Classificação de Dados

Parte 3

Prof. Guilherme Palermo Coelho

Roteiro

- Ensembles:
  - Motivação e Definições;
  - A questão da Diversidade;
  - Etapas de Construção.

Motivação e Definições

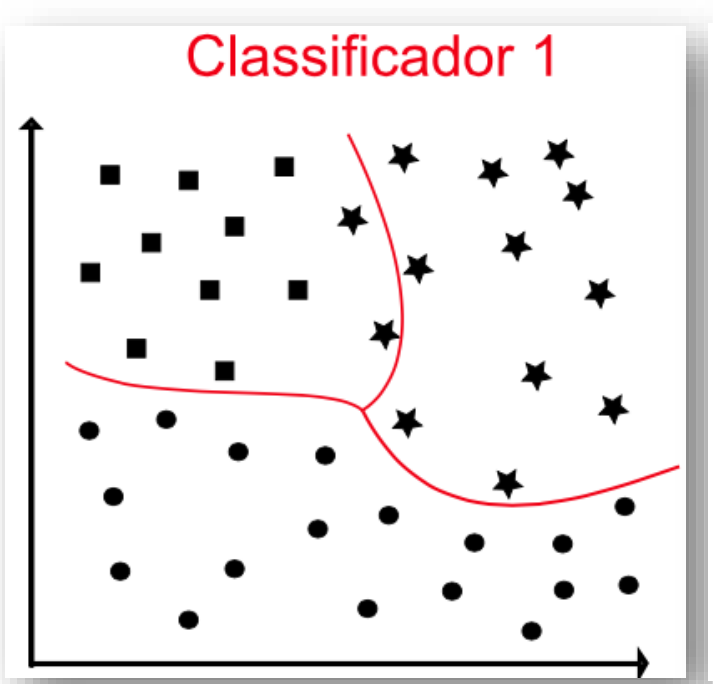
Motivação e Definições

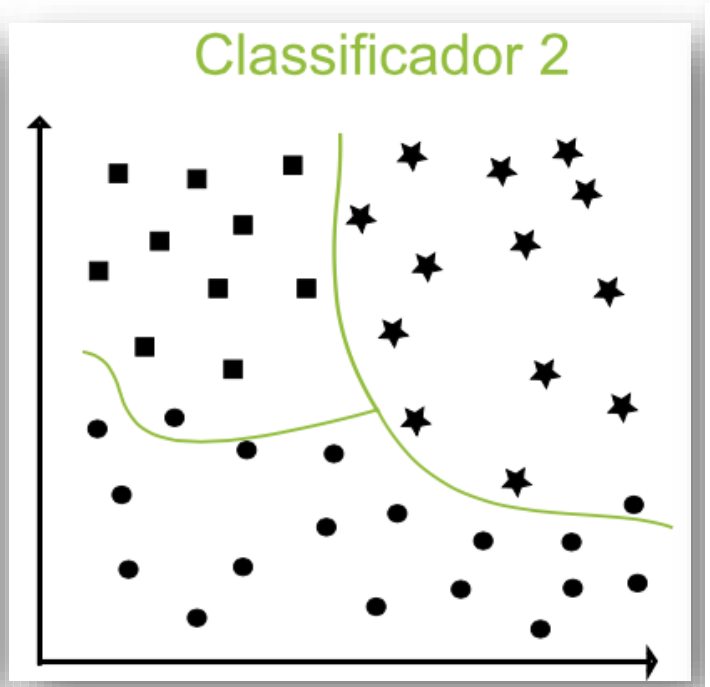
- Até o momento, vimos diferentes tipos de classificadores de dados;

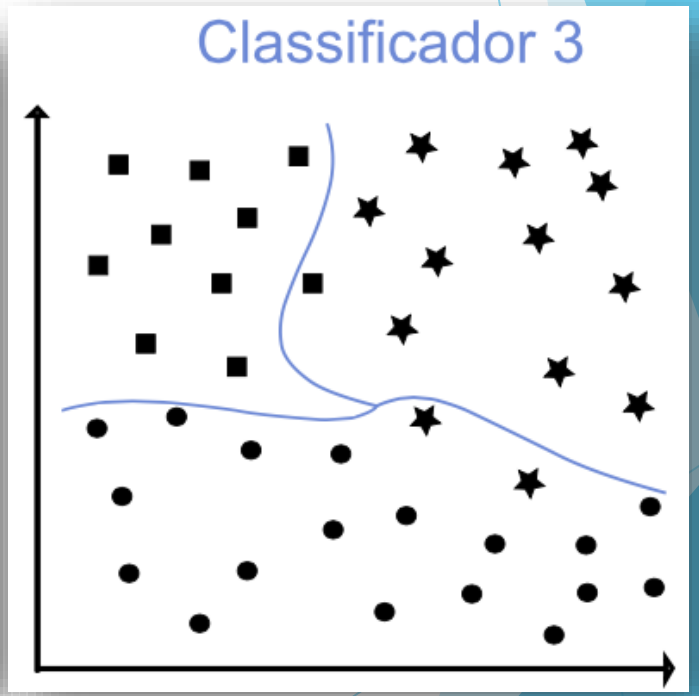
- Não é possível afirmar que um deles será **sempre o melhor** para todos os problemas de classificação:
  - Classificadores com estruturas diferentes;
  - Formas de construção do modelo diferentes;
  - Representação do modelo diferente;
  - Resultados com características diferentes.

