utilizando um modelo treinado.

- **Gini –** Estatística que representa a importância de um atributo na divisão da base de dados por uma árvore de decisão.
- **Delta Sp** – Estatística que representa a variação causada na saída do modelo quando um atributo é substituído por sua média.

TÉCNICAS DE EXTRAÇÃO DE ATRIBUTOS

Transformação do espaço de atributos ou criação de novos atributos baseados nos existentes.

- **Análise de Componentes Principais (PCA)** – transforma o espaço de atributos. As novas dimensões são ordenadas por quanto representam da variância da amostra.
- **Kernel PCA** – PCA não-linear, utilizando um kernel não-linear antes de “PCAr” os dados.
- **Fatoração de Matrizes Não-Negativas (NMF)** – método de decomposição de matrizes cujo objetivo é o mesmo das técnicas acima.

TREINAMENTO

PARADIGMAS DE APRENDIZADO

1) Aprendizado Supervisionado

Tarefas de classificação e regressão pertencem a esta categoria. O treinamento consiste em **encontrar parâmetros** para o modelo que **minimiza uma função de risco/erro** para uma amostra de treinamento, baseado na diferença entre os **valores previstos e reais**, para cada observação.

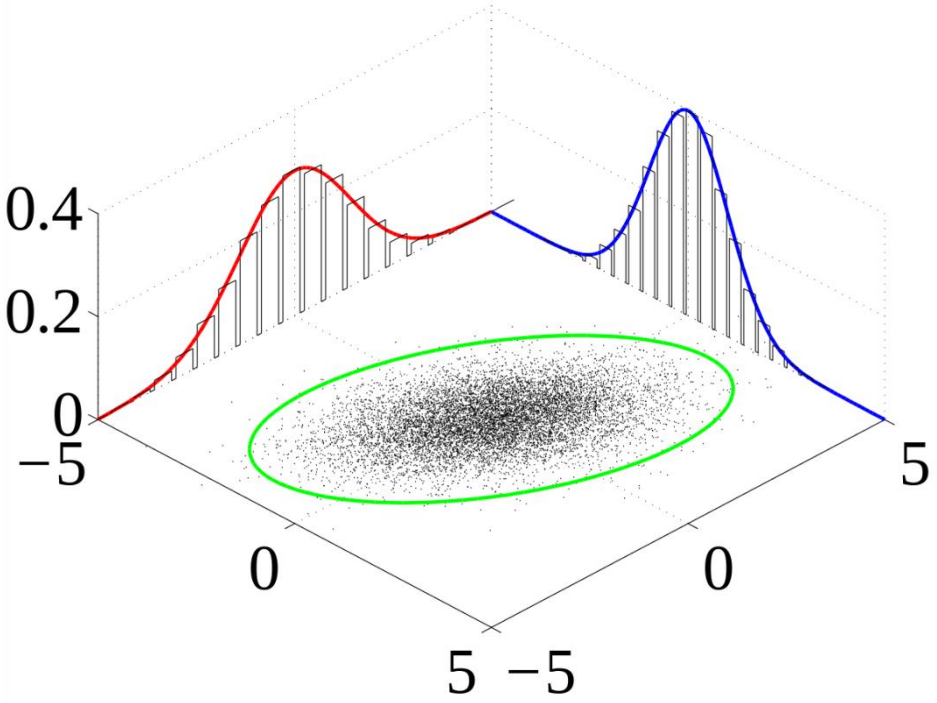
2) Aprendizado Não-Supervisionado

Agrupamento e modelagem probabilística pertencem a esta categoria. Não existe um **conhecimento "a priori" dos grupos** contidos nos dados. Algoritmos de agrupamento dependem fortemente de uma definição de "**distância**" ou "**similaridade**" entre as observações.

CLASSIFICAÇÃO: (1) ALGORITMOS BASEADOS EM DENSIDADE

Algoritmos que dependem da **função densidade de probabilidade** dos dados, ou aproximações locais, para determinar a classe de observações fora da amostra de treino.

