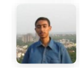
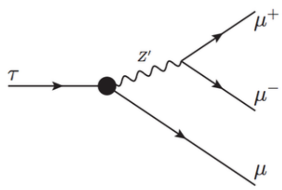
kaggle Search kaggle Competitions Datasets Kernels Discussion Jobs





Flavours of Physics: Finding τ → μμμ

Identify a rare decay phenomenon

$15,000 · 673 teams · 2 years ago

Overview Data Kernels Discussion Leaderboard Rules Late Submission

Competition Data

Edit

- check_agreement.csv....
- check_correlation.cs...
- sample_submission.cs...
- test.csv.zip
- training.csv.zip

test.csv.zip 260.35 MB

Download

Data Description

In this competition, you are given a list of collision events and their properties. You will then predict whether a τ → 3μ decay happened in this collision. This τ → 3μ is currently assumed by scientists **not** to happen, and the goal of this competition is to discover τ → 3μ happening more frequently than scientists currently can understand.

It is challenging to design a machine learning problem for something you have never observed before. Scientists at CERN developed the following designs to achieve the goal.

training.csv

This is a labelled dataset (the label 'signal' being '1' for signal events, '0' for background events) to train the classifier. Signal events have been simulated, while background events are real data.

This real data is collected by the LHCb detectors observing collisions of accelerated particles with a specific mass range in which τ → 3μ can't happen. We call these events "background" and label them 0.

- FlightDistance - Distance between τ and PV (primary vertex, the original protons collision point).
- FlightDistanceError - Error on FlightDistance.
- mass - reconstructed τ candidate invariant mass, which is **absent in the test samples**.
- LifeTime - Life time of tau candidate.

- IP - Impact Parameter of tau candidate.
- IPSig - Significance of Impact Parameter.
- VertexChi2 - $\chi 2$ of τ vertex.
- dira - Cosine of the angle between the τ momentum and line between PV and tau vertex.
- pt - transverse momentum of τ.
- DOCAone - Distance of Closest Approach between p0 and p1.
- DOCAtwo - Distance of Closest Approach between p1 and p2.
- DOCAthree - Distance of Closest Approach between p0 and p2.
- IP_p0p2 - Impact parameter of the p0 and p2 pair.
- IP_p1p2 - Impact parameter of the p1 and p2 pair.
- isolationa - Track isolation variable.
- isolationb - Track isolation variable.
- isolationc - Track isolation variable.
- isolationd - Track isolation variable.
- isolatione - Track isolation variable.
- isolationf - Track isolation variable.
- iso - Track isolation variable.
- CDF1 - Cone isolation variable.
- CDF2 - Cone isolation variable.
- CDF3 - Cone isolation variable.
- production - source of τ. This variable is **absent in the test samples**.
- ISO_SumBDT - Track isolation variable.
- p0_IsoBDT - Track isolation variable.
- p1_IsoBDT - Track isolation variable.
- p2_IsoBDT - Track isolation variable.
- p0_track_Chi2Dof - Quality of p0 muon track.
- p1_track_Chi2Dof - Quality of p1 muon track.
- p2_track_Chi2Dof - Quality of p2 muon track.
- p0_pt - Transverse momentum of p0 muon.
- p0_p - Momentum of p0 muon.
- p0_eta - Pseudorapidity of p0 muon.
- p0_IP - Impact parameter of p0 muon.
- p0_IPSig - Impact Parameter Significance of p0 muon.
- p1_pt - Transverse momentum of p1 muon.
- p1_p - Momentum of p1 muon.
- p1_eta - Pseudorapidity of p1 muon.
- p1_IP - Impact parameter of p1 muon.
- p1_IPSig - Impact Parameter Significance of p1 muon.
- p2_pt - Transverse momentum of p2 muon.
- p2_p - Momentum of p2 muon.

- p2_eta - Pseudorapidity of p2 muon.
- p2_IP - Impact parameter of p2 muon.
- p2_IPSig - Impact Parameter Significance of p2 muon.
- SPDhits - Number of hits in the SPD detector.
- min_ANNmuon - Muon identification. LHCb collaboration trains Artificial Neural Networks (ANN) from informations from RICH, ECAL, HCAL, Muon system to distinguish muons from other particles. This variables denotes the minimum of the three muons ANN. min ANNmuon should not be used for training. **This variable is absent in the test samples.**
- **signal - This is the target variable for you to predict in the test samples.**

test.csv

The test dataset has all the columns that training.csv has, except **mass**, **production**, **min_ANNmuon**, and **signal**.

The test dataset consists of a few parts:

1. simulated signal events for the τ → 3μ
2. real background data for the τ → 3μ
3. simulated events for the control channel, (ignored for scoring, used by agreement test)
4. real data for the control channel (ignored for scoring, used by agreement test)

You need to submit predictions for ALL the test entries. You will need to treat them all the same and predict as if they are all the same channel's collision events.

A submission is only scored after passing both the agreement test and the correlation test.

check_agreement.csv: Ds → φπ data

This dataset contains simulated and real events from the Control channel Ds → φπ to evaluate your simulated-real data of submission agreement locally. It contains the same columns as **test.csv** and `weight` column. For more details see agreement test.

check_correlation.csv

This dataset contains only real background events recorded at LHCb to evaluate your submission correlation with mass locally. It contains the same columns as **test.csv** and `mass` column to check correlation with. For more details see correlation test.