Applying k-means clustering in diHiggs signal/ background classification

Motivation

- Sometimes it may be hard for learning algorithms(like BDT/NN) to learn the relation between some variables in our training datasets

- Clustering is an unsupervised learning algorithm that can recognize similar events and put them in the same “cluster”

- Instead of using original variables, using the variables transformed by k-means clustering might give our learning algorithm a dataset that show more difference between signal and background

Approach

- To see whether k-means clustering is helpful in our case, we tried the following:
  - Run BDT with original variables
  - Run k-means clustering on original variables and feed the transformed variables to BDT
  - Run PCA (Principal Component Analysis) on the original variables and run k-means clustering on the PCA-tranformed variables, then feed the k-means-transformed variables into BDT

Original variables

Accuracy	Significance	N_sig	N_bkg
0.809	2.424+-0.229	114.0	2209.3

K-means clustering (without PCA)

- We are not sure about how many clusters is good to be informative, so we tried 15, 20 and 40 clusters

nClusters	Accuracy	Significance	N_sig	N_bkg
15	0.774	1.326+-0.030	2011.2	2.3E+06
20	0.778	1.436+-0.034	1787.5	1.6E+06
40	0.787	1.573+-0.036	1873.0	1.4E+06

K-means clustering (with PCA)

- PCA is used to reduce the number of variables by transforming our original variables to a new space but tried to keep most of the variance of our data
- We picked the number of variables that can keep 95% of the variance of our original data
- Then we do k-means clustering on the new variables and feed into BDT

nClusters	Accuracy	Significance	N_sig	N_bkg
15	0.773	1.309+-0.028	2186.7	2.8E+06
20	0.781	1.419+-0.031	2070.8	2.1E+06
40	0.783	1.452+-0.034	1865.9	1.7E+06

In conclusion

Method	Accuracy	Significance	N_sig	N_bkg
Original variables	0.809	2.424+-0.229	114.0	2209.3
15 Clusters	0.774	1.326+-0.030	2011.2	2.3E+06
15 Clusters(PCA)	0.773	1.309+-0.028	2186.7	2.8E+06
20 Clusters	0.778	1.436+-0.034	1787.5	1.6E+06
20 Clusters(PCA)	0.781	1.419+-0.031	2070.8	2.1E+06
40 Clusters	0.787	1.573+-0.036	1873.0	1.4E+06
40 Clusters(PCA)	0.783	1.452+-0.034	1865.9	1.7E+06