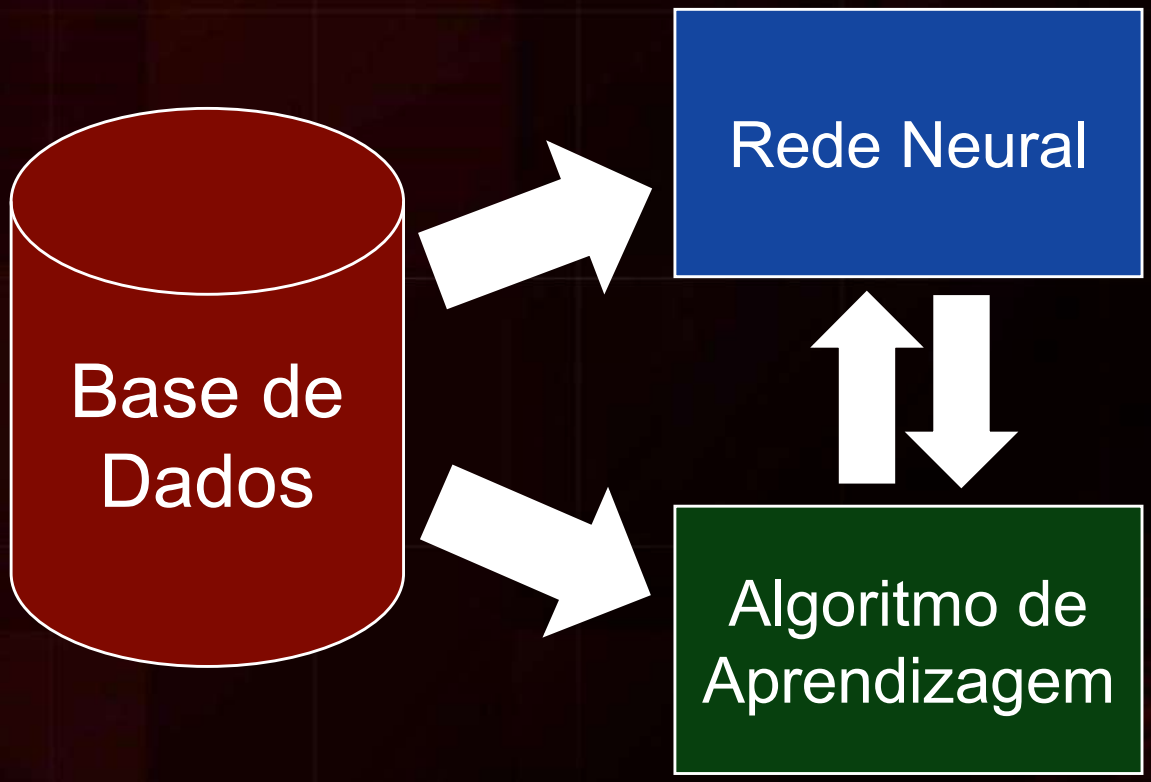
# REDES NEURAIS

## Preparação dos dados para treinamento

# TÓPICOS

1. Divisão do conjunto de dados
2. Validação Cruzada
3. Pré-processamento dos dados
    I. Transformação de dados categóricos em numéricos
    II. Normalização

