Métodos Computacionais em Física de Altas Energias

Exercício #3

Seleção de dados – Instruções gerais

Neste exercício, você irá utilizar dados reais do experimento LHCb para o decaimento $B^0 \to K^{*0}\mu^+\mu^-$ e dados simulados correspondentes. O objetivo é estudar a distribuição de massa do B^0, aplicar cortes no score de um classificador XGBoost, realizar ajustes de máxima verossimilhança usando `zfit`, otimizar a figura de mérito (FoM), calcular a significância do sinal via teorema de Wilks e, por fim, aplicar técnicas de reweighting com `XGBoost` para corrigir discrepâncias entre dados e simulação.

Dataset:

/shared_dataset/Met_Stat_HEP_AI_School/Kstarmumu_Data_2016_RM_xgb.csv

/shared_dataset/Met_Stat_HEP_AI_School/Kstarmumu_MC_xgb.csv

/shared_dataset/Met_Stat_HEP_AI_School/Kstarmumu_Data_2016_Jpsi.csv

Imports: *NumPy, pandas, matplotlib, scikit-learn, zfit, hepstats, xgboost*

Exercise 1: Resultado do treinamento em dados reais

a) Carregue o dataset real e visualize a distribuição da variável `xgb_score` em escala linear e logarítmica.

b) Aplique um corte inicial e compare as distribuições de massa B0_M antes e depois do corte.

Exercise 2: Inferência estatística

Vamos utilizar o resultado do treinamento para tentar limpar nosso sinal $B^0 \to K^{*0}\mu^+\mu^-$ nos dados reais e obter uma estimativa de sua contribuição em relação ao background. Neste exemplo vamos utilizar a ferramenta zfit.

a) Selecione a janela de massa $[5150, 5600]$ MeV.

b) Modele o sinal com uma Gaussiana e o background com uma exponencial.

c) Faça o ajuste estendido usando `ExtendedUnbinnedNLL` e obtenha o yield de sinal (N_S).

d) Plote os dados e as PDFs ajustadas.

Exercise 3: Figure-of-Merit

Para melhor decidir qual corte melhor otimiza o sinal nos dados, vamos definir uma métrica (tipicamente chamado de Figure-of-Merit – FoM):

$$\text{FoM} = \frac{S}{\sqrt{S+B}}, \tag{1}$$

Para obter S e B, vamos fazer uma série de etapas:

a) Cálculo da eficiência de sinal:

- Utilize o dataset de Monte Carlo para obter a eficiência relativa de sinal para vários cortes em `xgb_score`.
- Construa uma tabela com as colunas: corte, eficiência relativa e yield estimado.

b) Estimativa do background em função do corte:

- Ajuste uma exponencial aos sidebands (e.g. $[5400, 5600]$ MeV).
- Use a integral da exponencial para estimar o número de eventos de background na janela do sinal.
- Repita o procedimento para cada valor de corte e construa uma tabela com B_{est}.

c) Figure-of-Merit e otimização do corte:

- Calcule a FoM $S/\sqrt{S+B}$ para cada corte.
- Identifique o valor ótimo de `xgb_score` que maximiza a FoM.
- Plote a curva da FoM e destaque o ponto ótimo.

Exercise 4: Significância do sinal

Utilizando o teorema de Wilks, vamos demonstrar como obter a significância de um dado sinal.

a) Realize o ajuste completo (Gauss + Exp) e o ajuste nulo (somente Exp).

b) Calcule $\Delta(2\ln L)$ e a significância estatística $Z = \sqrt{2\Delta \ln L}$.

c) Compare os resultados para cortes diferentes.

Exercise 5: Pre-processamento para correções de Monte Carlo

a) Use `hepstats.splot.compute_sweights` para obter pesos de sinal e background a partir do melhor ajuste.

b) Adicione os pesos ao DataFrame e verifique que $\sum w_{\text{sig}} \approx N_S$.

c) Compare distribuições de variáveis cinemáticas (e.g. $B0\_PT$) entre:

- Dados sem peso;
- Dados ponderados por w_{sig};

- MC verdadeiro (label=1).

Exercise 6: Canal de controle e Reweighting com XGBoost

a) Selecione o canal de controle $B \to K^* J/\psi(\mu^+\mu^-)$ e ajuste sua massa.

b) Calcule os sWeights para o sinal de J/ψ.

c) Compare distribuições de variáveis ($B0\_PT$, $B0\_ENDVERTEX\_CHI2$) entre:

- Dados J/ψ (sWeighted);
- MC original do sinal.

d) Treine um reweighter baseado em `XGBClassifier` usando essas variáveis.

e) Avalie o desempenho com a curva ROC e o valor de AUC.

f) Plote as distribuições reponderadas e discuta se o reweighting melhorou o acordo entre dados e simulação.