Florestas Aleatórias

Ciência de Dados II

Professor: Gabriel Machado Lunardi
gabriel.lunardi@ufsm.br

Principais métodos preditivos

- ✓ Métodos baseados em distâncias
  - ✓ Algoritmo K-NN
- ✓ Métodos probabilísticos
  - ✓ Naive Bayes
  - ✓ Redes Bayesianas
- ✓ **Métodos baseados em procura**
  - ✓ **Árvores de decisão e regressão**
  - ✓ **Florestas Aleatórias (Random Forest)**
- ✓ Métodos baseados em otimização
  - ✓ Redes neurais artificiais
  - ✓ Máquinas de vetores de suporte (SVM)

Ensemble Learning

Combinação de vários modelos simples para criar um modelo melhor.

Tipos principais

- ✓**Bagging:** vários modelos independentes. Ex: Floresta Alea.
- ✓**Boosting:** modelos que aprendem com erros anteriores.
- ✓**Stacking:** combina modelos com um "meta-modelo".

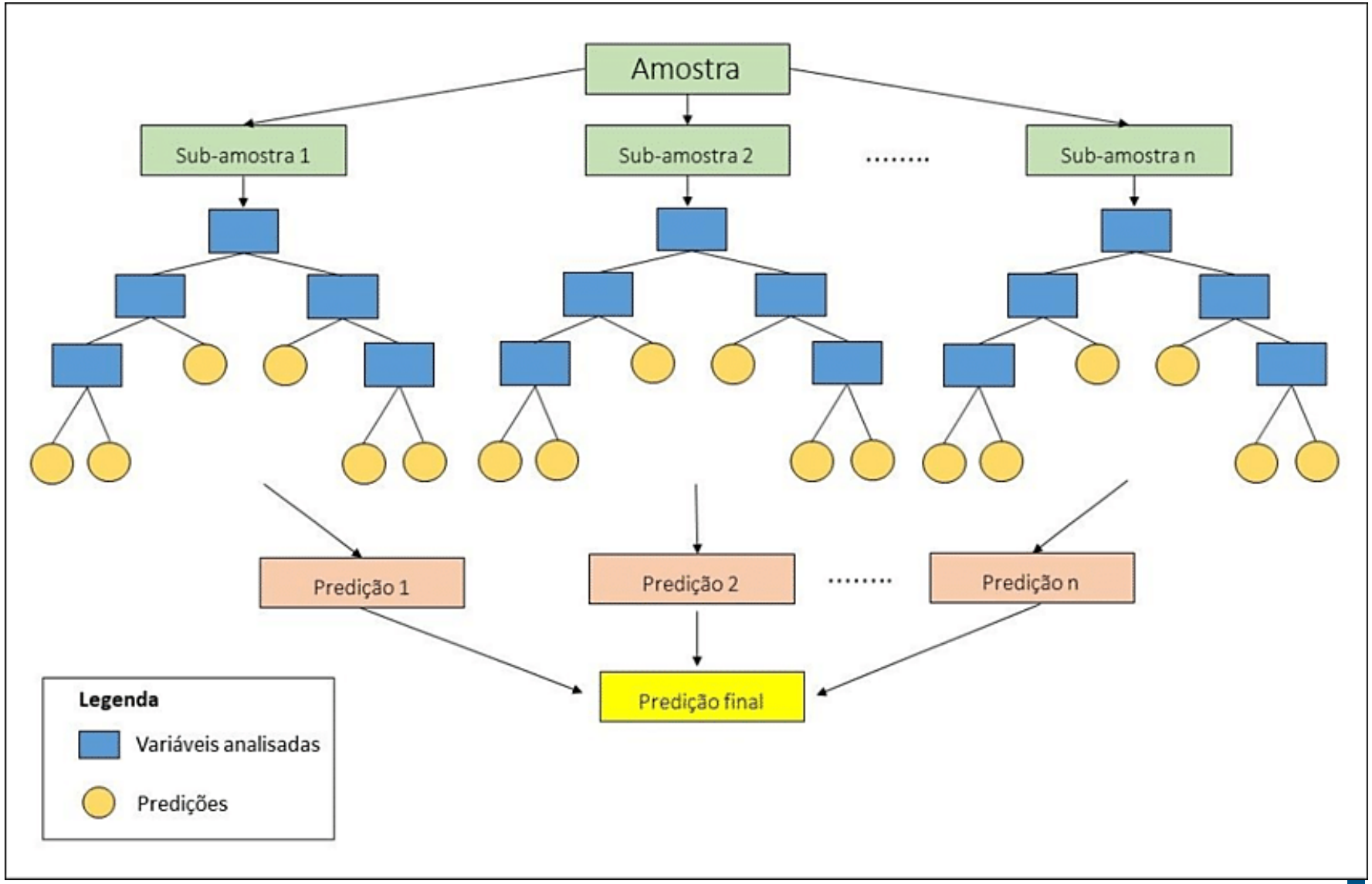
Floresta Aleatória (Random Forest)