1) Classificador Bayesiano
2) Classificador Bayesiano "Naïve"
3) K-Vizinhos mais próximos

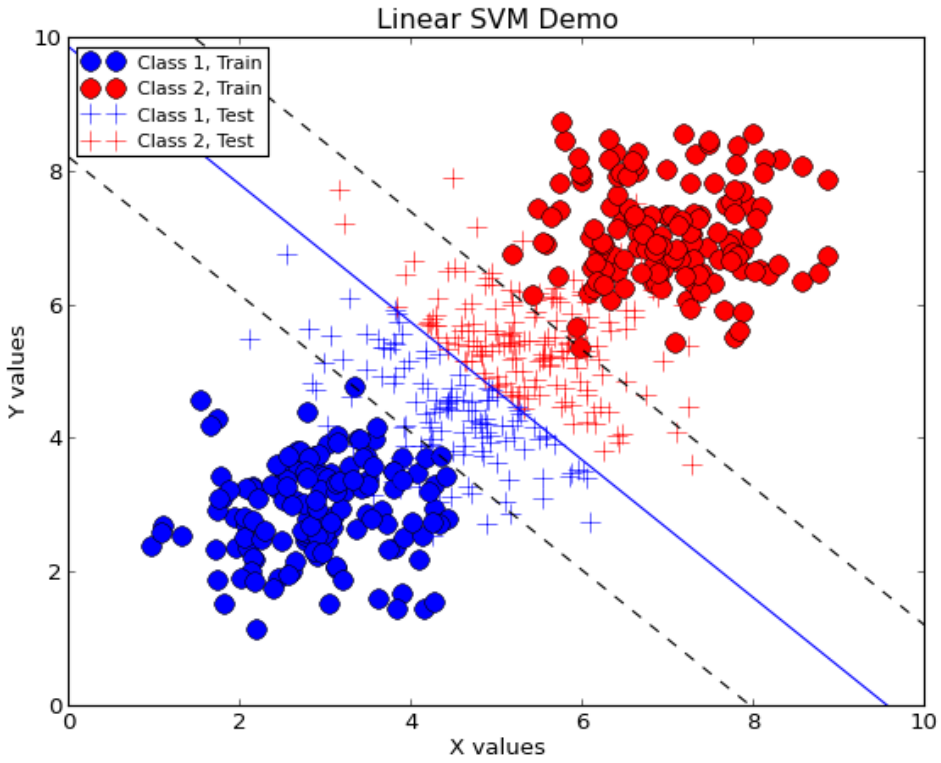


Algoritmos baseados em densidade dependem da **DENSIDADE (!!)**. Consequentemente, se beneficiam de um **conjunto grande de observações e de baixa esparsidade do espaço de atributos**. O Classificador Bayesiano é considerado o classificador "ótimo", mas é raramente utilizado, dada a dificuldade de estimar a função densidade de probabilidade dos dados. É normalmente utilizado como benchmark para comparação teórica entre os algoritmos de classificação.

CLASSIFICAÇÃO: (2) ALGORITMOS BASEADOS EM FUNÇÕES DISCRIMINANTES

Algoritmos que dependem da **estimação dos parâmetros de uma função** que é utilizada como **superfície de separação** entre as classes.

