Machine Learning

Homework3

By Mina Erfan

Dataset Info

- The dataset was created in 2014 by the University of Nottinghan, Ningbo, China
- The dataset was built from a collection of 1059 tracks covering 33 countries/area.
- The geographical location of origin was manually collected the information from the CD sleeve notes. The country of origin was determined by the artist's or artists' main country/area of residence.
- The position of each country's capital city (or the province of the area) have been taken by latitude and longitude as the absolute point of origin.
- The program MARSYAS[1] was used to extract audio features from the wave files.

Dataset Info

Data Set Characteristics:	Multivariate	**Number of Instances:**	1059	**Area:**	N/A
Attribute Characteristics:	Real	**Number of Attributes:**	68	**Date Donated**	2014-10-18
Associated Tasks:	Classification, Regression	**Missing Values?**	N/A	**Number of Web Hits:**	106939

Exp1: Estimators

Apply **Histogram estimator**, **Naïve Estimator**, and **Kernel Estimator** on your dataset for all possible values for **h** and report results. Then apply **K-NN** for values = 1, 3, 5, 7, 11 and compare the result with Exp1 HomeWork1.

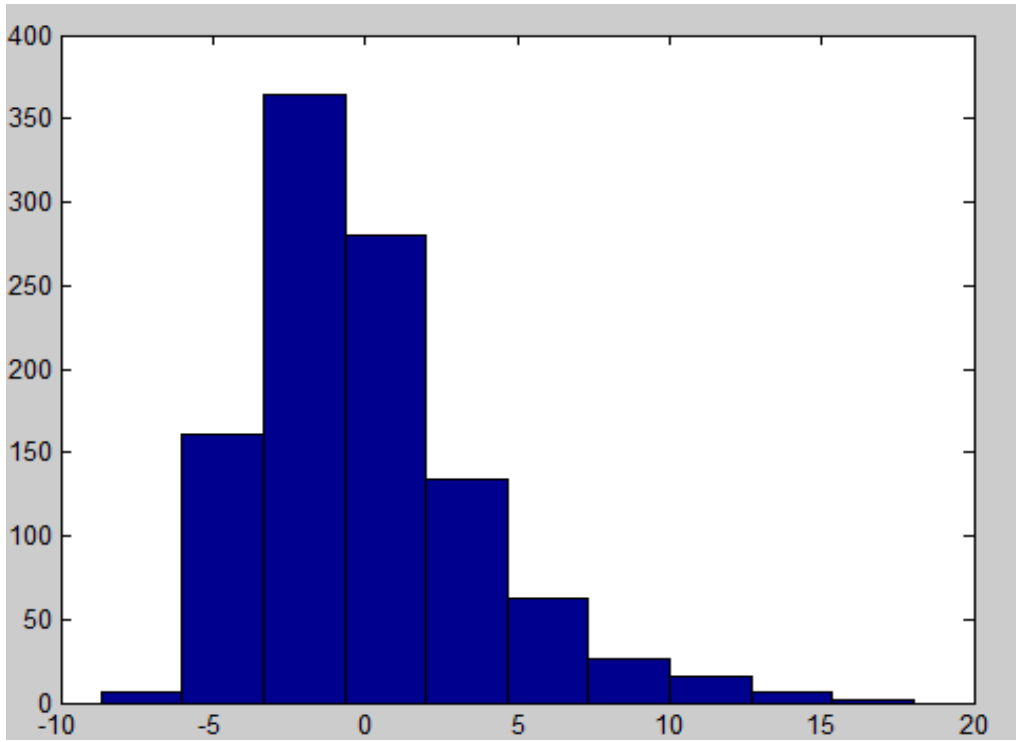
Histogram

- Hist Function in matlab is used for histogram estimator

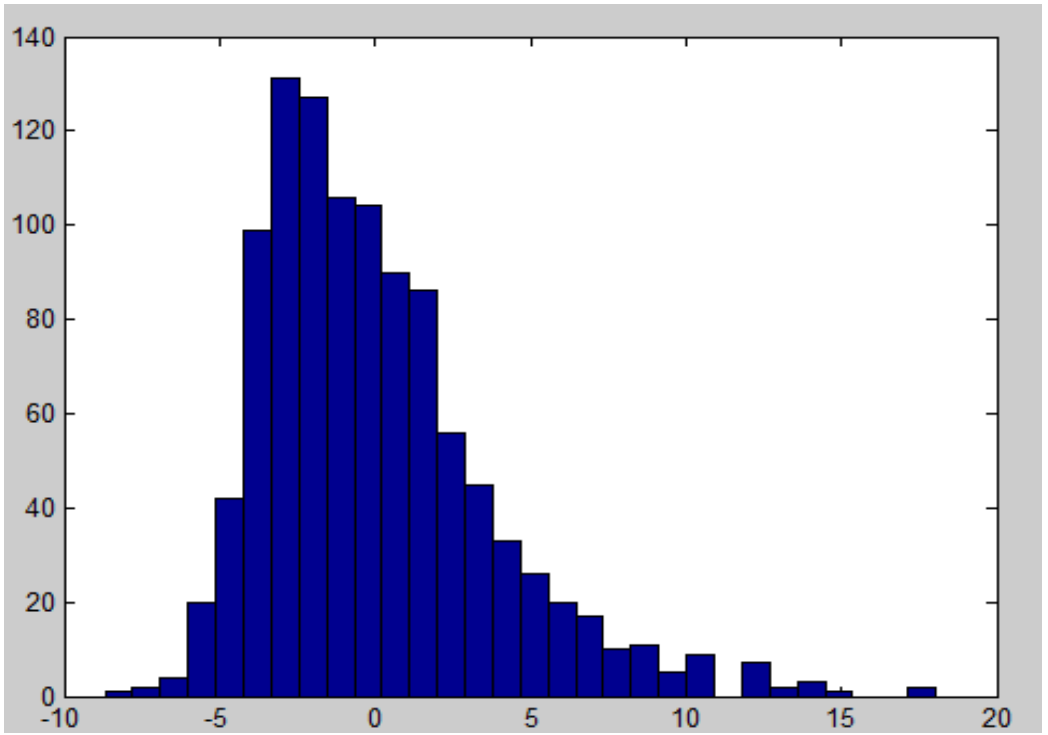
hist (data , nbins)

