Applying k-means clustering in diHiggs signal/ background classification

12/18/2019

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

- We used variables list as:
  [‘hh_mass', 'h1_mass', 'h2_mass', 'hh_pt', 'h1_pt', 'h2_pt', 'deltaR(h1, h2)', 'deltaR(h1 jets)', 'deltaR(h2 jets)', 'deltaPhi(h1, h2)', 'deltaPhi(h1 jets)', 'deltaPhi(h2 jets)', 'met', 'met_phi', 'scalarHT', 'nJets', 'nBTags']

- 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-transformed variables, then feed the k-means-transformed variables into BDT

Tuning the parameters for xgboost

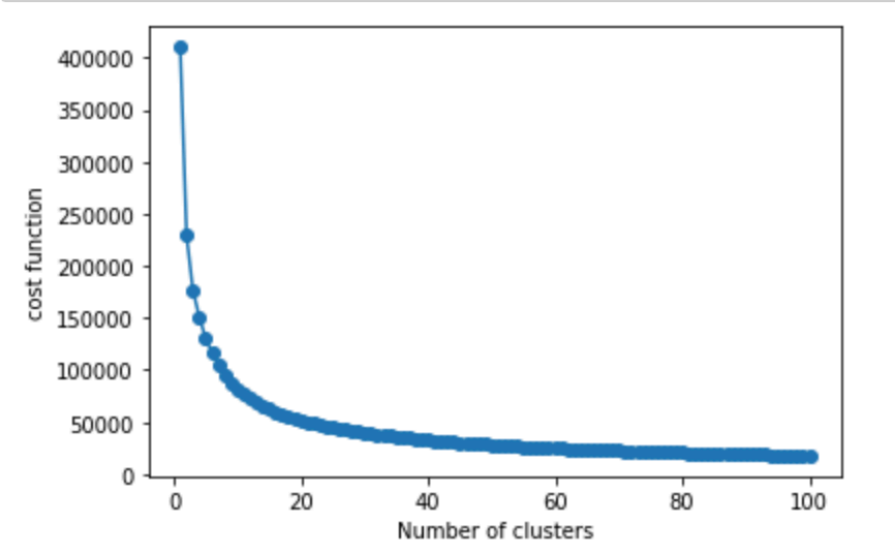
- Parameters in xgboost:
  - max_depth: Maximum depth of a tree
  - min_child_weight: minimum sum of instance weight (hessian) needed in a child
  - gamma: Minimum loss reduction required to make a further partition on a leaf node of the tree
  - subsample: Subsample ratio of the training instances
  - colsample_bytree: subsample ratio of columns when constructing each tree
  - alpha: L1 regularization term on weights
  - lambda: L2 regularization term on weights

- Run grid search in several steps:
  - Try different values of max_depth (in range [5,11]) and min_child_weight (in range [0,5])
  - Try different gamma (in range [0,1])
  - Try different subsample (in range [0.6,1]) and colsample_bytree (in range [0.6,1])
  - Try different alpha (in range [1e-05,100]) and lambda (in range [0,15])

- Then use the combination that gives us the best results on validation dataset

Choosing the number of clusters

- Normally we look at the plot of cost function vs. number of clusters to find an “elbow”, then use the corresponding number of clusters
- In this case, it’s hard to tell the position of elbow
- We tried 15, 20, 40 clusters, since there are something like an “elbow” in range [10,20], and we wonder whether more clusters gives us more information to do the classification

