BIG DATA (X-LARGE)

WSDM – KKBox's Music Recommendation Challenge

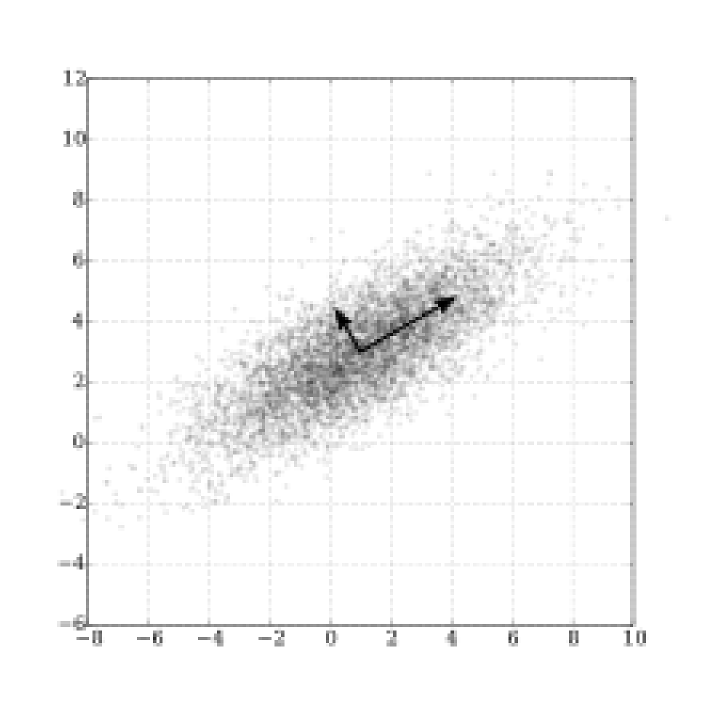
Modeling skills

PCA

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.



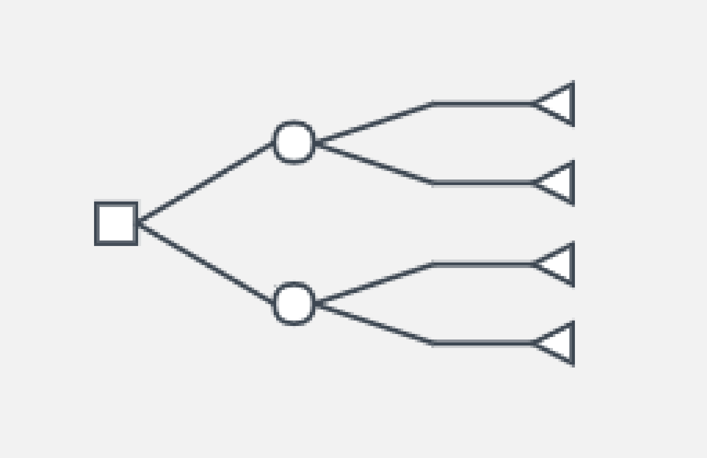
[picture 1]

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.



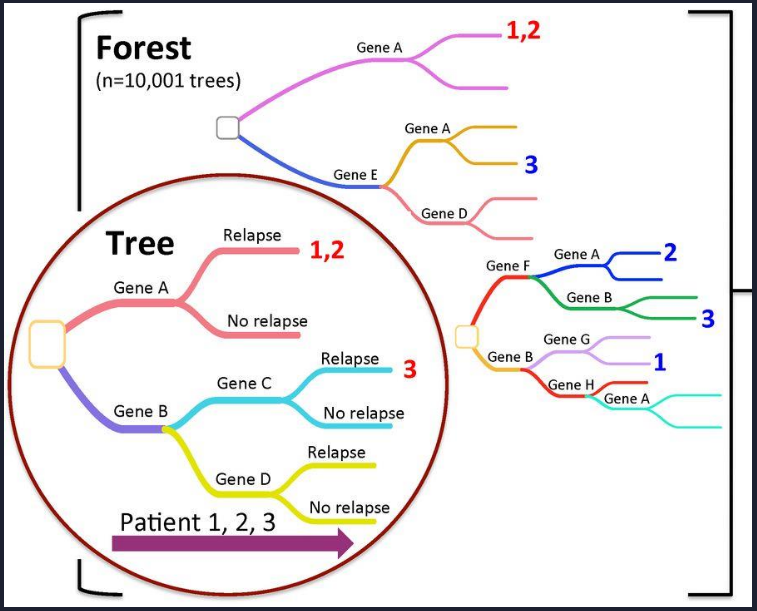
[picture 2] Decision tree diagram

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]