Motivação e Definições

- Classificadores diferentes:

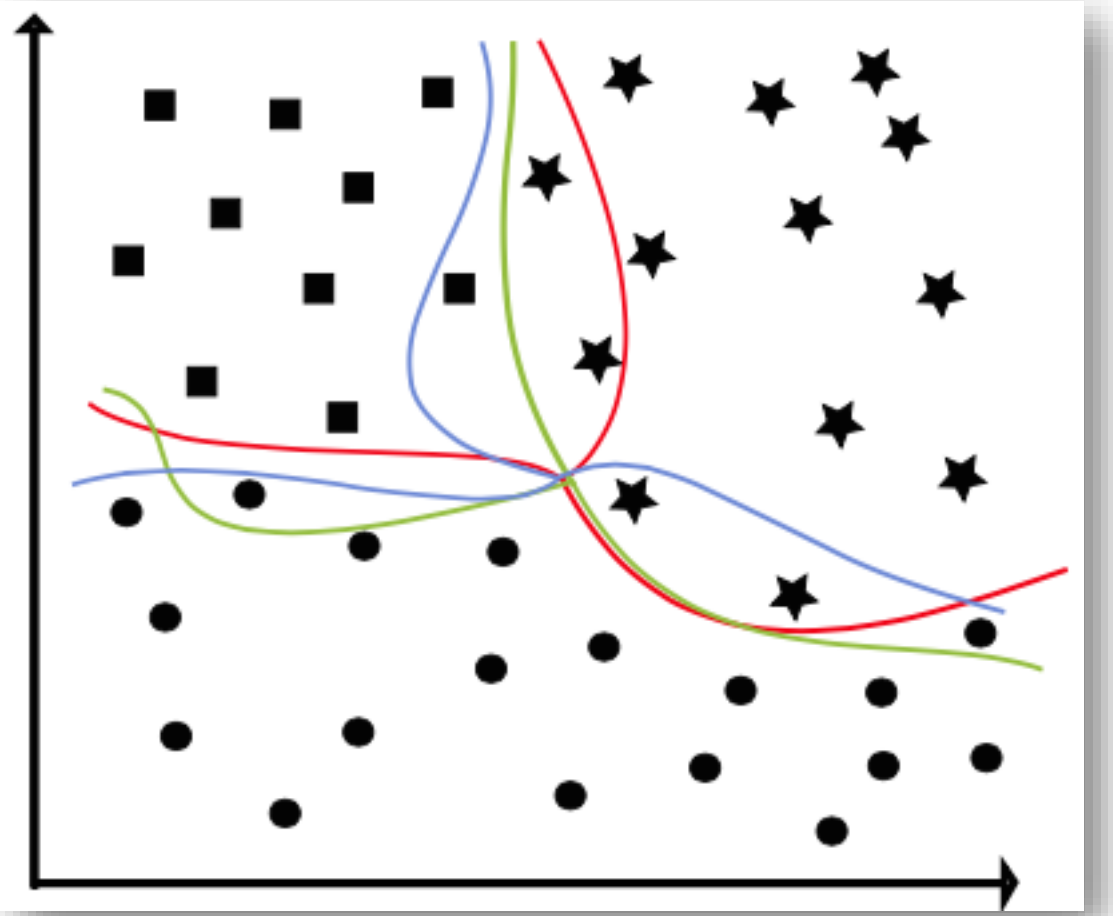






Motivação e Definições

- Classificadores diferentes:

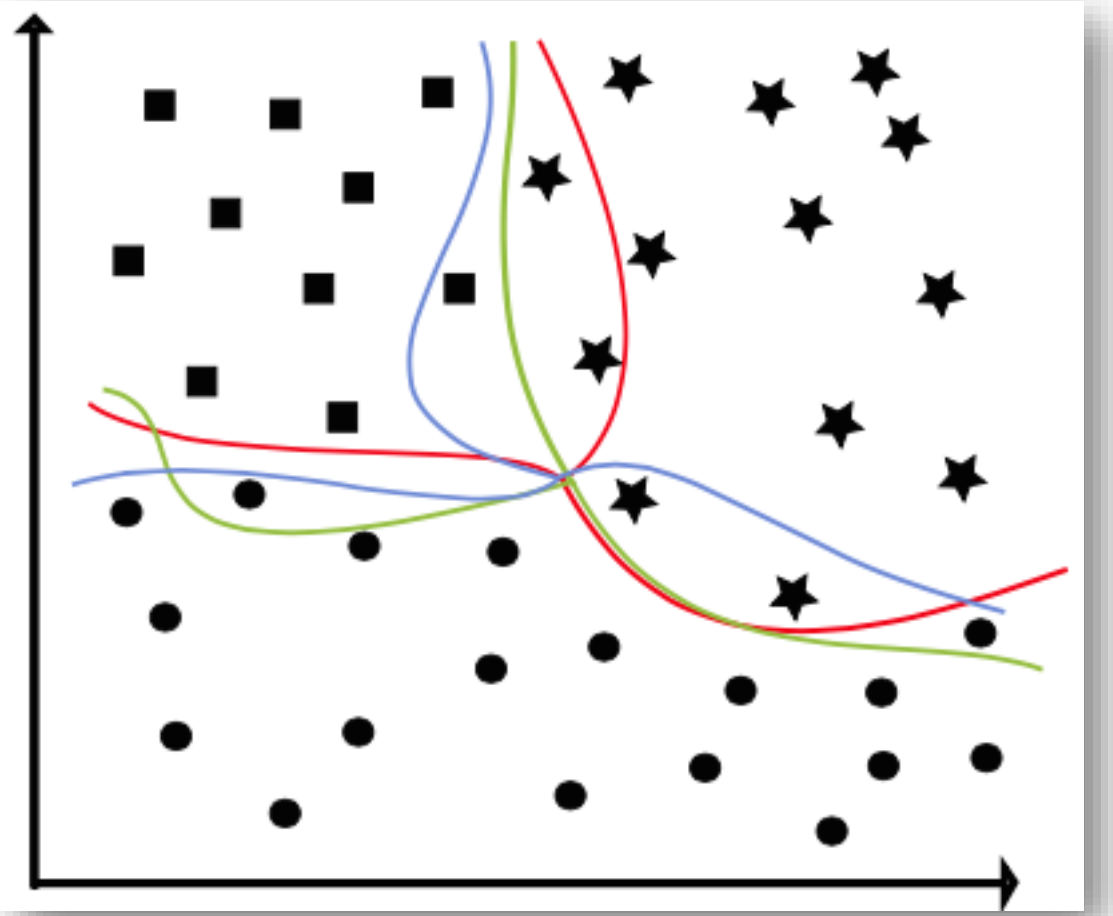


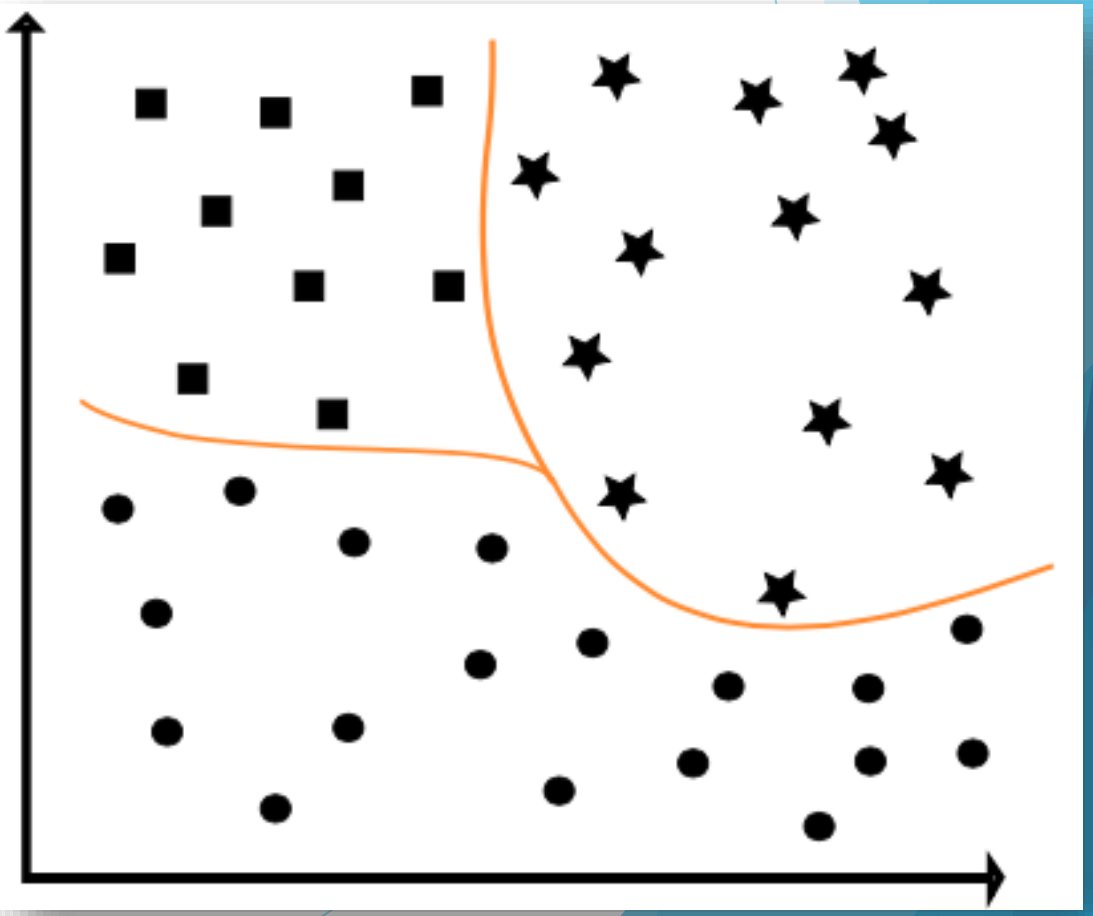
Por que não combinar as saídas dos classificadores?

Motivação e Definições

Isto é um *ensemble*!