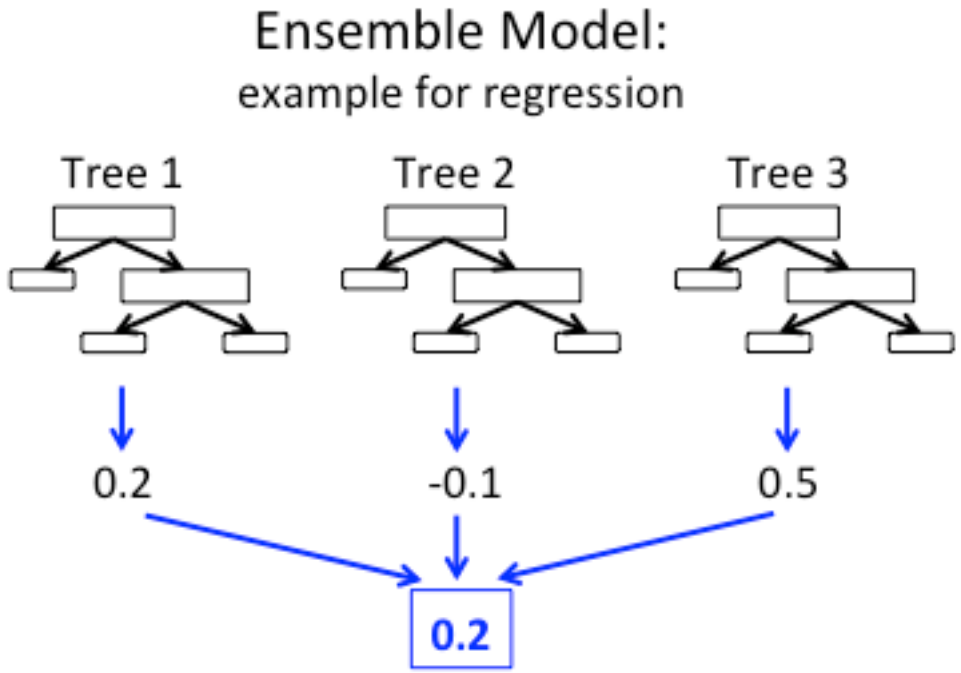
1) Funções Polinomiais
2) Regressão Logística
3) Máquina de Vetores Suporte
4) Redes Neurais
5) Árvores de Decisão



Algoritmos baseados em funções são **mais simples**, usualmente tem um **número menor de parâmetros** e não dependem em armazenar muitos dados para manter uma “memória”, como por exemplo K-vizinhos mais próximos.

CLASSIFICAÇÃO: (3) ALGORITMOS BASEADOS EM ENSEMBLE

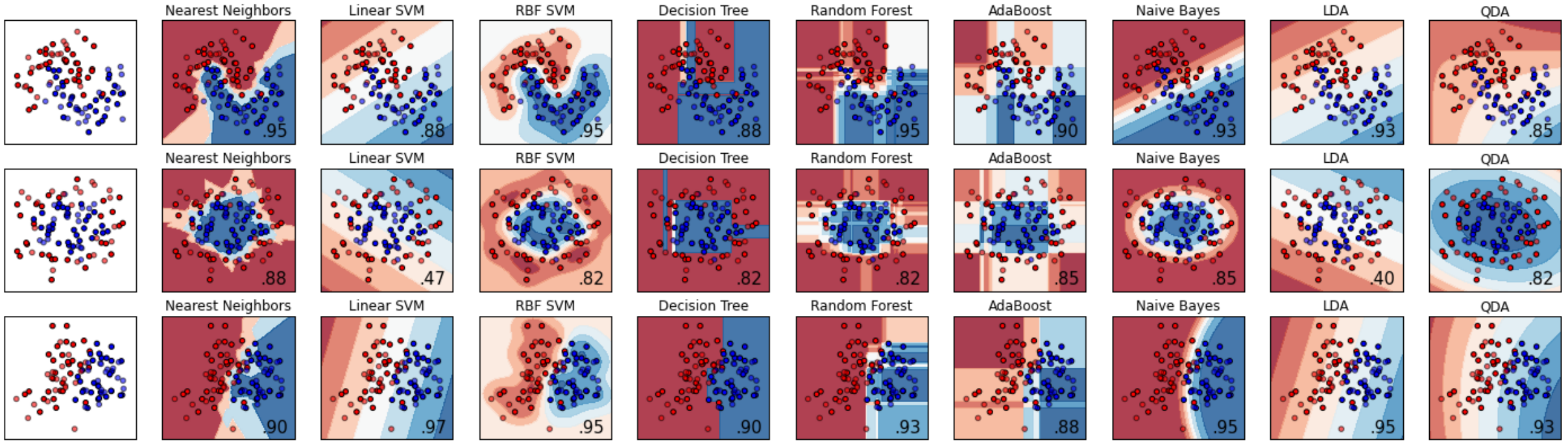
Algoritmos que **combinam modelos simples**, usualmente através de **votação ou ponderação**, para atingir maiores taxas de classificação.



