Lista de Exercícios – Correlações, Regressão e Validação Cruzada (k-Fold)

Disciplina: ICP363 – Introdução ao Aprendizado de Máquina

Revisão e Fundamentos

1. Explique, com suas palavras, o que é **Aprendizado Supervisionado**, destacando como o modelo aprende a partir de dados rotulados.
2. Relembre o conceito de **Regressão Linear** e discuta a diferença entre regressão simples e múltipla.
3. Defina **Correlação de Pearson** e **Correlação de Spearman**. Quais as principais diferenças em relação à sensibilidade a outliers e à linearidade da relação?
4. O que é **k-Fold Cross Validation**? Explique por que ela é importante na avaliação de modelos e como se diferencia de uma simples divisão treino-teste.

Correlação na prática com o dataset *Advertising*

Carregue o dataset `Advertising.csv` e calcule a matriz de correlação de Pearson e Spearman usando pandas. Compare os resultados com os métodos `pearsonr` e `spearmanr` do SciPy.

1. Quais variáveis têm correlação mais alta com `Sales`?
2. Há diferenças entre os valores de Pearson e Spearman? O que isso indica?
3. Interprete numericamente e conceitualmente o *p-valor* associado a cada correlação.

Experimentos visuais e análise exploratória

Baseando-se nos códigos dado no laboratório **Atividade Experimental 29_04_2025.pdf**: gere gráficos `pairplot` do Seaborn para observar relações bivariadas.

1. Observe visualmente se as relações são lineares, monotônicas ou não monotônicas.
2. Com base nos gráficos e nos coeficientes, qual correlação (Pearson ou Spearman) parece representar melhor os dados? Justifique.

Correlação e Regressão Linear com K-Fold Cross Validation

Use o `KFold` do `sklearn.model_selection` para realizar validação cruzada (5 folds) em uma regressão linear entre TV e Sales.

Listing 1: Validação cruzada com KFold e regressão linear

```
from sklearn.model_selection import KFold
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
from scipy.stats import pearsonr, spearmanr
import numpy as np

kf = KFold(n_splits=5, shuffle=True, random_state=42)
pearson_scores, spearman_scores, r2_scores = [], [], []

for train_index, test_index in kf.split(df):
    X_train, X_test = df[['TV']].iloc[train_index], df[['TV']].
        ↪ iloc[test_index]
    y_train, y_test = df['Sales'].iloc[train_index], df['Sales'].
        ↪ iloc[test_index]

    model = LinearRegression().fit(X_train, y_train)
    y_pred = model.predict(X_test)

    r2_scores.append(model.score(X_test, y_test))
    pearson_scores.append(pearsonr(y_test, y_pred)[0])
    spearman_scores.append(spearmanr(y_test, y_pred)[0])
```

- A correlação entre valores reais e previstos variou entre os folds?
- O comportamento de Pearson e Spearman foi semelhante?
- O que isso revela sobre a robustez da relação TV–Sales?
- Como o k-Fold ajuda a reduzir conclusões equivocadas sobre correlação?

Extensão: Impacto de Outliers nas Correlações

Listing 2: Adicionando ruído (outliers) aos dados

```
df_out = df.copy()
```

```
df_out.loc[np.random.choice(df.index, 5), 'Sales'] *= np.random.
    ↪ uniform(1.5, 3)
```

Recalcule as correlações Pearson e Spearman e analise:

- O que muda?
- Qual das duas medidas se mostrou mais estável frente aos outliers?
- Por que essa diferença é esperada teoricamente?
- Qual tipo de correlação é mais adequado para dados ruidosos?
- Como o k-Fold reforça a confiabilidade das análises?
- Quais suas principais descobertas numéricas e interpretativas?

Dataset alternativo (opcional):

`sns.load_dataset('tips')`

Analise a correlação entre `total_bill`, `tip` e `size`, comparando Pearson, Spearman e desempenho da regressão linear com k-Fold.

perceptron+k-fold

Para este exercício use o código perceptron que você implementou nos laboratórios anteriores. Use o dataset que simula o problema da porta lógica AND.

1. **Implementação do K-fold:** Usando o k-fold, com $k = 4$ para separar os dados:
2. Execute o código e registre os valores de acurácia para cada uma das 4 dobras (splits).
3. Calcule a acurácia média e o desvio padrão dos resultados. O que o desvio padrão nos diz sobre a estabilidade do modelo Perceptron neste dataset?
4. No material de revisão, a porta lógica AND foi resolvida com sucesso. O que os seus resultados de acurácia mostram sobre a capacidade do Perceptron de aprender essa relação? O modelo foi bem-sucedido em todas as dobras?
5. **Reflexão:**
    - O dataset da porta lógica AND é muito simples. Por que a validação cruzada ainda é uma boa prática, mesmo em casos como este?
    - Explique a diferença entre testar o modelo apenas uma vez com uma divisão `train_test_split` (como feito na regressão da atividade experimental) e testá-lo com K-fold. Por que o K-fold é considerado uma avaliação mais robusta e menos suscetível a "sorte"?