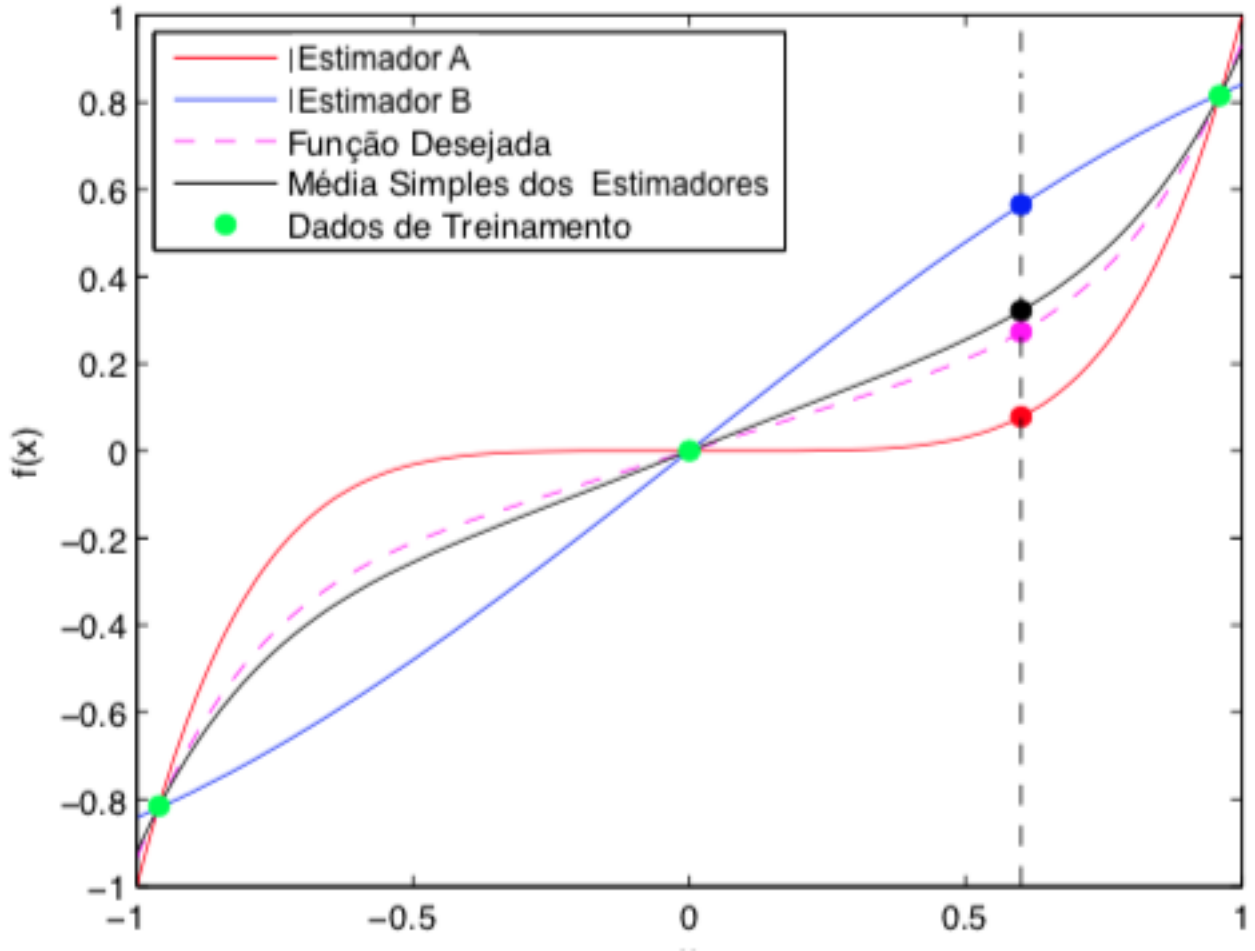
- Classificadores diferentes: saídas combinadas por **voto majoritário**





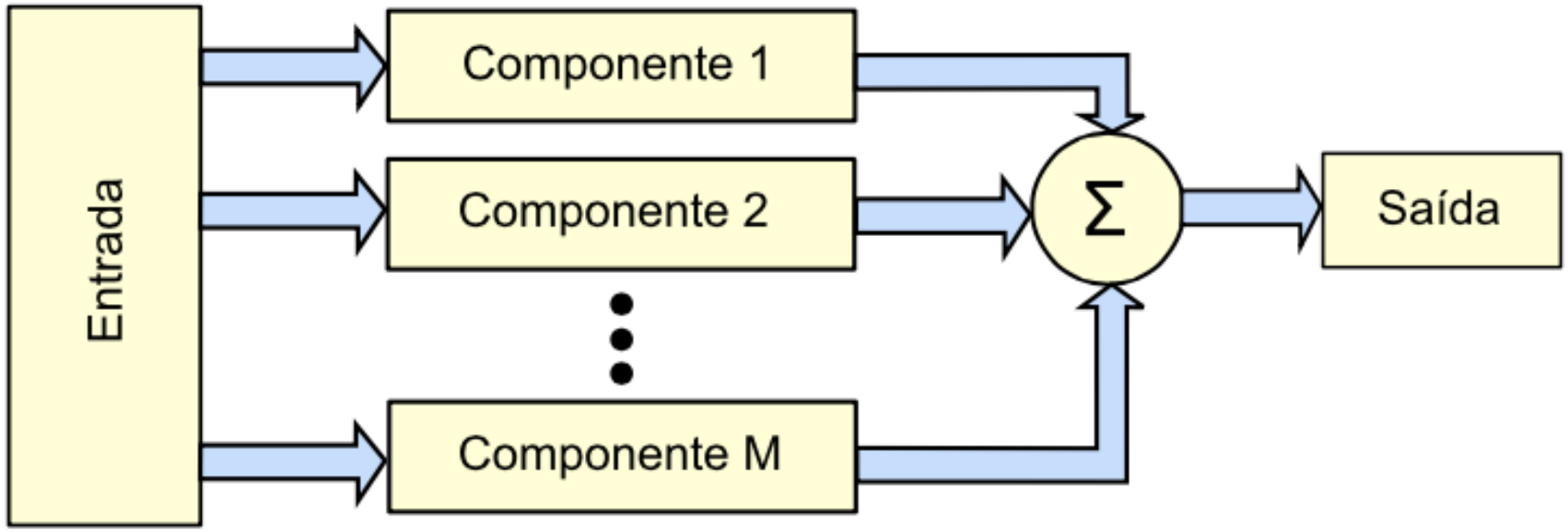
Motivação e Definições

- *Spoiler*: também é possível criar ensembles de **estimadores**



Ensembles

- Estrutura geral:



- Cada componente recebe a mesma entrada e gera uma saída;
- As saídas individuais (rótulos) são combinadas em uma única resposta para o problema.