1) Random Forest
2) Boosting

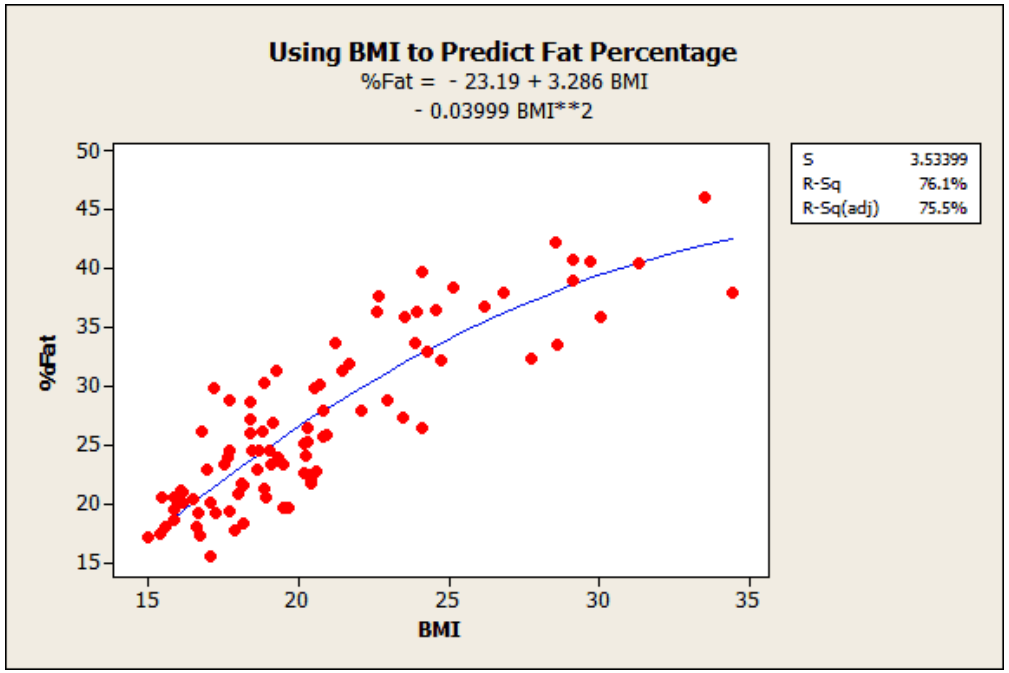
Boa **capacidade de generalização** gerado através de **arranjos complexos** de múltiplos modelos simples de machine learning.

COMPARAÇÃO DE ALGORITMOS DE CLASSIFICAÇÃO



REGRESSÃO: ALGORITMOS

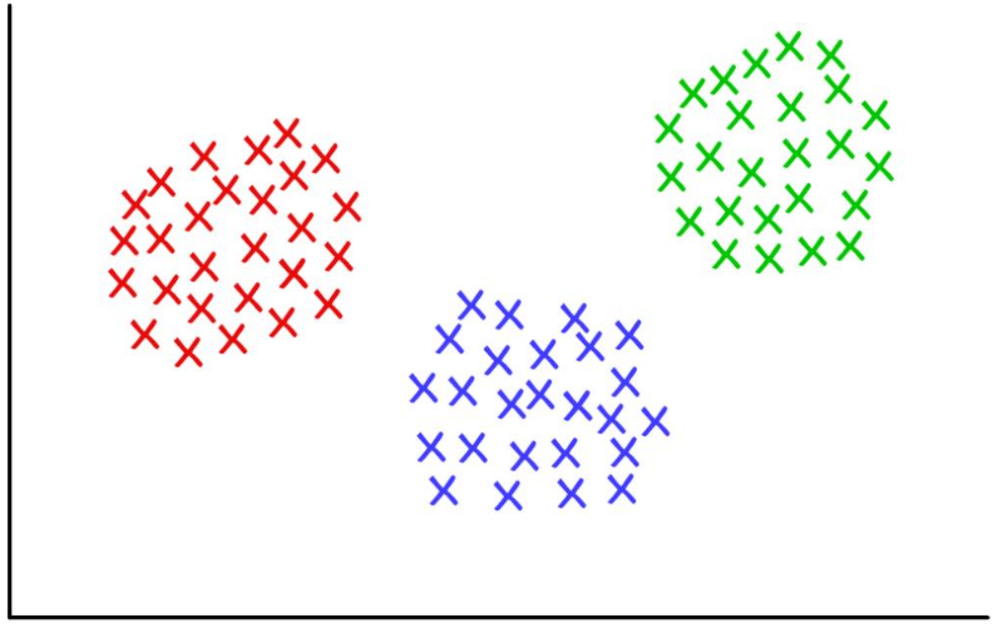
1) Modelos Lineares Generalizados
2) Regressão Não-Linear
3) Processos Gaussianos
4) Máquina de Vetores Suporte
5) Redes Neurais