- inputs
  - data : dataset
  - nbin : number of bins

Histogram

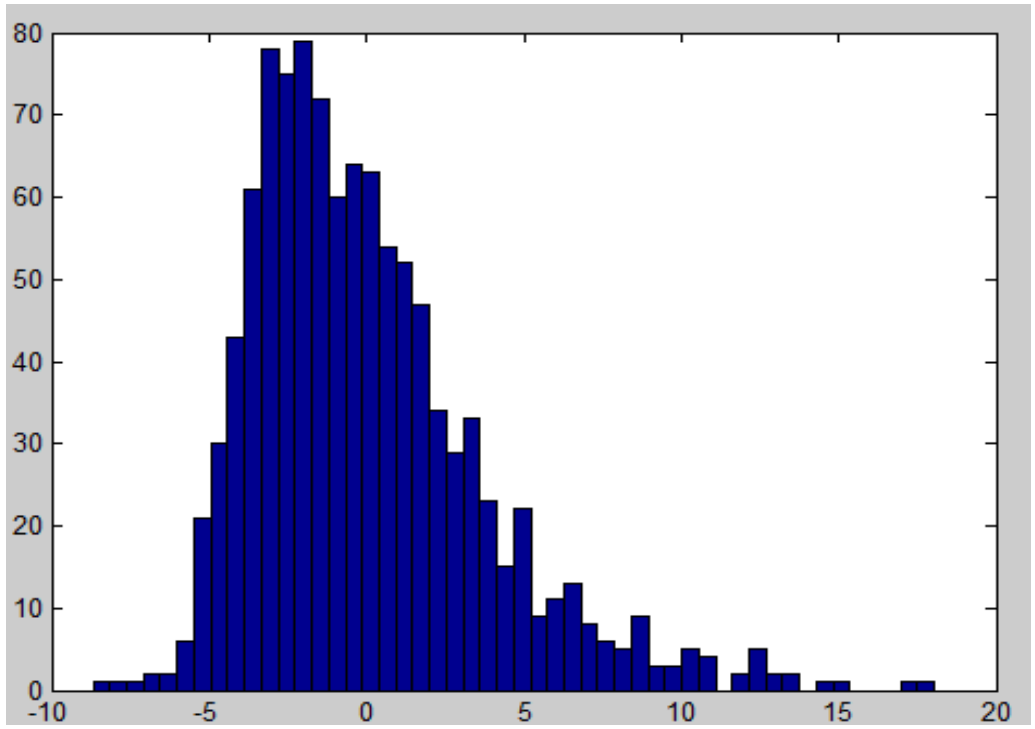


Histogram for 10 bin

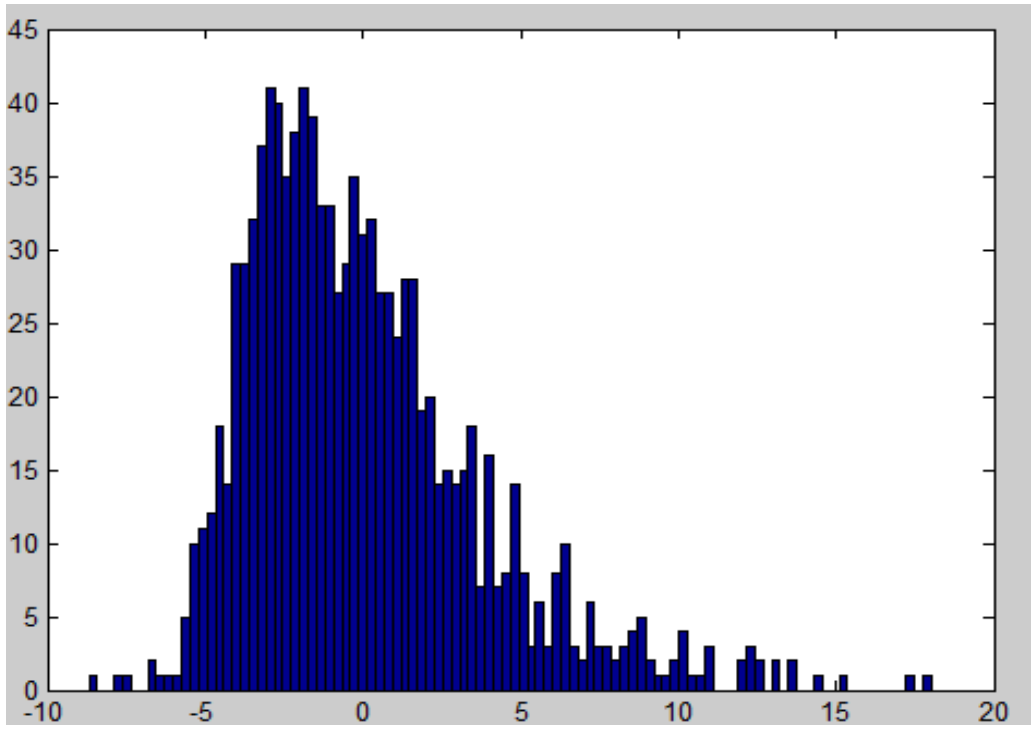


Histogram for 30 bin

Histogram



Histogram for 50 bin

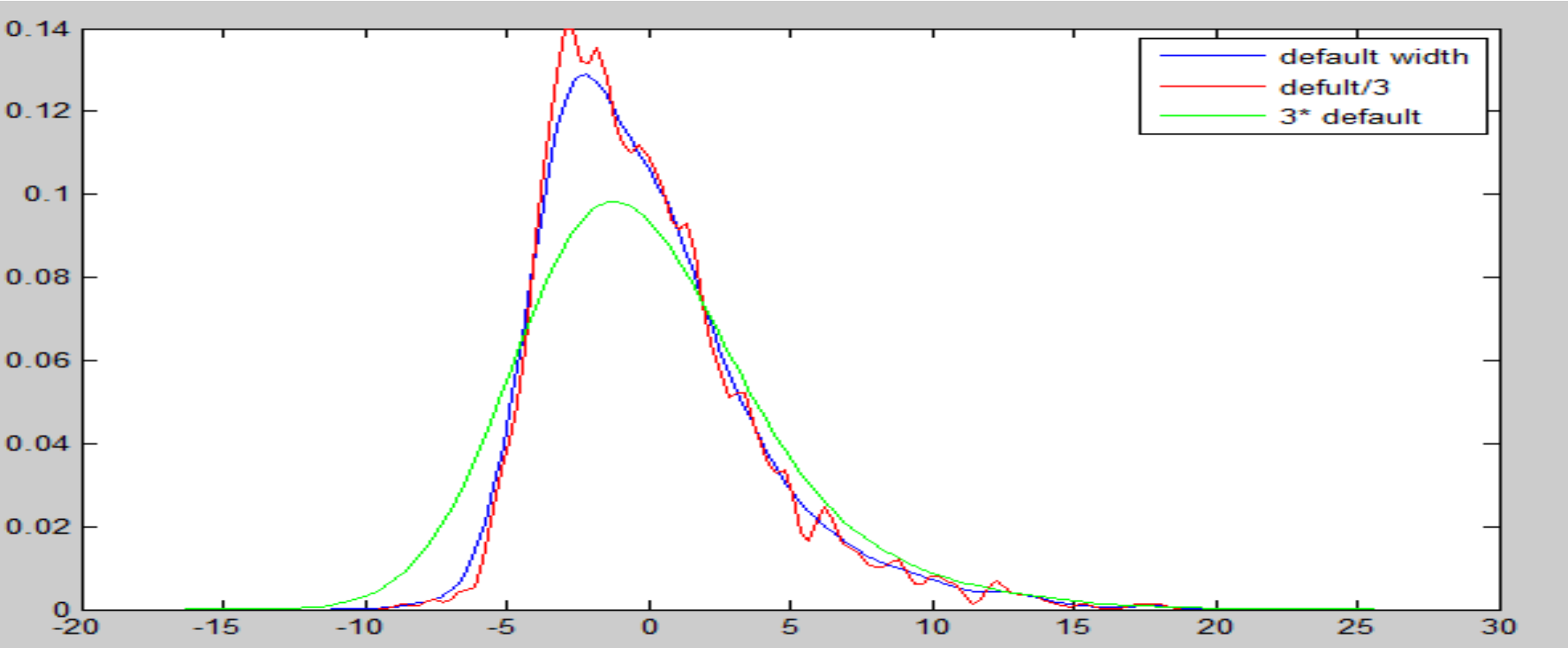


histogram for 100 bin

Kernel Estimator

برای اعمال روش kernel estimator از تابع ksdensity برای تخمین چگالی نمونه ها در نرم افزار متل اسفاده شده است.

Kernel Estimator



KNN

- Weka is used for KNN classification

weka.classifiers.lazy.IBK

KNN

	TP Rate	FP Rate	precision	Recall	F-Measure
KNN k=1	0.407	0.023	0.396	0.407	0.391
KNN k=3	0.367	0.027	0.394	0.367	0.355
KNN k=5	0.387	0.027	0.392	0.387	0.364
KNN k=7	0.383	0.027	0.388	0.383	0.359
KNN k=11	0.376	0.028	0.345	0.376	0.341

Exp2: Decision Tree