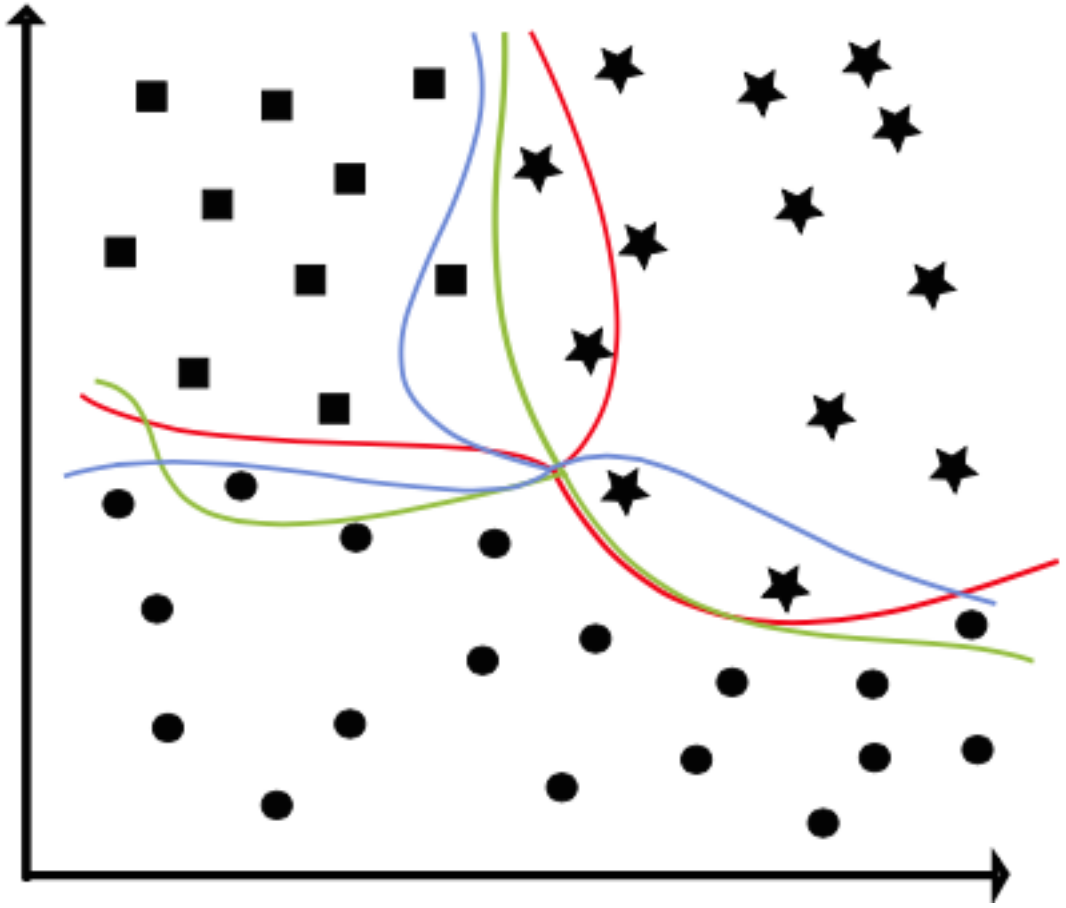
Ensembles

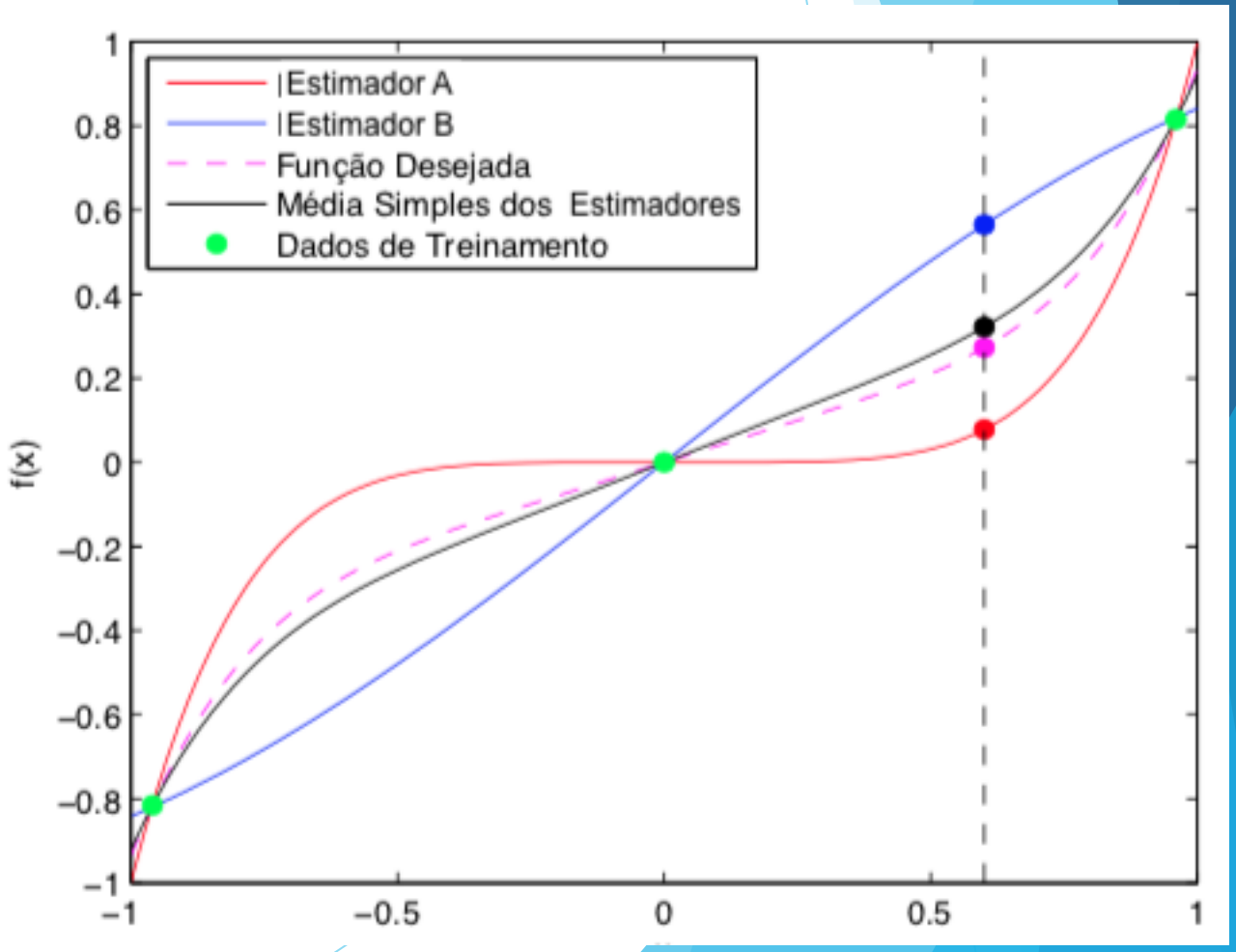
- Estudos mostraram que a **combinação de diferentes componentes** pode levar a ganhos significativos na capacidade de generalização do sistema;
  - **Capacidade de generalização**: responder bem a dados não vistos durante o treinamento;
- Exigência para os componentes:
  - Devem apresentar boa qualidade individualmente;
  - Devem apresentar **diversidade de erro**;
    - Não devem errar da mesma maneira.

A Questão da Diversidade

A Questão da Diversidade

- Diversidade de erro:





A Questão da Diversidade

- Formas de estimular a diversidade de erro:
  - Inicializar os algoritmos de treinamento com ***parâmetros diferentes*** (desde que seja possível - forma menos eficiente);
  - Utilizar componentes com ***arquiteturas diferentes*** (ex.: redes neurais com número diferente de neurônios nas camadas intermediárias);
  - Utilizar componentes baseados em ***paradigmas diferentes*** (ensemble heterogêneo);
  - Fornecer ***dados de treinamento*** (ligeiramente) ***diferentes*** para cada componente;
  - Treinar todos os componentes em conjunto, estimulando a diversidade.

A Questão da Diversidade - *Bagging*

- Uma das estratégias mais adotadas para estímulo de diversidade é conhecida como *bagging*;
- Supondo um conjunto *D* de dados, com $k = |D|$ amostras, para cada um dos *k* componentes do ensemble é feita uma amostragem em *D* de *k* amostras (com reposição);
  - Este tipo de amostragem é conhecido como "*bootstrap*";
  - Como a amostragem é feita com *reposição*, algumas amostras poderão ser **repetidas** em cada conjunto de treinamento, enquanto que outras podem **não estar presentes**;
    - Cada componente "aprende" aspectos ligeiramente diferentes do problema → **diversidade**.

