Data Analysis II - 2023

Exercise sheet no 3: March 21, 2023
Signal selection

- `pandas.DataFrame`
- `scikit-learn`

Exercise 1: Signal selection (20 Points)

The file `signal.txt` and `background.txt` each contain 10K values of seven features used to characterise signal and background for a particle decaying into three other particles: $P \to P1\,P2\,P3$. The signal is real particle decays and the background is random combinations of particles which do not represent anything physical.

(a) Make a histogram for each feature, overlaying the signal and background on the same plot, as seen in slide 8 of the lecture. (2 points)

(b) For each feature, calculate the Fisher score and use it to rank the importance of the features. (4 points)

(c) Select the three highest features and optimise rectangular selection to maximise the metric

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

defined in the lecture. (4 points)

(d) Use a BDT to optimise the accuracy with those three features. Be careful that you do not train the algorithm on the same data you use to calculate the signal and background efficiencies as this will cause an overfitting bias. There is an example skeleton script in OLAT which should help you get started. How does the accuracy compare to the rectangular selection? (8 points)

(e) Include the full set of seven features and recalculate the obtained accuracy using the BDT. How does the accuracy improve? (2 points)

For interest: Feature definitions

- PT1: Momentum component transverse to beam direction for first particle in decay.
- PT2: Momentum component transverse to beam direction for second particle in decay.
- P1: Momentum magnitude for first particle in decay.
- P2: Momentum magnitude for second particle in decay.
- TotalPT: Vectoral sum of the entire decay.
- VertexChisq: χ^2 of decay vertex fit (geometric quality). Low values are good!

- Isolation: Estimate of how isolated the signal is.

Deadline for submission: Tuesday, 4 April 2023 18:00
Form: Please submit your solutions to OLAT. The solutions should be submitted as a single python script which creates the answers to the questions.