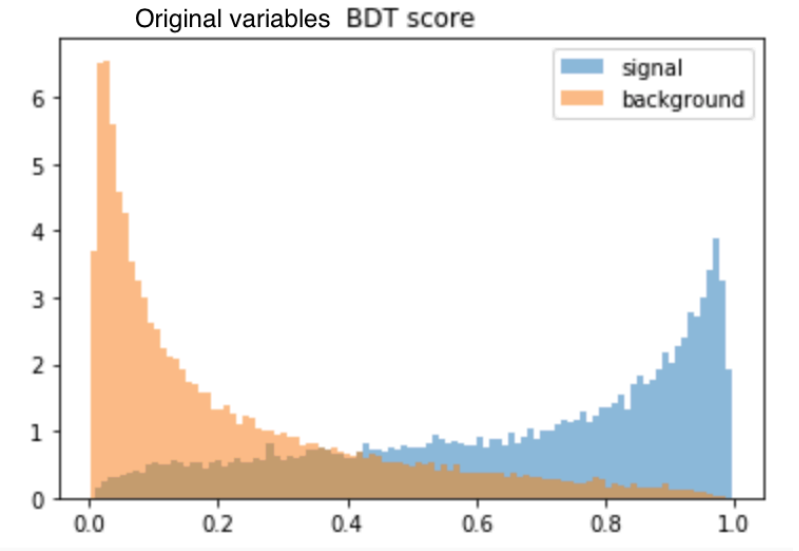


Doing PCA

- 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

Original variables

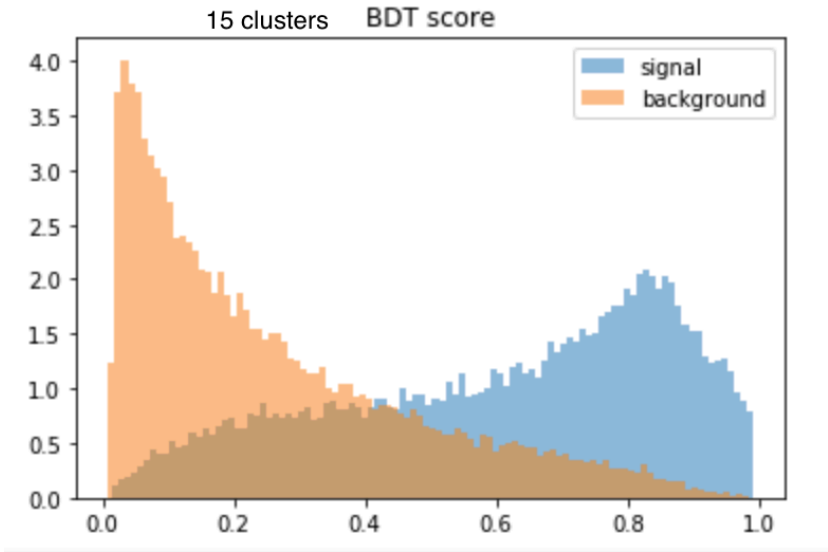
Parameter	Value
eta	0.1
n_estimator	5000
max_depth	11
min_child_weight	0
subsample	1.0
colsample_bytree	1.0
gamma	0.5
reg_alpha	1E-05
reg_lambda	9
scale_pos_weight	1



Accuracy	Significance	N_sig	N_bkg
0.809	2.16+-0.09	1230.6	3.2E+05

K-means clustering (15 clusters without PCA)

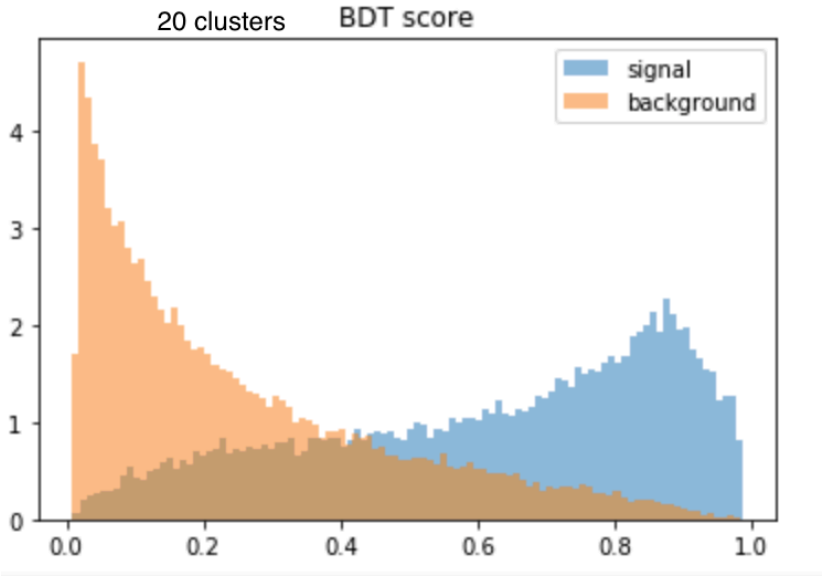
Parameter	Value
eta	0.1
n_estimator	5000
max_depth	11
min_child_weight	0
subsample	1.0
colsample_bytree	1.0
gamma	0.5
reg_alpha	1E-05
reg_lambda	9
scale_pos_weight	1



Accuracy	Significance	N_sig	N_bkg
0.774	1.326+-0.025	2011.2	2.3E+06

K-means clustering (20 clusters without PCA)

