Machine learning

Kong Lingteng

2018.9.7

Machine learning

- Machine learning is a field of computer science that uses statistical techniques to give computer systems the ability to "learn“.
- Approaches: Decision tree learning, Artificial neural networks, etc.

Decision tree

- Decision tree is a kind of classification method based on tree structure. A decision tree has one root node, serval internal nodes and serval leaf nodes.
- Information entropy is a common method to measure sample purity.

$$\mathrm{Ent(D)} = -\sum_{k=1}^{|y|} p_k \log_2 p_k$$

(the proportion of k in set D is pk)

- The purity of the sample is higher when Ent(D) is smaller.
- When using property "a" to classify sample "D", we can use information gain to measure the improvement of purity.

$$Gain(D,a) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v)$$

- ID3 decision tree method is based on information gain.
- However, information gain prefers properties which have more Available values. We can use gain ratio to solve this problem.

$$SplitInfo_{A}(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2(\frac{|D_j|}{D}) \qquad GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)}$$

- C4.5 decision tree method is based on gain ratio.

- Former work: using decision tree to classify iris
- How to accomplish decision in python:
- Key function:
- SelectKBest(score_func=<function f_classif>, k=10): Select features according to the k highest scores.
- DecisionTreeClassifier: A decision tree classifier.

Boosting

- Ensemble learning method can get classification results by constructing many learners to make a vote on classification results. Individual learner is taken from training set using existing algorithms like C4.5 decision tree algorithm.
- Working mechanism of boosting : Get a basic learner from Initial training set, then adjust the training sample distribution based on the performance of learner, make the samples with which basic learner had something wrong got more attention in the following process, and then get the next basic learner based on the adjusted samples.

AdaBoost algorithm

- Use the linear combination of basic learner to minimize the exponential loss function l_{exp}
- $\mathrm{H}(x) = \sum_{t=1}^{T} \alpha_t \, h_t(x)$ $\qquad$ $l_{exp}(H|D) = E_{x\sim D}\left[e^{-f(x)H(x)}\right]$
- Input:
- training set D={(x1,y1),(x2,y2),⋯,(xm,ym)}
- Basic learner E
- Iterations T
- Process:
- 1.D1(x)=1/m.
- 2.for t=1,2,⋯,T do
- 3. ht=E(D,Dt)
- 4. $et = P_{x\sim Dt}\left(ht(x)! = f(X)\right)$
- 5. if et>0.5 then break
- 6. $\alpha_t = \frac{1}{2}\ln(\frac{1-et}{et})$
- 7. $D_{t+1}(x) = \frac{D_t(X)}{Z_t} \times \begin{cases} \exp(-\alpha_t),\ if\ h_t(x) = f(x) \\ exp(\alpha_t), if\ h_t(x) \neq f(x) \end{cases} = \frac{D_t(x)\exp(-\alpha_t f(x) h_t(x))}{Z_t}$
- 8.end for
- output: $\mathrm{H(x)} = \mathrm{sign}(\sum_{t=1}^{T} \alpha_t h_t(x))$