Uma floresta de árvores de decisão!

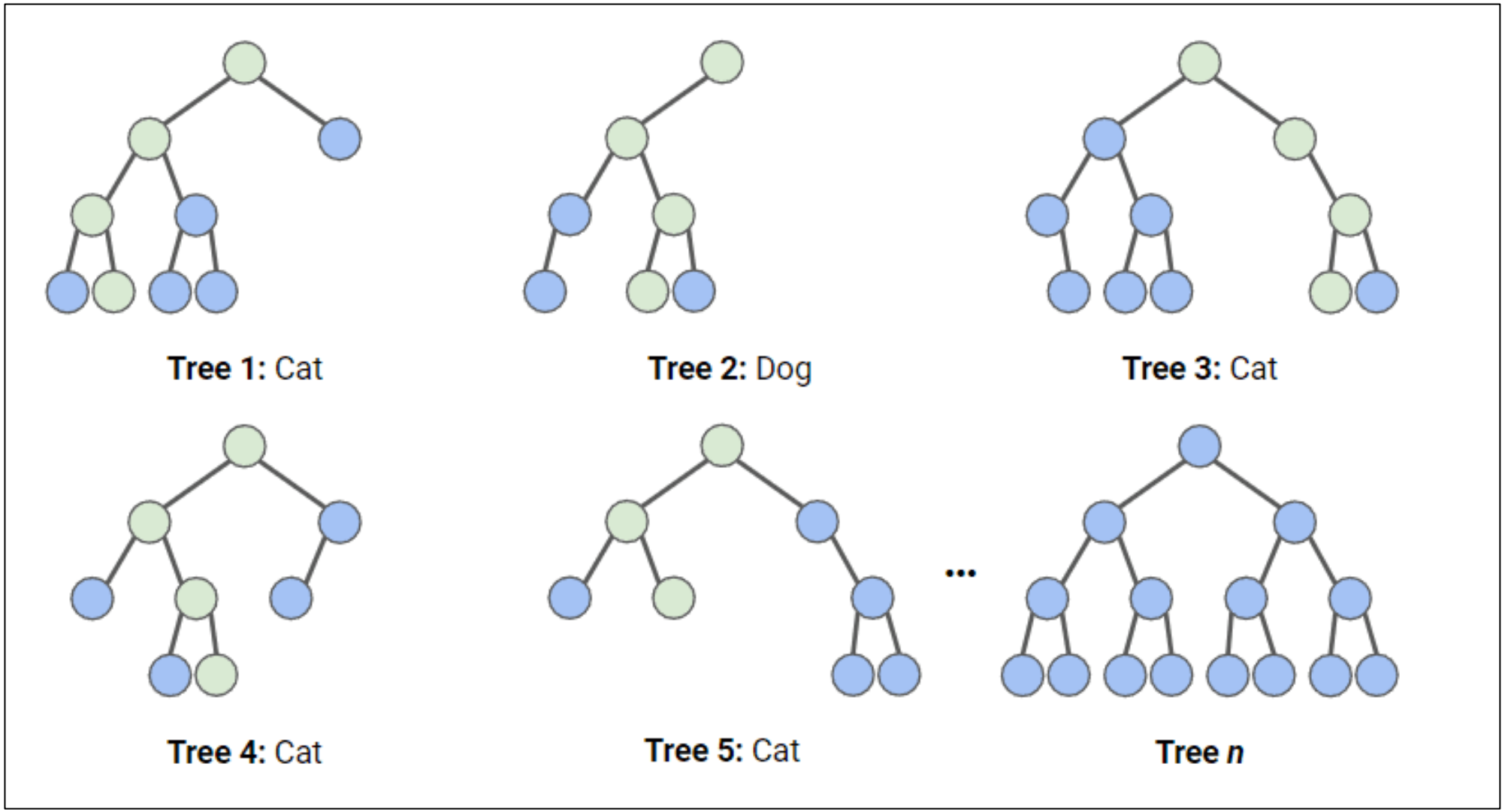
- ✓ Método de **bagging** que usa múltiplas árvores de decisão.
- ✓ Cada árvore é treinada em:
  - ✓ **Um subconjunto aleatório dos dados (bootstrap).**
  - ✓ **Um subconjunto aleatório das features.**