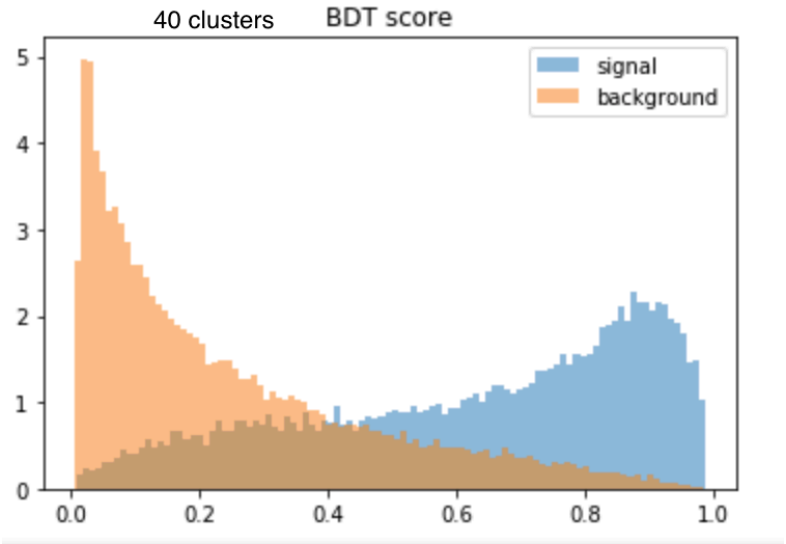
Parameter	Value
eta	0.1
n_estimator	5000
max_depth	11
min_child_weight	0
subsample	0.9
colsample_bytree	0.8
gamma	0
reg_alpha	1E-05
reg_lambda	9
scale_pos_weight	1



Accuracy	Significance	N_sig	N_bkg
0.778	1.436+-0.032	1787.5	1.5E+06

K-means clustering (40 clusters without PCA)

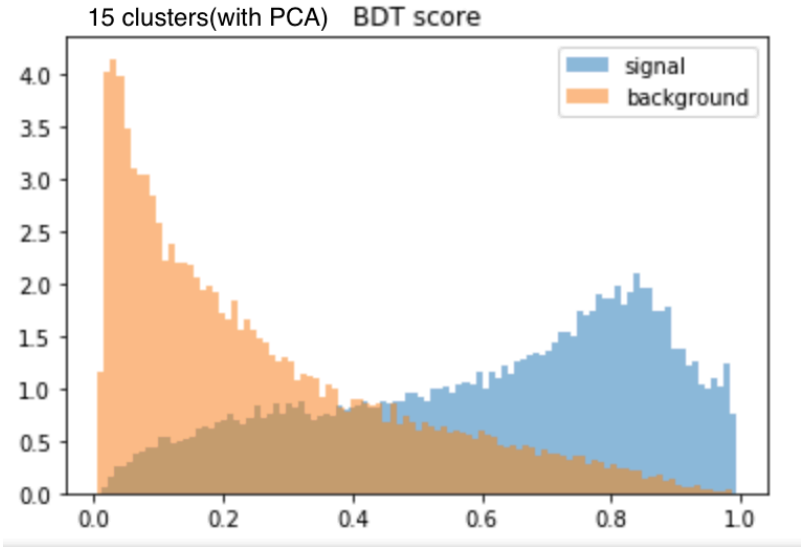
Parameter	Value
eta	0.1
n_estimator	5000
max_depth	11
min_child_weight	0
subsample	0.9
colsample_bytree	0.9
gamma	0.5
reg_alpha	1
reg_lambda	12
scale_pos_weight	1



Accuracy	Significance	N_sig	N_bkg
0.787	1.573+-0.036	1873.0	1.4E+06

K-means clustering (15 clusters without PCA)