Use univariate c4.5/j.48 to retrieve a set of learned rules and present them. Then report all unused features with their ranks from most significant to the least.

.

Decision Tree

- Weka is used for classification with decision tree (unprune=true)

weka.classifiers.trees.j48

Decision Tree

266	Number of leaves
523	Size of the decision tree
7	Unused features

Correctly classified	TP Rate	FP Rate	precision	Recall	F-Measure
29.17	0.292	0.027	0.302	0.292	0.294

Rank of attributes	Attributes
1	attr36
2	attr38, attr53
3	attr54, attr4, attr30, attr55
4	attr40, attr2, attr42, attr62, attr3
5	attr41, attr8, attr37, attr5, attr1, attr6, attr33, attr57, attr25, attr7
6	attr60 , attr14, attr21, attr26, attr29, attr11, attr31
7	attr22, attr43, attr24, attr15, attr57, attr57, attr35, attr65, attr28, attr9, attr47
8	attr39, attr66, attr20, attr16, attr44, attr58, attr54, attr63, attr64, attr50
9	attr18, attr10, attr12, attr52, attr13
10	attr23, attr59
11	attr17, attr68, attr45, attr49
12	attr19

Exp3: Decision Tree-Pruning

Prune the decision tree for different values of θ_t and report the size of the new tree. Then derive results from pruned tree, and compare those with previous results.

Decision Tree-Pruning

- Weka is used for KNN classification (unprune=false)

weka.classifiers.lazy.IBK

	Tree size	TP Rate	FP Rate	precision	Recall	F-Measure
θ = 0.15	515	0.294	0.027	0.303	0.294	0.296
θ = 0.25	519	0.293	0.027	0.302	0.293	0.295
θ = 0.35	519	0.293	0.027	0.302	0.293	0.295
θ = 0.45	519	0.293	0.027	0.302	0.293	0.295
θ = 0.55	523	0.292	0.027	0.302	0.292	0.294
θ = 0.65	523	0.292	0.027	0.302	0.292	0.294
θ = 0.75	523	0.292	0.027	0.302	0.292	0.294
θ = 0.85	523	0.292	0.027	0.302	0.292	0.294
θ = 0.95	523	0.292	0.027	0.302	0.292	0.294
unproun	523	0.292	0.027	0.302	0.292	0.294

Thank You