Ensemble (Boosting)

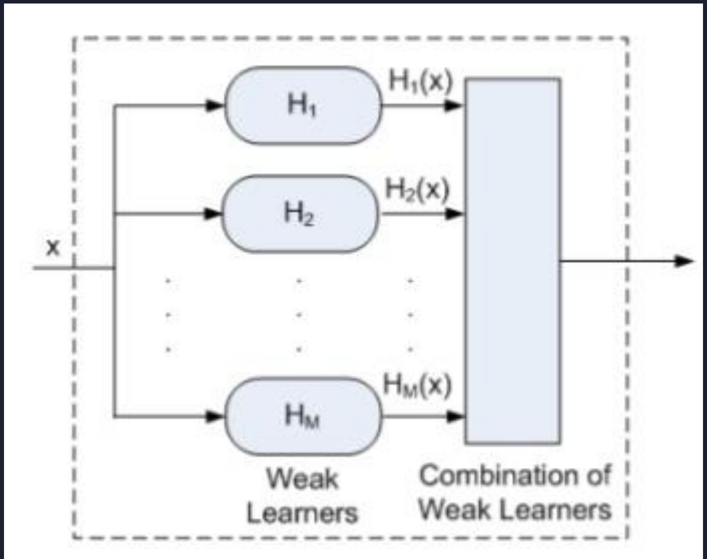
Ensemble learning

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

Implementation

PCA

PCA(Principal Component Analysis)

A method of reducing the variance of several variables into new variables that are made up of linear combinations of several highly correlated variables. In this case, the principal component means a direction vector having the largest variance of data in that direction.

PCA

Covariance

The general variance means the arithmetic balancing of the squared deviation of the sample data extracted from the population, ie the degree spreading from the mean. On the other hand, covariance in probability and statistics is a measure of the degree of correlation of two random variables. The covariance of X and y indicates how much the scattering of x and y is scattered with each other.

The variance-covariance matrix

- The variance-covariance matrix is a square matrix containing variance and covariance associated with several variables.

The diagonal elements of the matrix contain the variance of each variable, and the elements other than the diagonal contain covariances between all possible variable pairs.

PCA

Eigenvalue and eigenvector

When a matrix A is transformed into a linear transformation, a nonzero vector whose conversion result by the linear transformation A is multiplied by its own constant is called an eigenvector, and this constant eigenvalue is called an eigenvalue. That is, a non-zero column vector v satisfying Av =λV for A is defined as an eigenvector, and a constant λ Is defined as an eigenvalue.

PCA

Apply to PCA

Two components are used to decompose covariance matrices created to see the relationship of input data. The eigenvector is a vector representing a direction in which the variance of data is large, that is, the tendency of the original data indicates a direction or the like. The eigenvalue represents the magnitude of the variance. Therefore, the larger the eigenvalue is, the stronger the tendency of the data can be represented. So the higher the unique value, the better.