机器学习

周志华 著

清华大学出版社

周志华 著

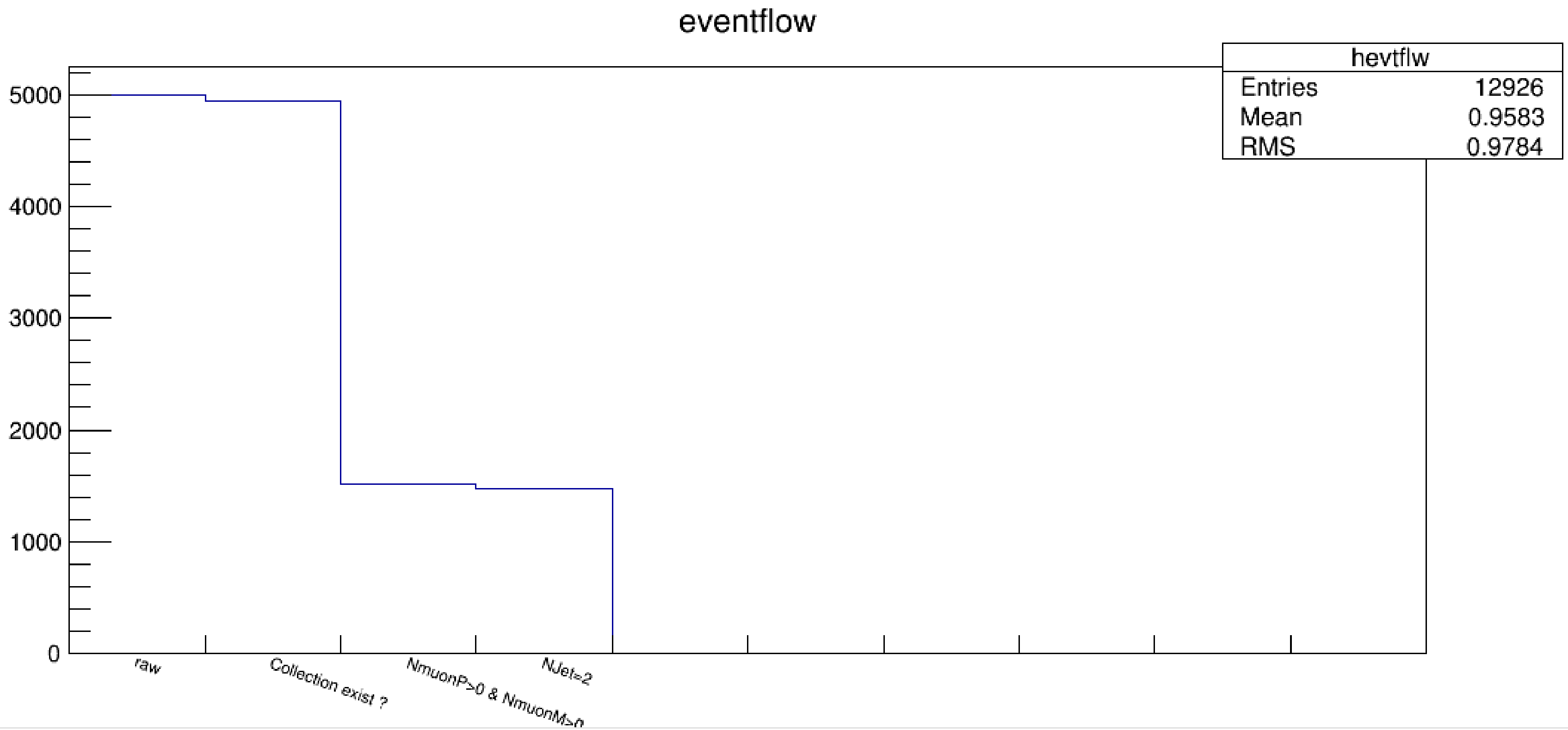
MACHINE LEARNING

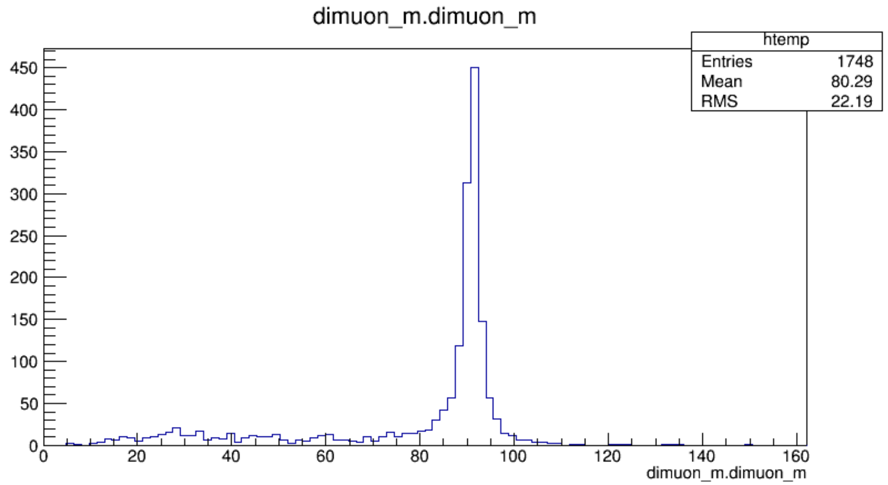
机器学习

清华大学出版社

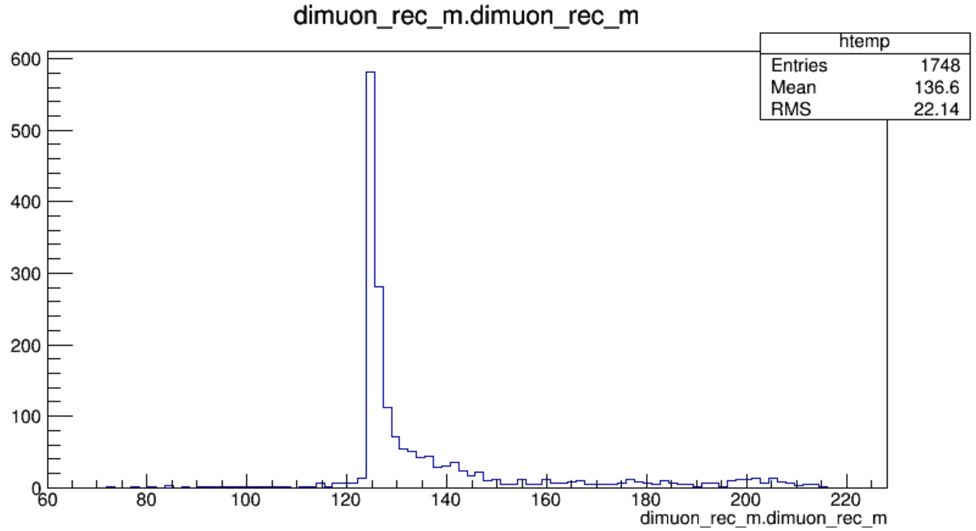
Hig2zz

- Run signal sample