Etapas de Construção

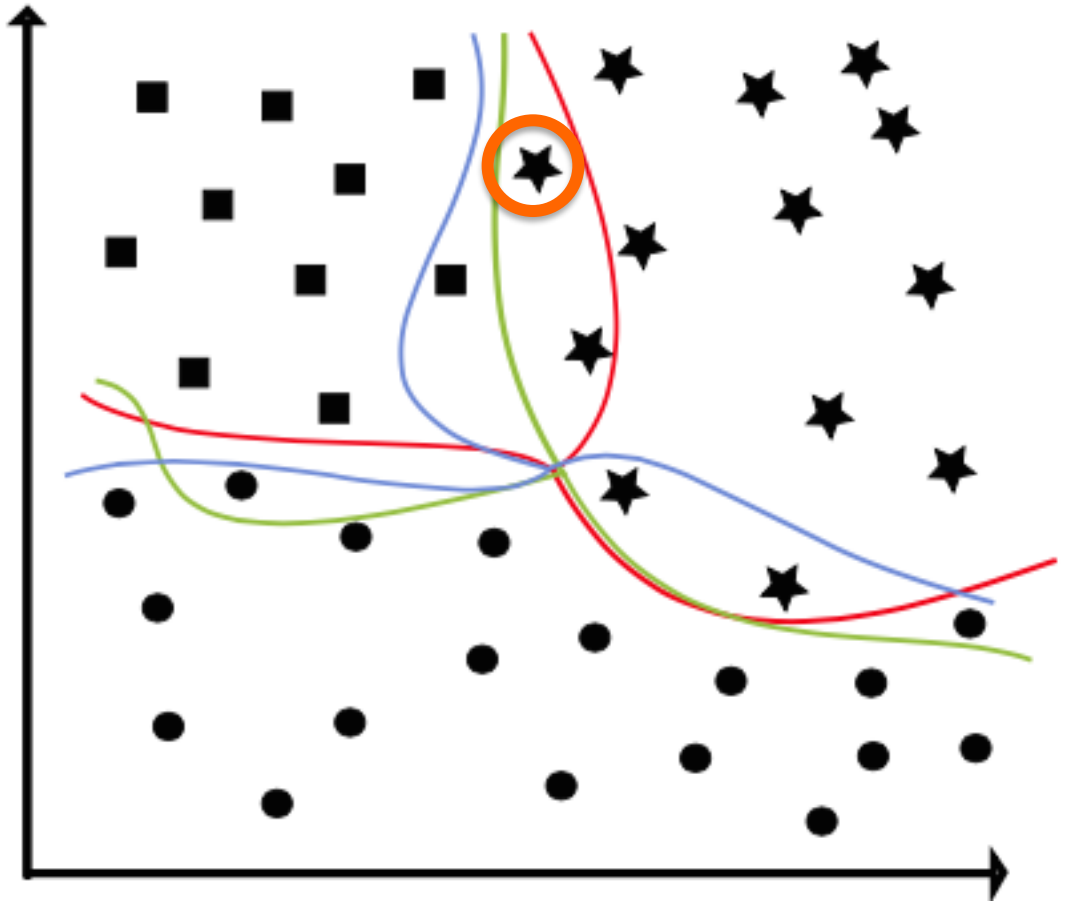
Etapas de Construção de um Ensemble

- Etapas de construção de um ensemble:
  - **Geração** (treinamento) de “candidatos” a componentes;
  - **Seleção** dos componentes (opcional);
  - Definição da estratégia de **combinação**;

- Caso seja necessário ajustar parâmetros para a estratégia de combinação, recomenda-se a utilização de uma parte do conjunto dos dados não utilizada no treinamento;
  - O mesmo vale para a etapa de seleção;
  - É preciso muitos dados!

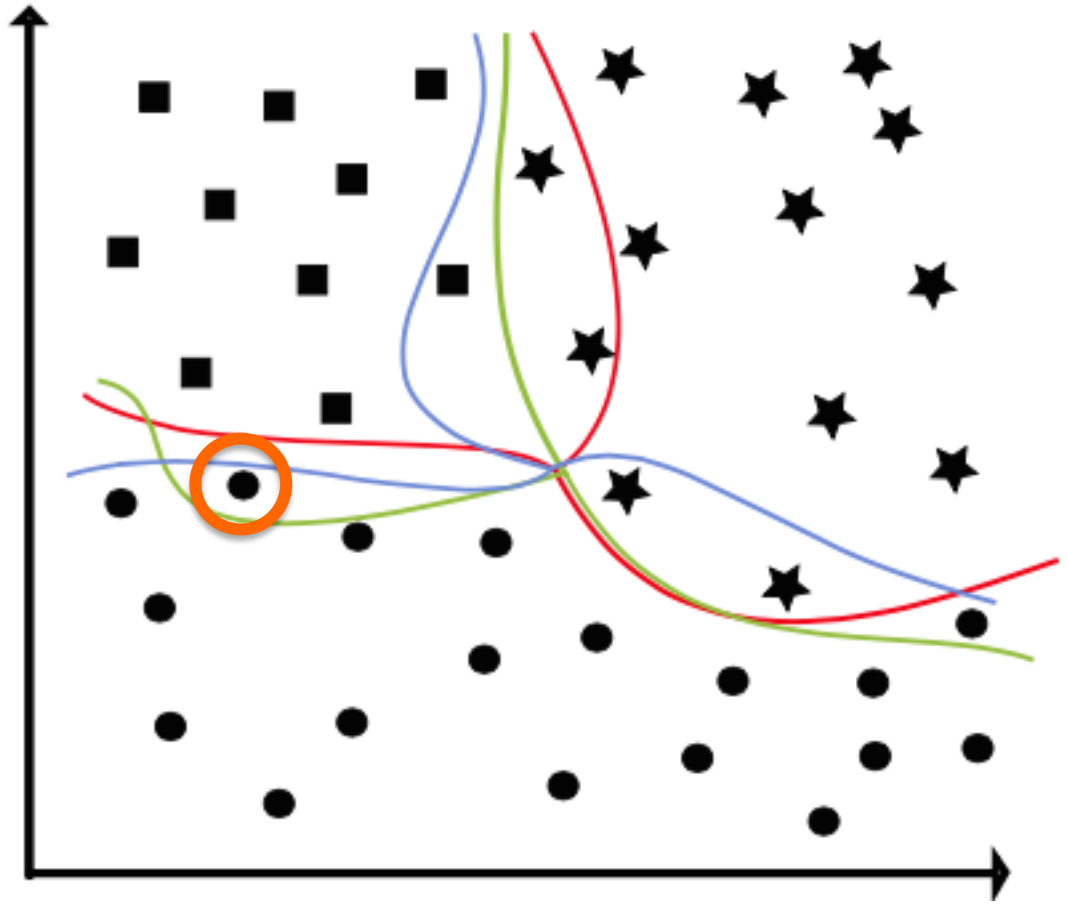
Etapas de Construção de um Ensemble

- Em problemas de ***classificação***, o **voto majoritário** é uma das abordagens mais diretas:
  - Para cada nova amostra dos dados, conta-se as indicações de rótulos (votos) de cada componente do *ensemble*;
  - Atribui-se à amostra o rótulo que obteve o maior número de votos.