- ✓ A decisão final é por votação (classificação) ou média (regressão).

Floresta Aleatória



Floresta Aleatória



Conceito de Bootstrap

Técnica de **reamostragem com reposição** para criar múltiplos subconjuntos de dados a partir de um único dataset.

Urna com bolas numeradas (dados).

- ✓ Sorteamos uma bola
- ✓ anotamos o número
- ✓ **devolvemos** (reposição).

Repetimos isso várias vezes para formar um novo conjunto de dados.

Conceito de Bootstrap

Técnica de **reamostragem com reposição** para criar múltiplos subconjuntos de dados a partir de um único dataset.

Urna com bolas numeradas (dados).

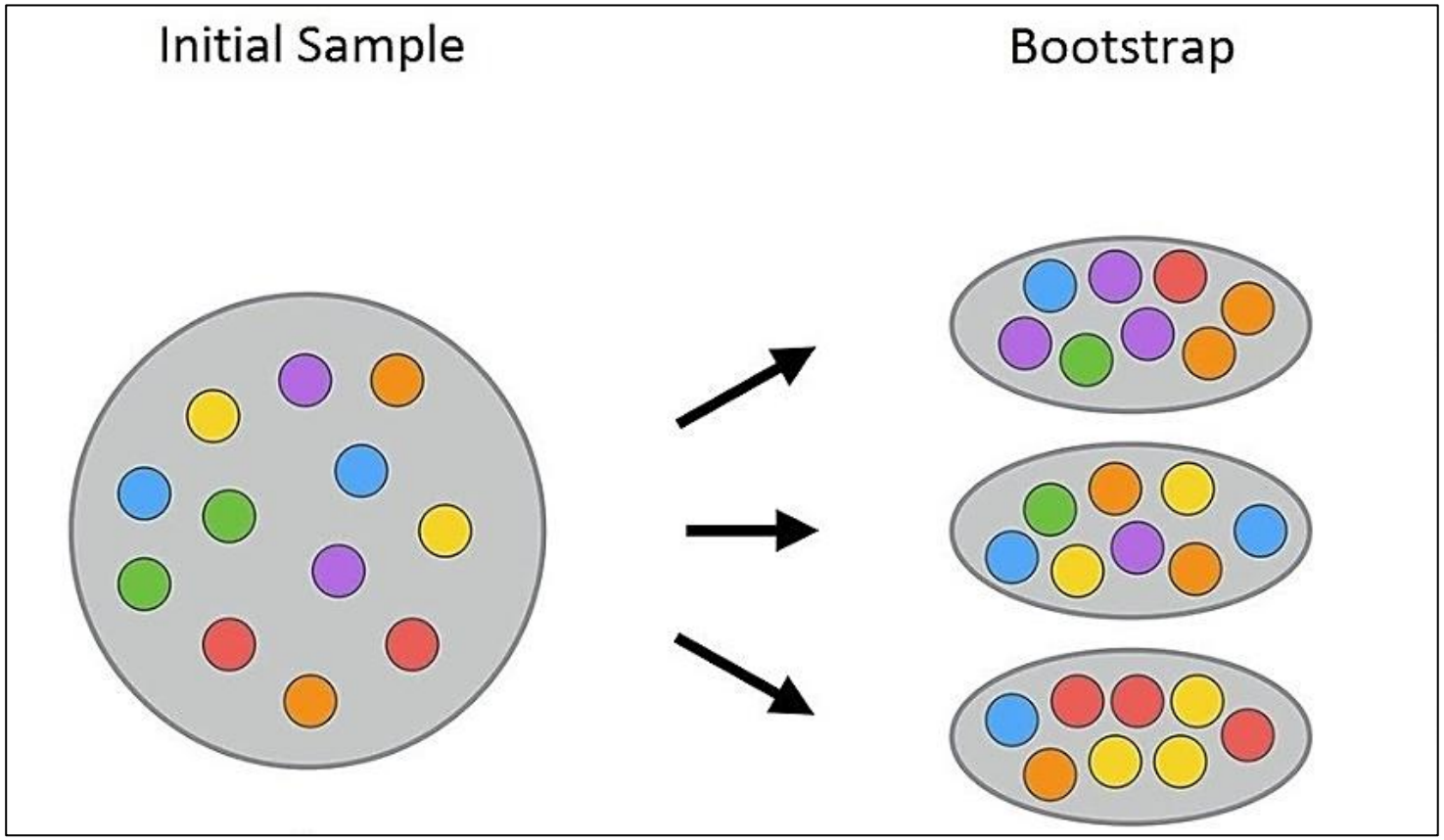
- ✓ Sorteamos uma bola
- ✓ anotamos o número
- ✓ **devolvemos** (reposição).

Repetimos isso várias vezes para formar um novo conjunto de dados.

Cuidado! Bootstrap é diferente de dividir os dados em treino e teste!

Conceito de Bootstrap

Técnica de **reamostragem com reposição** para criar múltiplos subconjuntos de dados a partir de um único dataset.



Dados repetidos não são um problema?

✗ Duplicatas no **dataset original** = problema (enviesa o modelo)

✓ Duplicatas no **bootstrap** = estratégia (cria diversidade controlada)

 **Treinador**

Divide o time em 3 grupos para treinar:

Grupo 1: 2 atacantes + 1 zagueiro (ênfase em ataque)

Grupo 2: 2 zagueiros + 1 atacante (ênfase em defesa)

Grupo 3: 2 laterais + 1 meia (ênfase em alas)

Resultado

✓ Time completo aprende **todos os aspectos do jogo**

✓ Cada grupo contribui com seu "viés especializado"

Benefícios do bootstraping

- ✓ Cada árvore recebe uma variação única dos dados.
- ✓ As repetições criam "versões levemente diferentes" do dataset original.
- ✓ Evita que todas as árvores aprendam exatamente os mesmos padrões
  - ✓ Redução do *overfitting*

Quando usar uma Floresta Aleatória?

- ✓ Dados complexos (não lineares, muitas features).
- ✓ Precisão é mais importante que interpretabilidade.
- ✓ Evitar overfitting (comparado a uma única árvore).

Quando não usar uma Floresta Aleatória?

- ✓Se interpretabilidade for mais importante.
- ✓Muitos dados e tempo limitado para processamento.
- ✓Dados muitos simples (poucas feautures/instâncias).