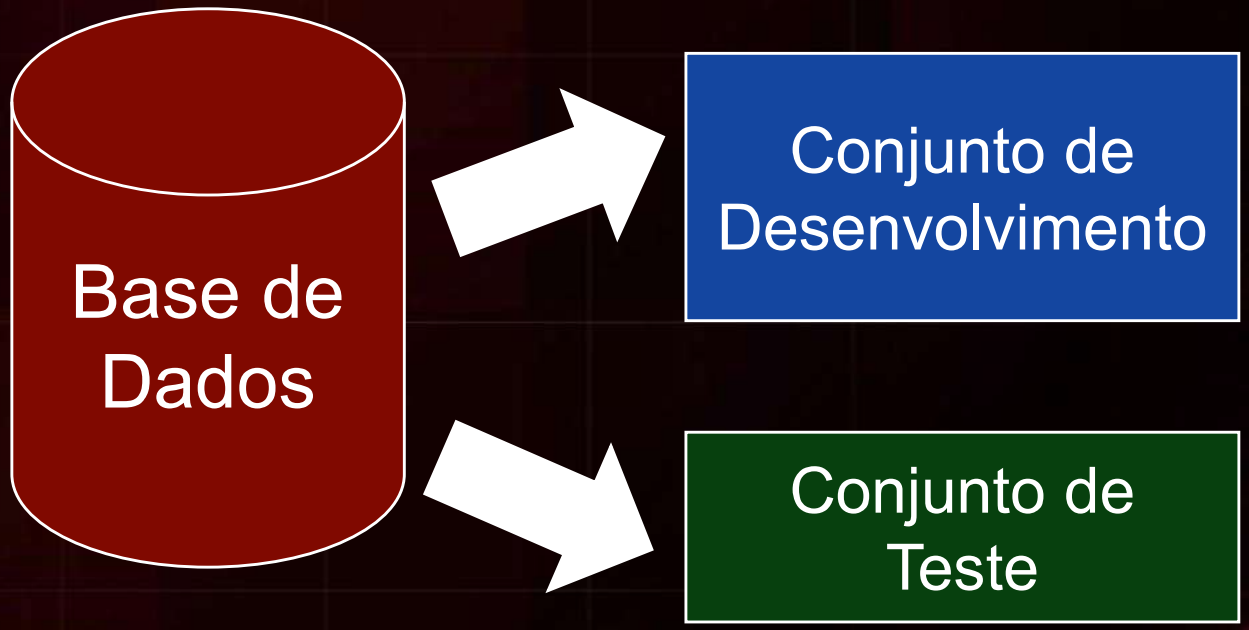
# DIVISÃO DO CONJUNTO DE DADOS




**AJUSTE DOS MOD**

1. **Hiperparâme**
2. **Parâmetros**

# DIVISÃO DO CONJUNTO DE DADOS




**CONJUNTO DE DESENVOLVIMENTO**

- **Usado para configu parâmetros e hiperparâmetros do**

**CONJUNTO DE TESTE**

- **Utilizado para avaliar a resposta do modelo final já treinado**
- **Não pode ser usado durante o desenvolvimento**

# DIVISÃO DO CONJUNTO DE DADOS

Conjunto de Desenvolvimento

Conjunto de Treino

Conjunto de Validação

- **Os exemplos dos conjuntos de treino e validação devem vir da mesma distribuição**
- **Treino → ajuste dos parâmetros**
- **Validação → ajuste dos hiperparâmetros**