Algoritmos de regressão geralmente são modelados combinando uma **parte determinística e uma parte aleatória**. Os parâmetros correspondente à parte determinística são encontrados utilizando estimadores como máxima verossimilhança ou máximo a posteriori (MAP).

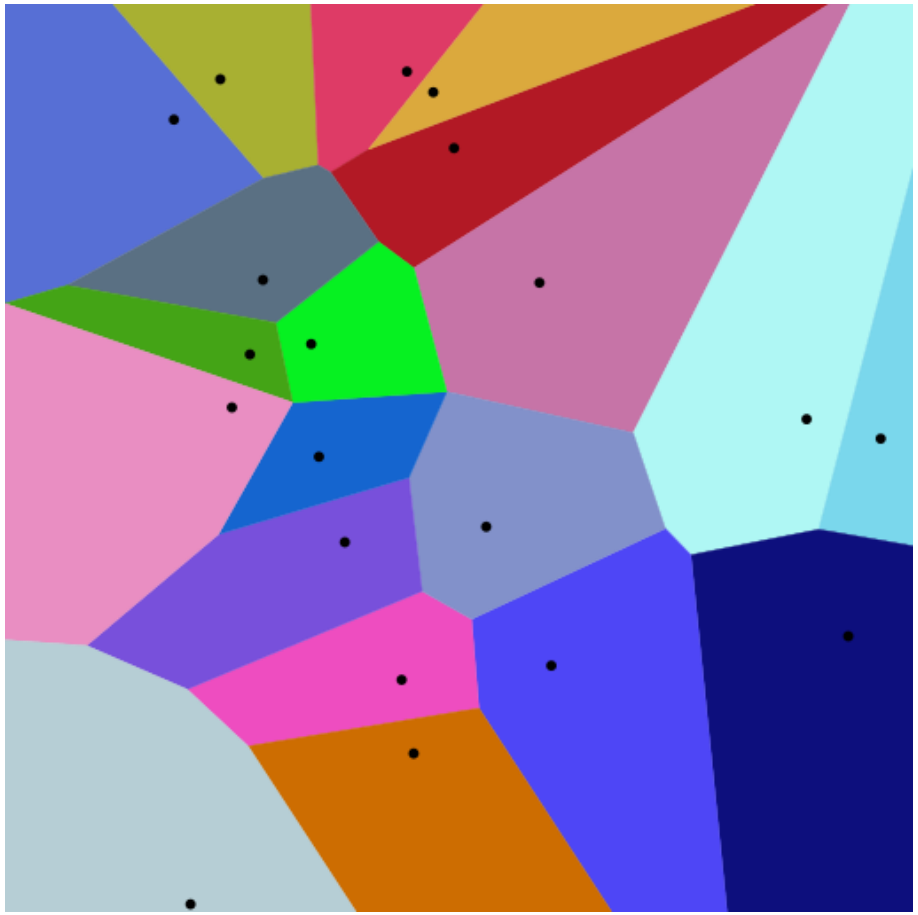
AGRUPAMENTO: ALGORITMOS

1) K-Means
2) Mapa Auto-Organizável
3) Hieráquico
4) DBSCAN



Além da escolha do algoritmo, os resultados do agrupamento dependem diretamente dos atributos e da **métrica escolhida para definir similaridade** entre os objetos.