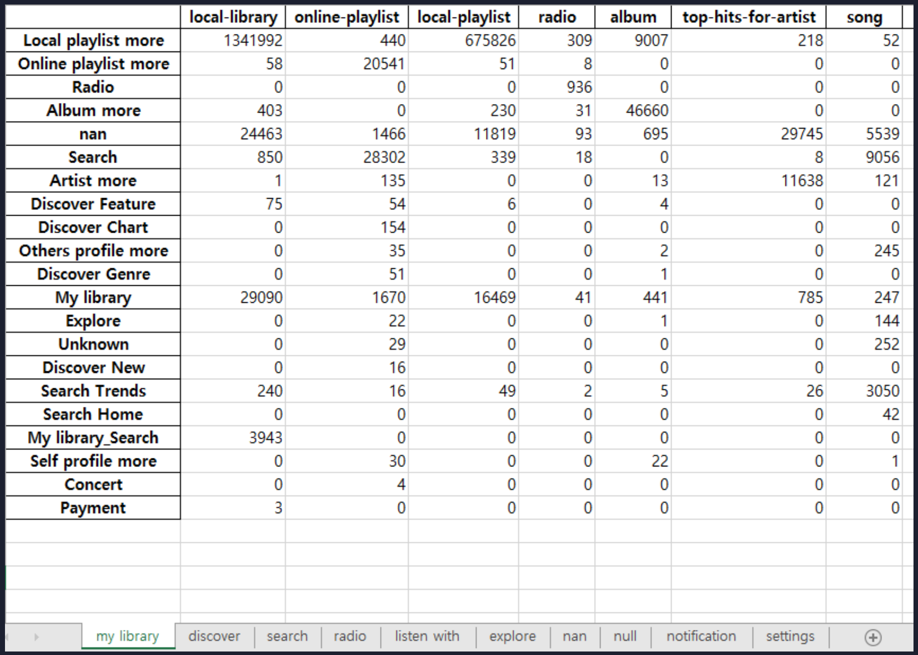
PCA

Overall process

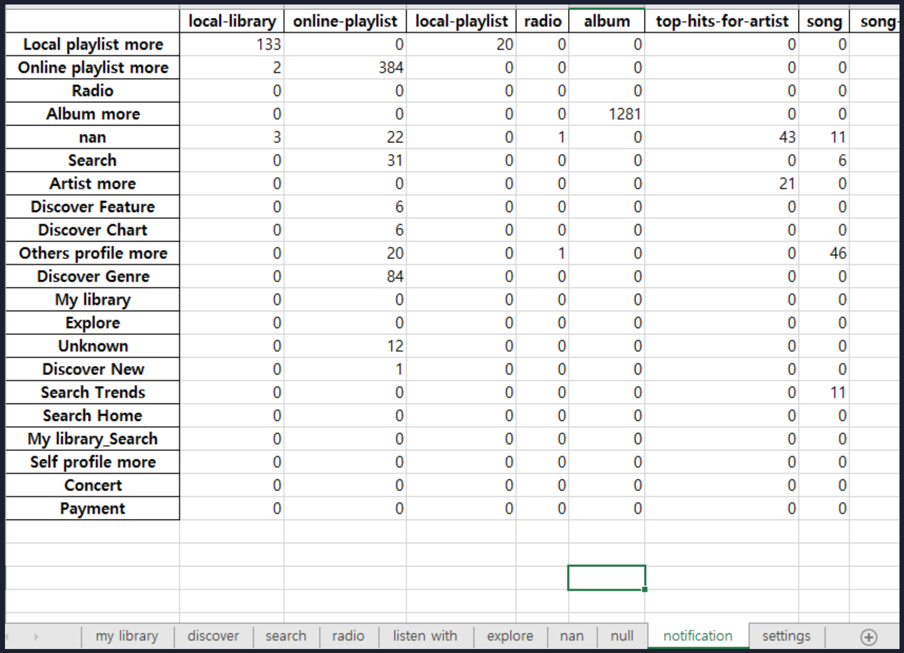
1. Standardize data
2. Create a covariance matrix
3. Decomposes the covariance matrix into eigenvalues and eigenvectors.
4. if there are k eigenvalues that are large enough to represent a feature and there are k pairs, the dimension of the data is reduced to k. At this time, k vectors are basis of the new data dimension.

P C A

Reduce demensions



	local-library	online-playlist	local-playlist	radio	album	top-hits-for-artist	song
Local playlist more	1341992	440	675826	309	9007	218	52
Online playlist more	58	20541	51	8	0	0	0
Radio	0	0	0	936	0	0	0
Album more	403	0	230	31	46660	0	0
nan	24463	1466	11819	93	695	29745	5539
Search	850	28302	339	18	0	8	9056
Artist more	1	135	0	0	13	11638	121
Discover Feature	75	54	6	0	4	0	0
Discover Chart	0	154	0	0	0	0	0
Others profile more	0	35	0	0	2	0	245
Discover Genre	0	51	0	0	1	0	0
My library	29090	1670	16469	41	441	785	247
Explore	0	22	0	0	1	0	144
Unknown	0	29	0	0	0	0	252
Discover New	0	16	0	0	0	0	0
Search Trends	240	16	49	2	5	26	3050
Search Home	0	0	0	0	0	0	42
My library_Search	3943	0	0	0	0	0	0
Self profile more	0	30	0	0	22	0	1
Concert	0	4	0	0	0	0	0
Payment	3	0	0	0	0	0	0