# DIVISÃO DO CONJUNTO DE DADOS

Conjunto de Treino

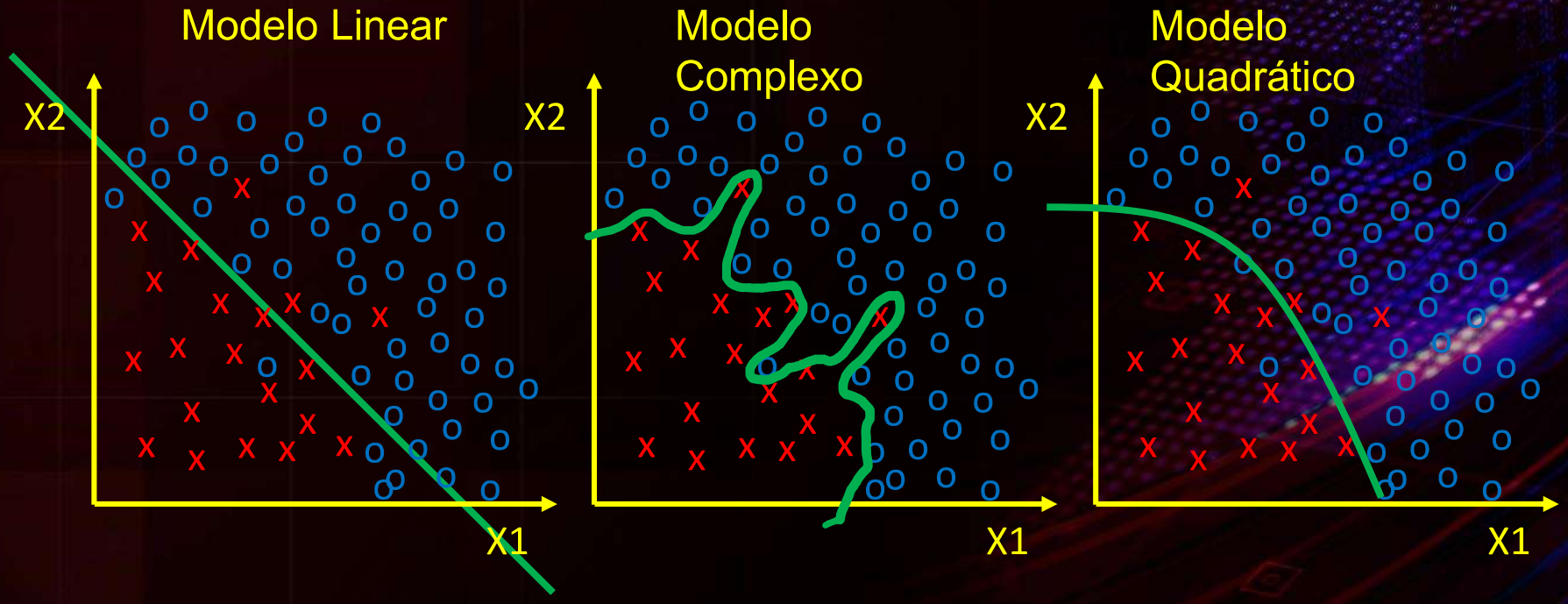
Conjunto de Validação

1. **Um modelo é configurado (seleção dos hiperparâmetros, i.e. número de camadas/neurônios)**
2. **O modelo é treinado com o conjunto de treino (parâmetros)**
3. **O modelo é avaliado com o conjunto de validação**
4. **Se a validação é ruim, os hiperparâmetros devem ser ajustados e o treino reiniciado (passo 1)**