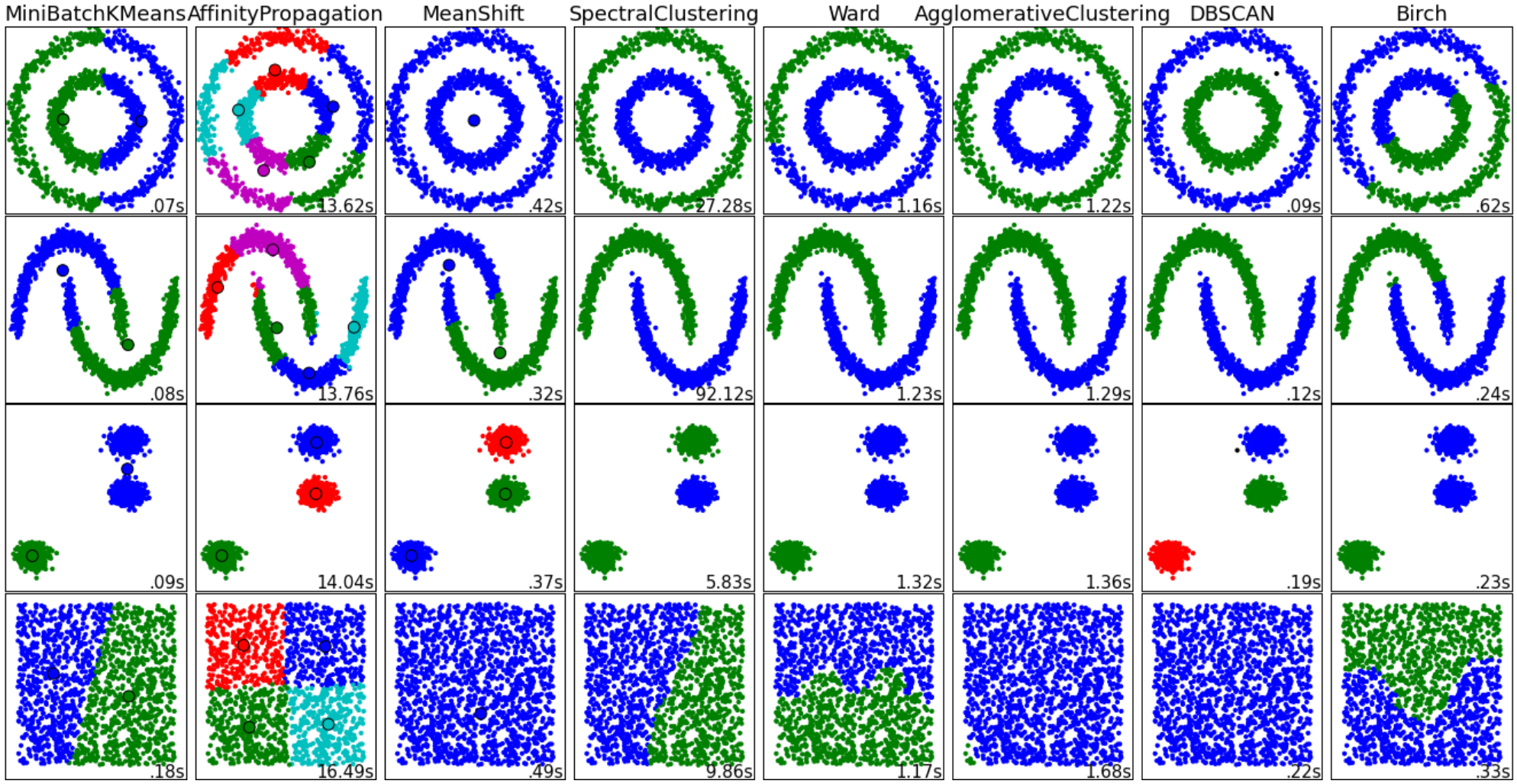
AGRUPAMENTO: DIAGRAMA DE VORONOI



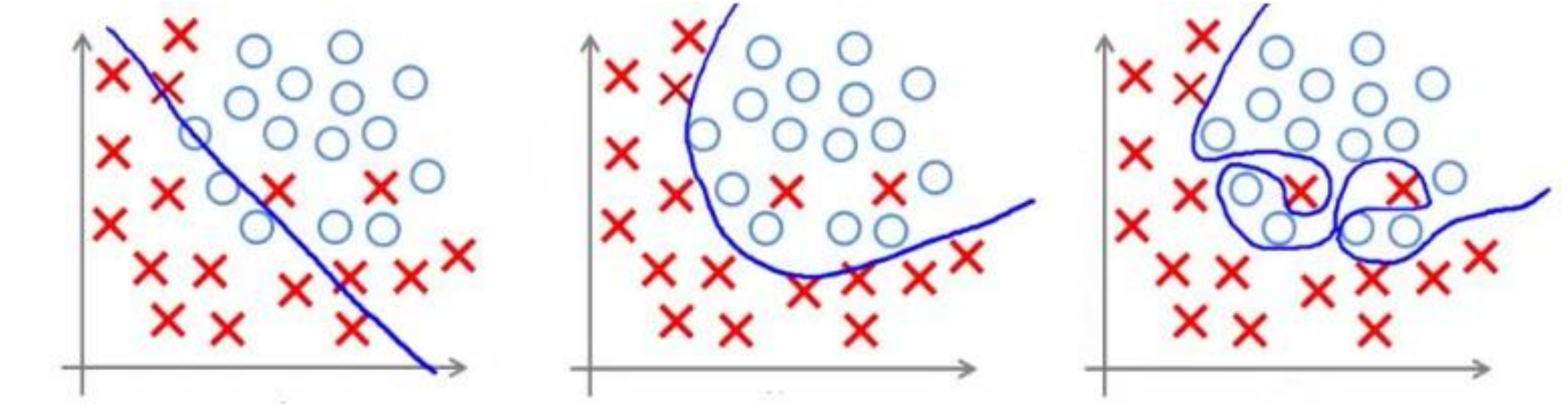
Distância Euclideana

Distância de Manhattan

AGRUPAMENTO: COMPARAÇÃO DE ALGORITMOS



COM ASSEGURAR GENERALIZAÇÃO?



Under-fitting

(too simple to explain the variance)

Appropriate-fitting

Over-fitting

(forcefitting -- too good to be true)

VALIDAÇÃO CRUZADA: EVITAR OVERFITTING AO ESCOLHER OS PARÂMETROS

LEAVE ONE OUT

- Uma única observação é deixada de fora a cada treinamento. N treinamentos são realizados para calcular a estatística de erro.

K FOLDS

- Amostra é dividida em K conjuntos. K treinamentos são realizados, mantendo um conjunto como fora-da-amostra.

BOOTSTRAPPING

- O algoritmo itera, amostrando aleatoriamente M observações, para a quantidade Q desejada de treinamentos.

FIGURAS DE MÉRITO PARA AVALIAR O RESULTADO

Verdadeiro Positivo (True Positive/TP)

- Classe positiva, classificação correta.

Verdadeiro Negativo (True Negative/TN)

- Classe negativa, classificação correta.

Falso Positivo (False Positive/FP)

- Negativo, classificado como positivo.

Falso Negativo(False Negative/FN)

- Positivo, classificado como negativo.