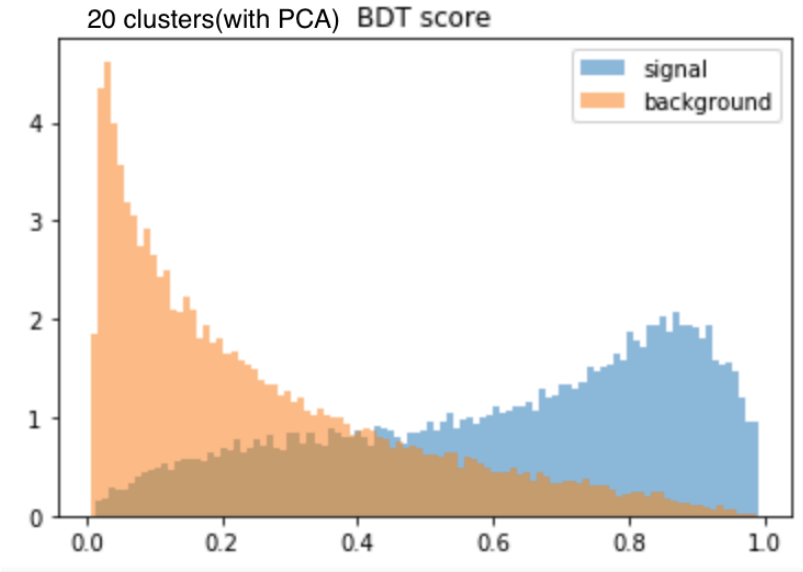
Parameter	Value
eta	0.1
n_estimator	5000
max_depth	11
min_child_weight	0
subsample	1.0
colsample_bytree	1.0
gamma	0
reg_alpha	1
reg_lambda	6
scale_pos_weight	1



Accuracy	Significance	N_sig	N_bkg
0.773	1.309+-0.023	2186.7	2.8E+06

K-means clustering (20 clusters without PCA)

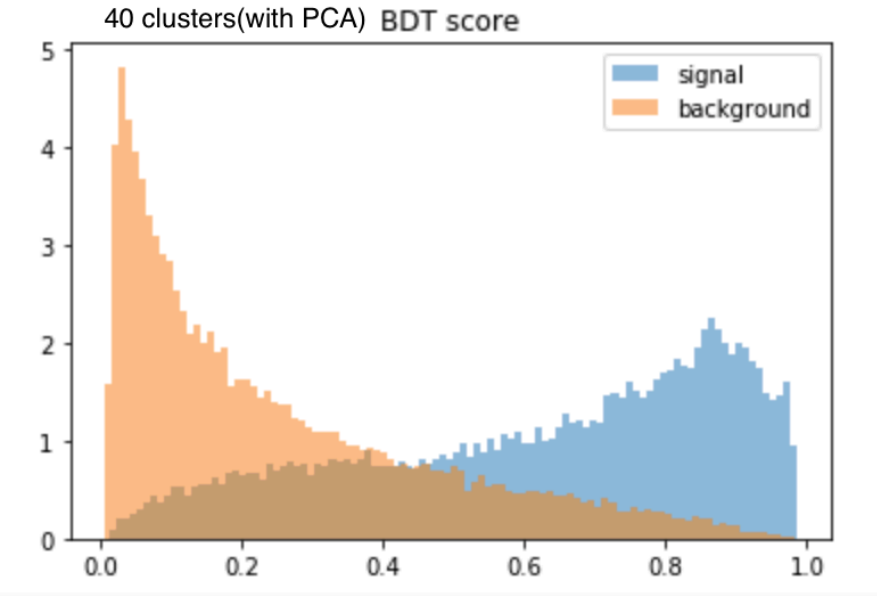
Parameter	Value
eta	0.1
n_estimator	5000
max_depth	11
min_child_weight	0
subsample	0.9
colsample_bytree	0.8
gamma	0
reg_alpha	1E-05
reg_lambda	9
scale_pos_weight	1



Accuracy	Significance	N_sig	N_bkg
0.778	1.436+-0.032	1787.5	1.5E+06

K-means clustering (40 clusters without PCA)

Parameter	Value
eta	0.1
n_estimator	5000
max_depth	11
min_child_weight	0
subsample	0.9
colsample_bytree	0.9
gamma	0.5
reg_alpha	1
reg_lambda	12
scale_pos_weight	1



Accuracy	Significance	N_sig	N_bkg
0.783	1.452+-0.031	1865.9	1.7E+06

In conclusion

Method	Accuracy	Significance	N_sig	N_bkg
Original variables	0.809	2.16+-0.09	1230.6	3.2E+05
15 Clusters	0.774	1.326+-0.025	2011.2	2.3E+06
15 Clusters(PCA)	0.773	1.309+-0.023	2186.7	2.8E+06
20 Clusters	0.778	1.436+-0.032	1787.5	1.5E+06
20 Clusters(PCA)	0.781	1.419+-0.028	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+-0.031	1865.9	1.7E+06