# EXEMPLO

Conjunto de Treino

Conjunto de Validação




Alto viés - Underfitting

Alta variância Overfitting

Modelo adequado

# QUESTÕES

Conjunto de Treino

Conjunto de Validação

1. **Qual deve ser o tamanho de cada conjunto?**

**R: Depende**

**Regra 80 / 20 – bases pequenas**
**Para bases maiores, uma fração pequena pode ser considerada**

# QUESTÕES

Conjunto de Treino

Conjunto de Validação

**2. Como deve ser feita a divisão?**

**R: Seleção aleatória dos dados**

**Devemos garantir que ambos os subconjuntos sejam representativos**

# QUESTÕES

Conjunto de Treino

Conjunto de Validação

**3. O conjunto de teste é necessário?**

**R: Resposta curta: sim**

**Na prática: podemos reportar o erro de validação como uma estimativa do erro de teste**

# QUESTÕES

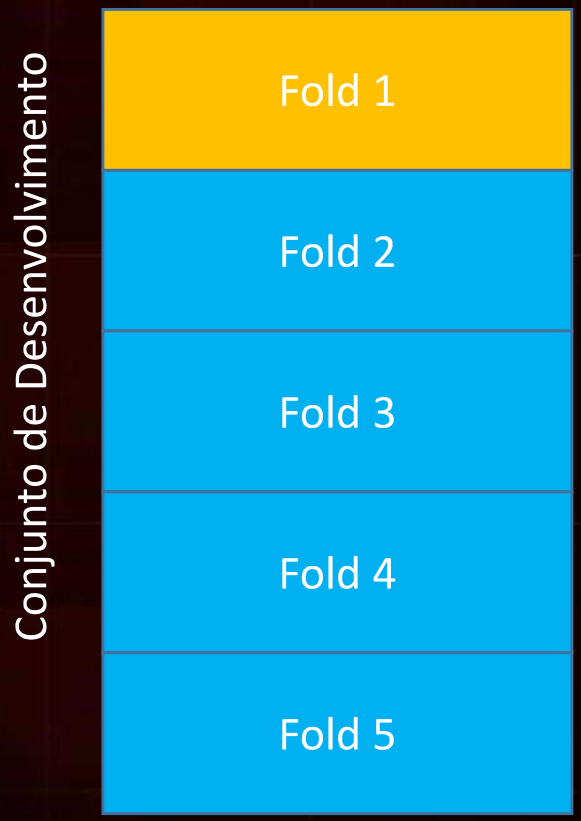
Conjunto de Treino

Conjunto de Validação

**4. Existem problemas com a abordagem baseada na divisão treino/validação?**

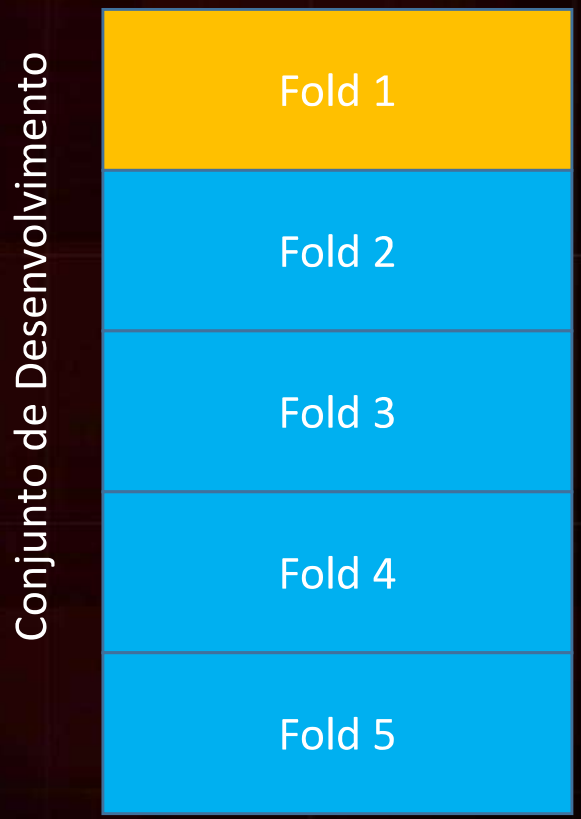
**R: A divisão do conjunto pode não ser adequada. Por exemplo, podemos ter amostras do conjunto de validação não representadas no conjunto de treino**

# VALIDAÇÃO CRUZADA




1. O modelo é treinado com 4 *folds* (azul) e validado com o *fold* extra (amarelo)
2. O processo é executado para todas as combinações
3. O erro de validação é o erro médio das execuções
4. k-fold cross-validation

# VALIDAÇÃO CRUZADA




## OUTRAS ABORDAGENS

1. Validação Cruzada Estratificada
2. Leave-One-Out
3. Estratificação por agrupamento (clustering)
4. Bootstrapping

# O CONJUNTO DE DADOS

- O conjunto de dados é a base para desenvolvimento dos modelos
- Dados reais podem conter problemas:

1. Atributos faltantes
2. Dados duplicados
3. Ruídos
4. Escalas incompatíveis, etc.

# PRÉ-PROCESSAMENTO DOS DADOS

- **O PRÉ-PROCESSAMENTO TEM POR OBJETIVO:**

  - ➢ **Melhorar a qualidade dos dados**
  - ➢ **Facilitar a aplicação de uma data técnica de aprendizado de máquina**
  - ➢ **Acelerar o processo de treinamento**