Matriz de Confusão para Classificação de Gatos

		Prediction	
		Cat	Dog
Actual	Cat	15 TP	35 FN
	Dog	40 FP	10 TN

Acurácia

- (TP+TN)/(P+N)

Taxa de Erro

- 1-Acurácia

Sensibilidade/Eficiência (aka recall)

- TP/(TP+FN)

Especificidade

- TN/(TN+FP)

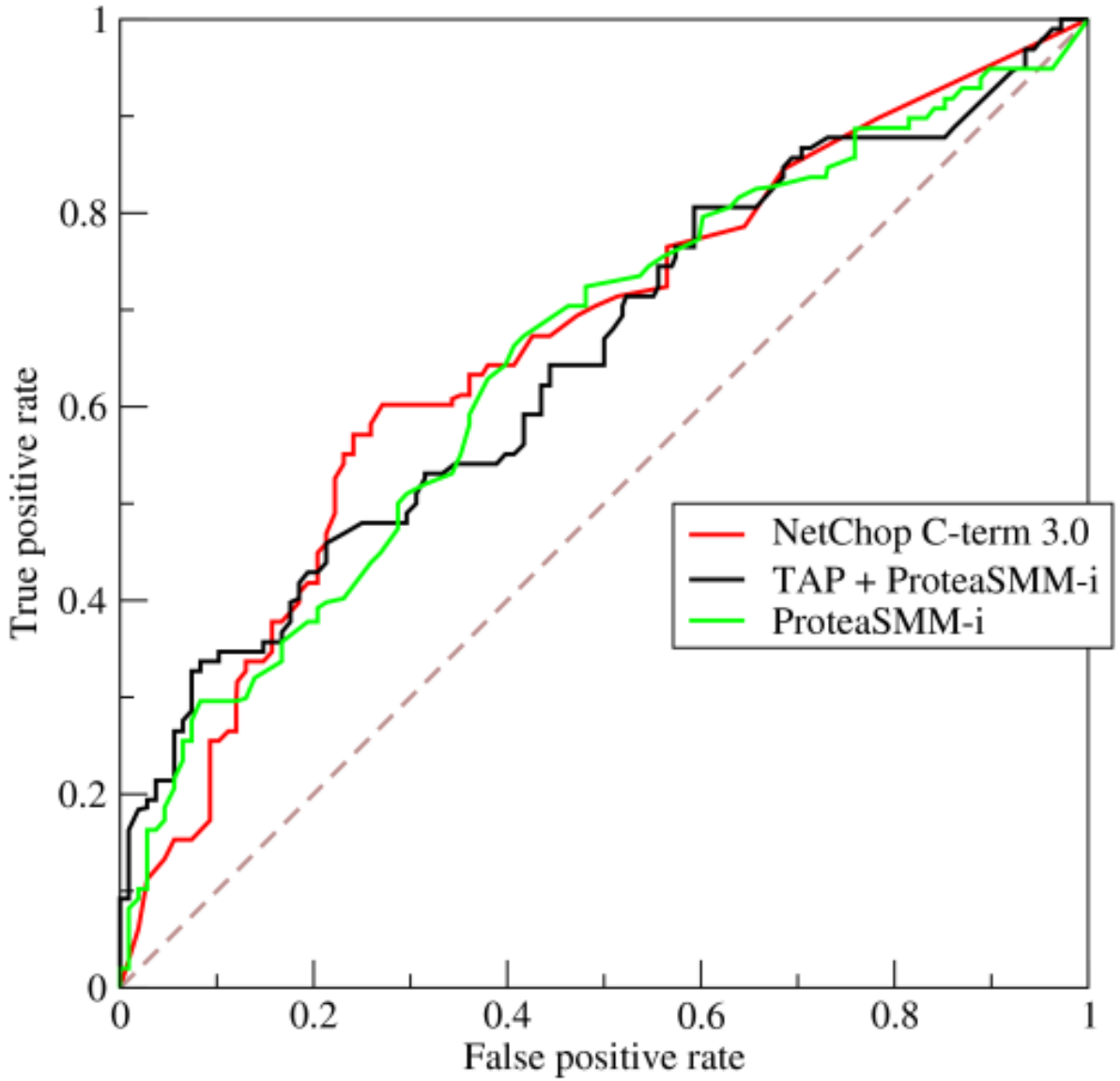
Precisão

- TP/(TP+FP)

Produto Sp

- SQRT[SQRT(R1*R2) * (R1 + R2)/2]

CALIBRANDO A SAÍDA DO MODELO – CURVA ROC



PERGUNTAS?