Métodos Computacionais em Física de Altas Energias

Exercício #1

Classificação binária – Instruções gerais
Dados: simulação de eventos no contexto do experimento LHCb com uma coluna de rótulos binários (*label*=1 para *sinal*, 0 para *contaminação/ruído*) e múltiplas *features* (*e.g.* cinemática).

Dataset:/shared_dataset/Met_Stat_HEP_AI_School/Kstarmumu_Combined_10K.csv
Imports: *numpy, pandas, scikit-learn, matplotlib*

Exercise 1: Explorando os dados

O objetivo da primeira etapa deste exercício é explorar o dataset e entender as características básicas que podem ser utilizadas em seguida para um aprendizado de máquina.

a) Estude visualmente as *features* disponíveis e compare suas distribuições de sinal e contaminação por meio de histogramas. Use cores ou marcadores diferentes para representar as duas classes.

b) Descreva quaisquer tendências visuais, fronteiras ou padrões de separabilidade entre os dois grupos. O que você observa sobre a distribuição geral dos eventos?

Exercise 2: Classificação baseada em cortes

a) Defina uma regra simples de classificação na forma de um limiar sobre uma variável escolhida (por exemplo, classificar como sinal se $y < y_{\text{limiar}}$). Justifique a escolha do limiar adotado.

b) Aplique essa regra para classificar cada evento do conjunto de dados.

c) Quantifique o desempenho da regra calculando:

- Eficiência de sinal (*true positive rate*);
- Rejeição da contaminação (*true negative rate*).

Exercise 3: Decisão linear

a) Construa uma regra de decisão linear da forma $y = m \cdot x + b$. Explique como foram determinados os valores da inclinação m e do intercepto b.

b) Use essa reta para classificar cada ponto do conjunto de dados.

c) Avalie o desempenho do classificador usando as mesmas métricas da etapa anterior: eficiência de sinal (*true positive rate*) e rejeição de contaminação (*true negative rate*).

d) Compare os resultados: como esse classificador linear se comporta em relação à regra baseada em limiar?

Exercise 4: Avaliação de desempenho

a) Construa as curvas ROC (*Receiver Operating Characteristic*) para ambos os classificadores – o baseado em corte e o linear manual.

b) Calcule a área sob cada curva (AUC, *Area Under the Curve*).

c) Compare os valores obtidos: qual método apresenta melhor separação entre sinal e fundo, e por que isso pode ocorrer?

Exercise 5: Regularização e *overfitting*

Vamos estudar um exemplo de regressão gerando uma amostra sintética baseada por exemplo em um $\sin(x)$ adicionando um ruído gaussiano, *e.g.* value_1 * np.random.randn(value_2).

a) Gere modelos de regressão polinomial de ordem crescente e avalie o erro de treino e o erro de teste em função da complexidade do modelo. Plote ambos em um mesmo gráfico e discuta o comportamento observado.

b) Aplique regularização do tipo L1 (Lasso) e L2 (Ridge) ao modelo e mostre como essas técnicas afetam os coeficientes e reduzem o *overfitting*.

c) Avalie o erro de treino e o erro de teste em função da complexidade do modelo. Plote ambos em um mesmo gráfico e discuta o comportamento observado.