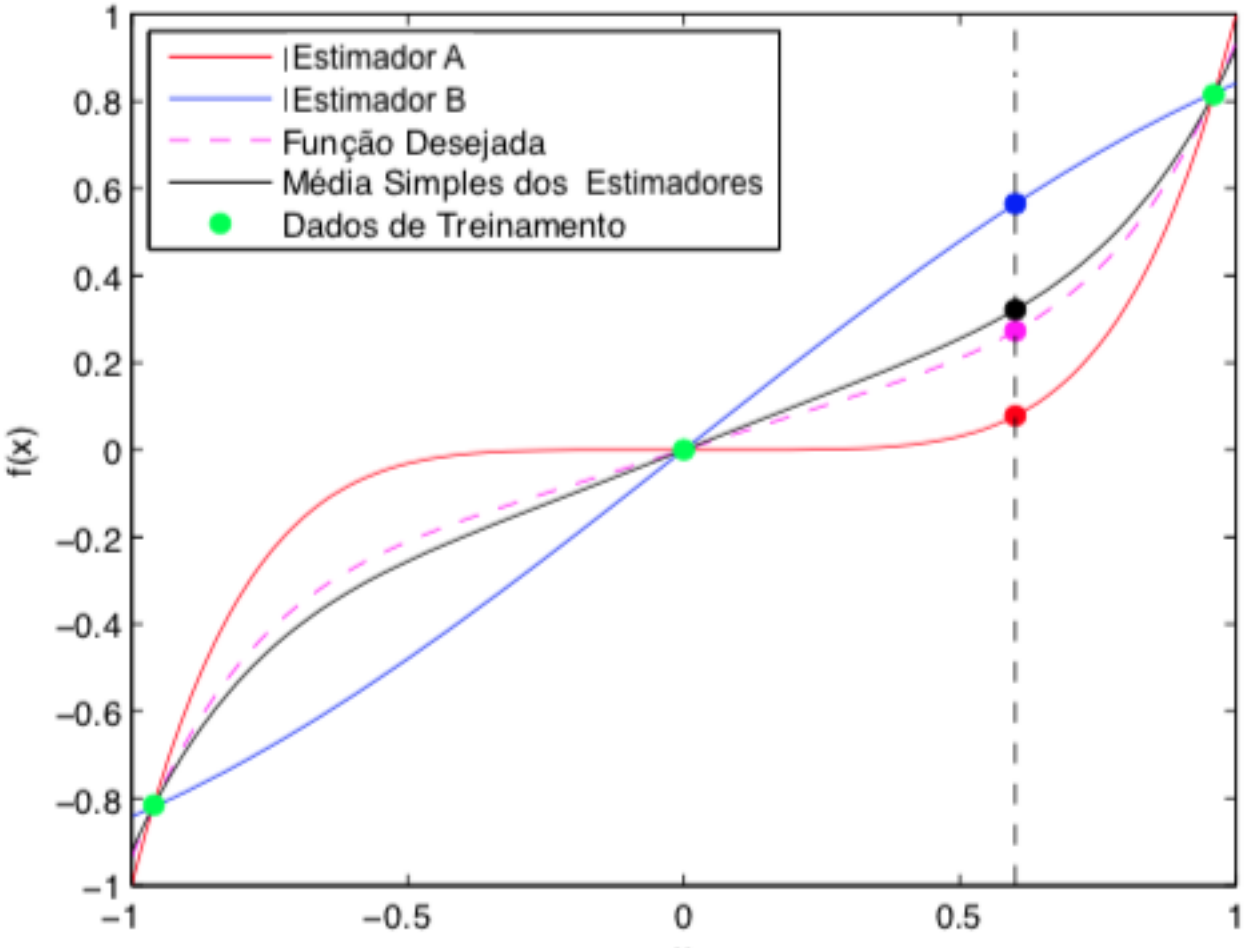
Etapas de Construção de um Ensemble

- Em problemas de ***classificação***, o **voto majoritário** é uma das abordagens mais diretas:
  - Para cada nova amostra dos dados, conta-se as indicações de rótulos (votos) de cada componente do *ensemble*;
  - Atribui-se à amostra o rótulo que obteve o maior número de votos.



Etapas de Construção de um Ensemble

- Em problemas de ***estimação***, a **média simples** é uma abordagens mais diretas:
  - Para cada nova amostra dos dados, soma-se as estimativas de cada modelo e divide-se pelo número de modelos.
- Alternativa: **média ponderada**.
  - Atribuir pesos às estimativas de cada modelo.



Referências Bibliográficas

Referências Bibliográficas

Han, J. & Kamber, M. “Data Mining: Concepts and Techniques”, Elsevier, 2006.

Witten, I. H., Frank E. & Hall, M. A. “Data Mining: Practical Machine Learning Tools and Techniques”, Elsevier, 2011.

Coelho, G. P. “Geração, Seleção e Combinação de Componentes para Ensembles de Redes Neurais Aplicadas a Problemas de Classificação”. Dissertação de Mestrado, Faculdade de Engenharia Elétrica e de Computação (FEEC), Unicamp, 2006.