- MaxRecordNumber" 5000

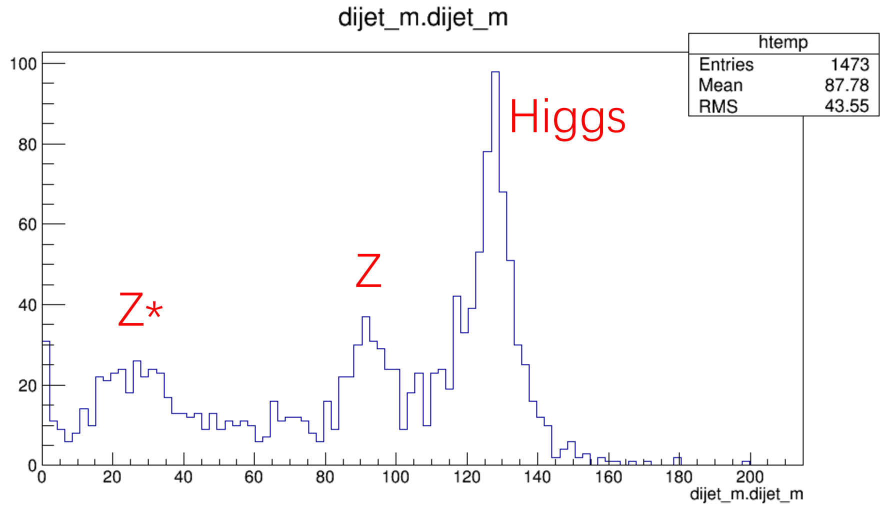




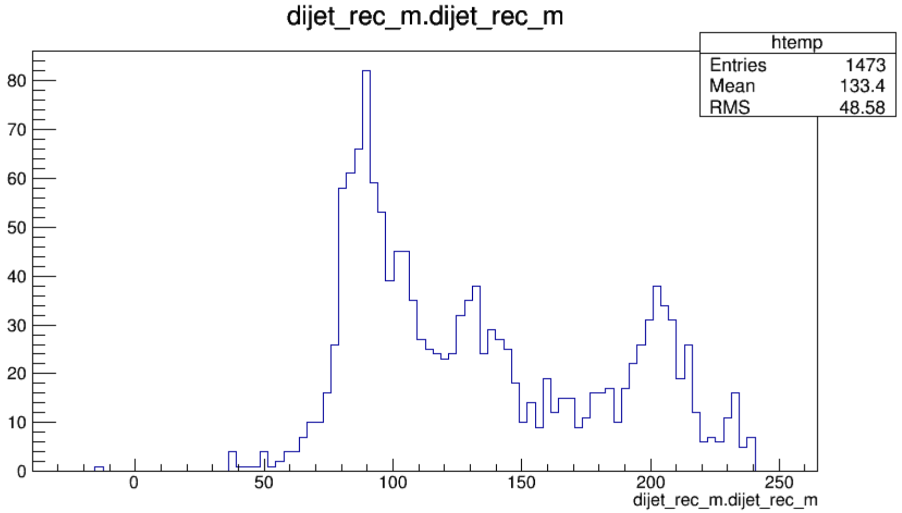
invariant mass of two-muons



recoil mas of two-muons



invariant mass of two-jets



recoil mas of two-jets