	local-library	online-playlist	local-playlist	radio	album	top-hits-for-artist	song	song-
Local playlist more	133	0	20	0	0	0	0	
Online playlist more	2	384	0	0	0	0	0	
Radio	0	0	0	0	0	0	0	
Album more	0	0	0	0	1281	0	0	
nan	3	22	0	1	0	43	11	
Search	0	31	0	0	0	0	6	
Artist more	0	0	0	0	0	21	0	
Discover Feature	0	6	0	0	0	0	0	
Discover Chart	0	6	0	0	0	0	0	
Others profile more	0	20	0	1	0	0	46	
Discover Genre	0	84	0	0	0	0	0	
My library	0	0	0	0	0	0	0	
Explore	0	0	0	0	0	0	0	
Unknown	0	12	0	0	0	0	0	
Discover New	0	1	0	0	0	0	0	
Search Trends	0	0	0	0	0	0	11	
Search Home	0	0	0	0	0	0	0	
My library_Search	0	0	0	0	0	0	0	
Self profile more	0	0	0	0	0	0	0	
Concert	0	0	0	0	0	0	0	
Payment	0	0	0	0	0	0	0	

PCA

Categorical principal components analysis(CATPCA)

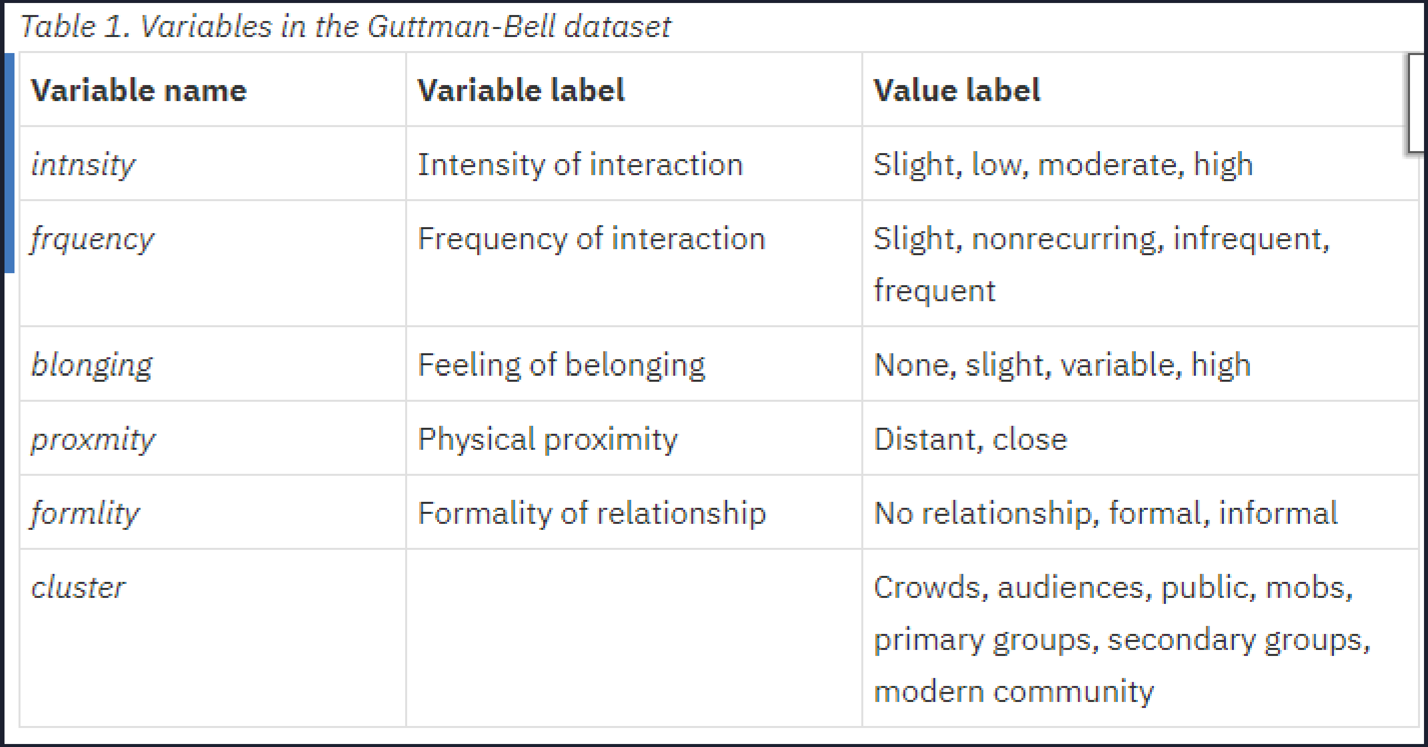


Table 1. Variables in the Guttman-Bell dataset

Variable name	Variable label	Value label
intnsity	Intensity of interaction	Slight, low, moderate, high
frquency	Frequency of interaction	Slight, nonrecurring, infrequent, frequent
blonging	Feeling of belonging	None, slight, variable, high
proxmity	Physical proximity	Distant, close
formlity	Formality of relationship	No relationship, formal, informal
cluster		Crowds, audiences, public, mobs, primary groups, secondary groups, modern community