# PRÉ-PROCESSAMENTO DOS DADOS

1. Seleção e eliminação de atributos
   - Conhecimento do especialista
   - Atributos com variância próxima de zero
2. Redução de dimensionalidade
3. Amostragem dos dados
   - Redução do conjunto
   - Redução da redundância (duplicações)
4. Balanceamento de dados
5. Limpeza dos dados
   - Dados incompletos, inconsistentes, redundantes

# PRÉ-PROCESSAMENTO DOS DADOS


6. **Transformação dos dados**
   - **Conversão categórico-numérico**
   - **Normalização (reescala)**

# CONVERSÃO CATEGÓRICO-NUMÉRIC

- **CATEGÓRICOS COM DOIS VALORES**
  - **Nominal: 0 – não presente / 1 – presente**
    - **Exemplo: Febre? (0 ou 1)**
  - **Ordinal: 0 – menor valor / 1 – o segundo valor**

- **CATEGÓRICOS COM MAIS VALORES**

| Nominal | Codigo 1-de-c |
|---|---|
| Professor | 0 0 0 1 |
| Médico | 0 0 1 0 |
| Engenheiro | 0 1 0 0 |
| Músico | 1 0 0 0 |

| Ordial | Escalar |
|---|---|
| Ruim | 0 |
| Regular | 1 |
| Bom | 2 |
| Ótimo | 3 |

# TRANSFORMAÇÃO DOS ATRIBUTOS

- **Atributos podem assumir valores com escalas incompatíveis:**

| Atributo 01 | Atributo 02 | Classe |
|---|---|---|
| -0,01 | -1238940 | 0 |
| 0,02 | 87232667 | 1 |
| 0,03 | 9229893 | 1 |
| -0,01 | 2187287722 | 0 |

# TRANSFORMAÇÃO DOS ATRIBUTOS

- A normalização deve ser aplicada por atributo
- Tem por objetivo equilibrar a amplitude de valores (escala) de atributos distintos
- DUAS FORMAS PRINCIPAIS:
  - POR AMPLITUDE: transforma os dados em um intervalo fixo (min-max)
  - POR DISTRIBUIÇÃO: transforma os dados de tal forma que a média seja zero e o desvio um

# EXEMPLO NORMALIZAÇÃO

- **Por Amplitude**

$$v_{novo} = min + \frac{v_{atual} - menor}{maior - menor}(max - min)$$

| | | | | | |
|---|---|---|---|---|---|
| -5 | -4 | -2 | 2 | 3 | 5 |

Assumindo max = 1 e min = 0, temos:

| | | | | | |
|---|---|---|---|---|---|
| 0 | 0,1 | 0,3 | 0,7 | 0,8 | 1 |

# EXEMPLO NORMALIZAÇÃO

- **Por Distribuição**

$$v_{novo} = \frac{v_{atual} - \mu}{\sigma}$$

| | | | | | |
|---|---|---|---|---|---|
| -5 | -4 | -2 | 2 | 3 | 5 |

Média zero e desvio 1

| | | | | | |
|---|---|---|---|---|---|
| -1,187 | -0,941 | -0,5 | 0,53 | 0,78 | 1,27 |

# VANTAGENS E DESVANTAGENS

- A normalização deve ser usado com cuidado, pois pode reduzir (ou eliminar) a importância de um dado atributo
- A normalização por amplitude restringe os valores em um intervalo fixo, adequada quando precisamos estabelecer esses limites
- A normalização por distribuição tem tolerância maior a *outliers*

# O QUE VIMOS?

- Aprendemos a separar os dados para ajuste dos parâmetros, hiperparâmetros e teste dos modelos

- Conhecemos algumas técnicas de pré-processamento de dados, como a transformação categórico-numérico e a normalização

# PRÓXIMA VIDEOAULA

- Revisitar o neurônio MCP

- Introduzir o Perceptron e o Adaline