BIG DATA (X-LARGE)

WSDM – KKBox's Music Recommendation Challenge

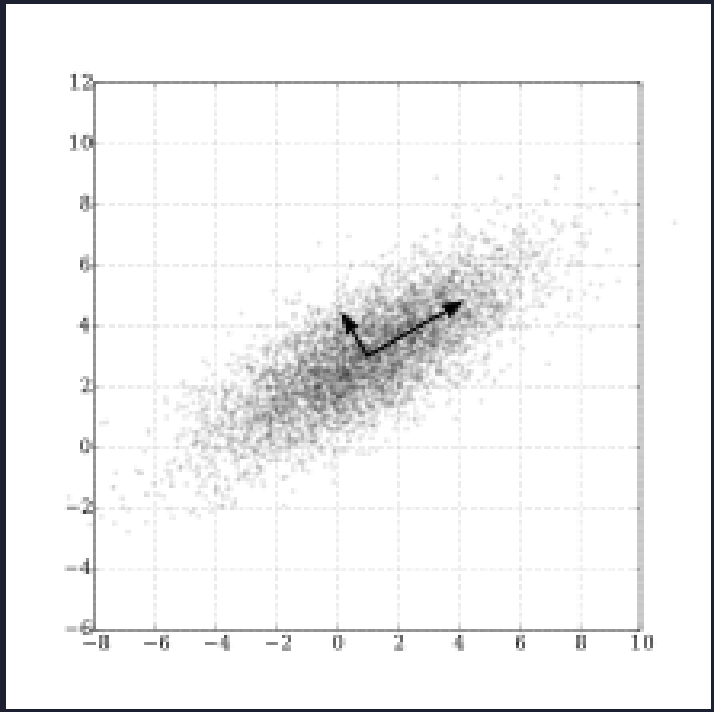
Modeling skills

P C A

PCA (Principal Component Analysis)

A statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components (or sometimes, principal modes of variation).

[picture 1] PCA of a multivariate Gaussian distribution centered at (1,3) with a standard deviation of 3 in roughly the (0.866, 0.5) direction and of 1 in the orthogonal direction. The vectors shown are the eigenvectors of the covariance matrix scaled by the square root of the corresponding eigenvalue, and shifted so their tails are at the mean.



[picture 1]

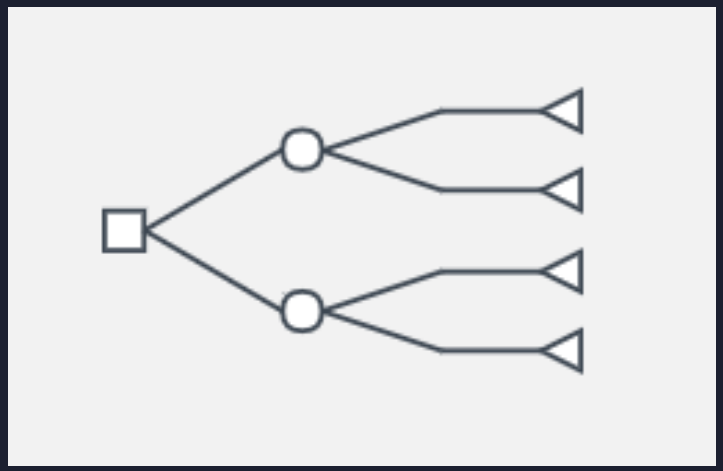
→이렇게 적용하면 되겠네~

Decision tree

Decision tree

A decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

[picture 2] A decision tree typically starts with a single node, which branches into possible outcomes. Each of those outcomes leads to additional nodes, which branch off into other possibilities. This gives it a treelike shape.