Ensemble (Boosting)

Boosting: General

“General and effective method for producing an accurate prediction rule by combining rough” rules

A procedure to “seek for additional opinions” for incorrect classifications

AdaBoost, usually for Classification Problems
- Python Library: scikit-learn

Gradient Boosting, usually for Regression Problems
- Python Library: lightgbm

Ensemble (Boosting)

AdaBoost

Boosting technique for classification problems
Proposed by Freund, in chemistry fields

Versions

- Discrete AdaBoost
- Real AdaBoost
- Gentle AdaBoost
- AdaBoost.MH for multiclass classification

Ensemble (Boosting)

Discrete AdaBoost

```
Repeat T times with initial weight $1/N$,
    Sample M records from training set using the weight.
    Obtain a weak learner(ex: shallow decision tree) from samples.
    For each records, $err=0$ if correct, 1 if incorrect.
    $E$ = Sum all errors of the records, multiplied by record weight.

    We have Confidence Index $C=\log(1-E)-\log(E)$
    Multiply weights of incorrect records by $\exp(C)=(1-E)/E$
    Mormalize the weights.

Let the ensemble of obtained learners classify the test data.
```

Ensemble (Boosting)

CART classification vs. AdaBoost [1]

A: Classification and regression tree (CART)
Making a overfitted decision tree first, and pruning it

B: Adaboost with 3 iterations

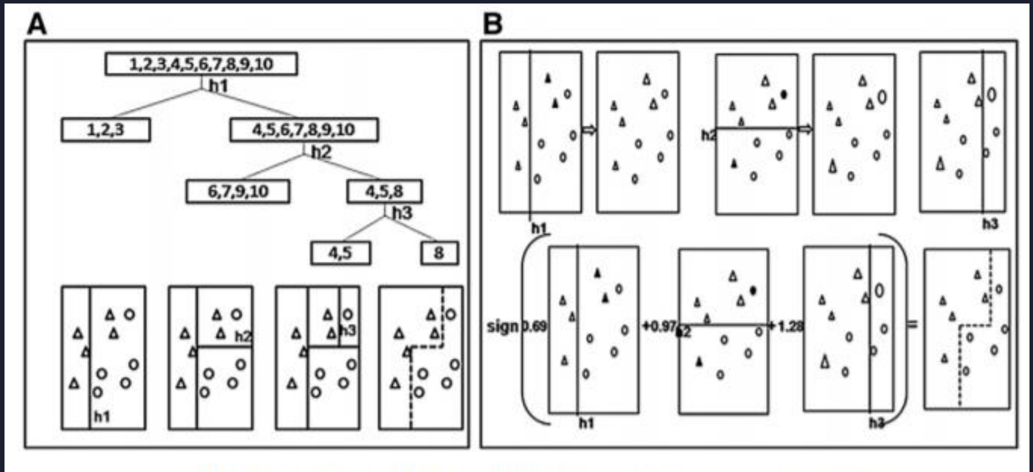


Fig. 4. (A) The overall classification process of CART. (B) The overall iteration process for AdaBoost.

Ensemble (Boosting)

Issues on AdaBoost

Finding the optimal number of iterations is important.

- Low T does not reflect the solution space will.
- High T leads to overfitting.

Issue on misclassified(noisy) data

- As boosting increases weights of incorrect records, misclassified data exerts significant influence on the models as the iterations goes by.
- Workarounds: BrownBoost (giving up the noisy data), …….

Reference

Reference

[Pages]	[Reference]
p.3	https://en.wikipedia.org/wiki/Principal_component_analysis
p.4	https://en.wikipedia.org/wiki/Decision_tree https://www.lucidchart.com/pages/decision-tree
p.5	https://en.wikipedia.org/wiki/Random_forest https://www.lucidchart.com/pages/decision-tree
p.6	https://en.wikipedia.org/wiki/Ensemble_learning https://stats.stackexchange.com/questions/18891/bagging-boosting-and-stacking-in-machine-learning
8-14p	http://yamalab.tistory.com/32
15-19p	D. Cao et al., "The boosting: A new idea of building models", Chemometrics and Intelligent Laboratory Systems, vol. 100, no. 1, pp. 1-11, 2010.

THANKYOU

WSDM – KKBox's Music Recommendation Challenge