[picture 2] Decision tree diagram

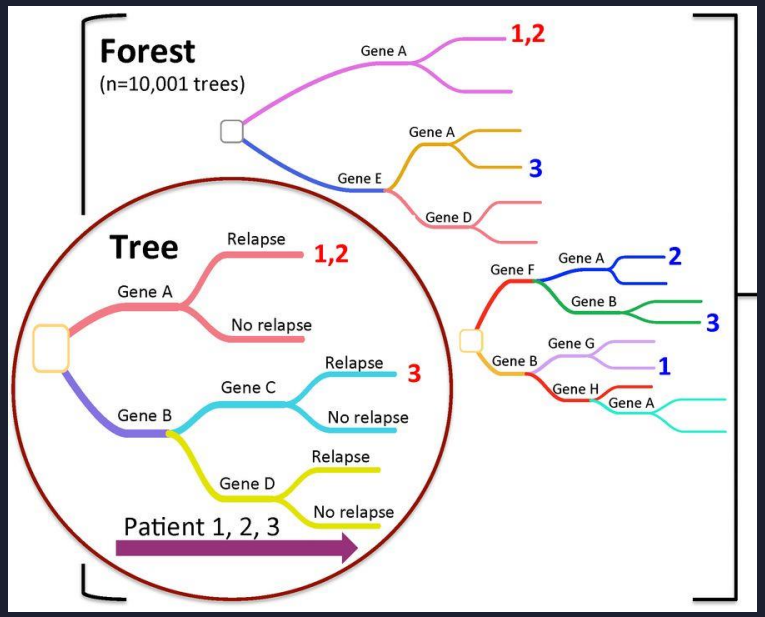
→이렇게 적용하면 되겠네~

Random Forest classifier

Random forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

[picture 3] A random forest classifier consists of multiple trees designed to increase the classification rate. For increased accuracy, sometimes multiple trees (decision trees) are used together in ensemble methods.



[picture 3]

→이렇게 적용하면 되겠네~

Ensemble (Boosting)

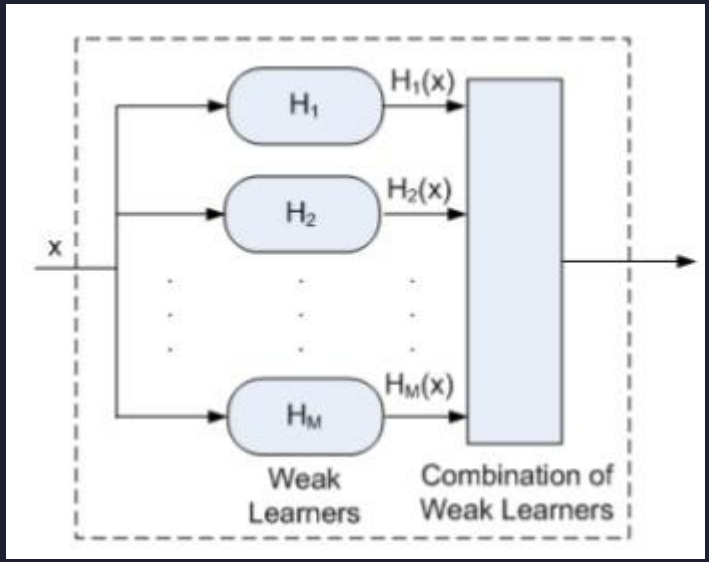
Ensemble learning

Ensemble learning use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone.

Boosting

Meta-algorithm. It approaches to combine several machine learning techniques into one predictive model in order to decrease the bias. Boosting is a two-step approach, where one first uses subsets of the original data to produce a series of averagely performing models and then "boosts" their performance by combining them together using a particular cost function.

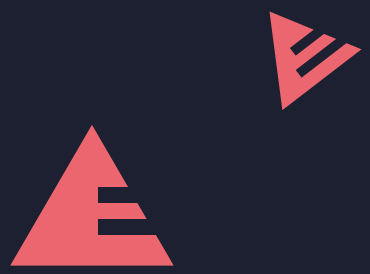
[picture 4] Boosted trees can be used for regression and classification trees.



[picture 4] Boosting

→이렇게 적용하면 되겠네~

Libraries



Python

Random Forest.. Boosting..?

R

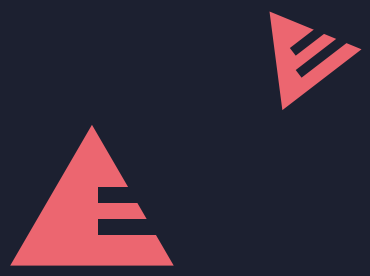
PCA..?

Hypotheses

Hypotheses

1.

Extra data



Extra data

조사하다 보니 더 있었으면 좋겠는 data!

THANKYOU

WSDM – KKBox's Music Recommendation Challenge