Machine Learning Homework1

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:Bayes Classifier

Bayesian Classifier:

- Apply Bayesian classifier for all different settings of given covariance matrix and compare results:
- All classes reside on a single diagonal COV
- Each class on a separate diagonal COV
- All classes on a single common COV
- Each class on a dedicated COV

Bayes Classifier using matlab

- Training set: 70% of samples
- Testing set: 30% of samples
- Using classify function in matlab

Classify(test , train , group , type)

- ✓ Test :testing set
- ✓ Train: training set
- ✓ Group : number of classes
- ✓ Type: Linear/ diaglinear/ quadratic/dia quadratic

Results

	TP Rate	FP Rate	Precision	Recall	F-Measure
linear	0.38	0.02	0.32	0.38	0.33
diaglinear	0.31	0.02	0.29	0.31	0.27
diagquadratic	0.29	0.02	0.31	0.29	0.28
quadratic	----	----	----	----	----

Quadratic : error > The covariance matrix of each group in TRAINING must be positive definite. -> number of samples in each class in fewer than number of features

Exp2: Missing values

Predicting the Missing Value:

Consider the strongest property. Randomly delete values of such property at 5%, 10%, 15%, 20% and 25% rate respectively. Then Learn a Regressor and use it to estimate deleted values. Reapply Bayesian classifier by replacing missing values with estimated values. Compare results.

Creating missing value

1. Finding vital features with Weka:
   - Weka/explorer/select attributes/ choose InfoGainAttributeEval
   - "InfoGainAttributeEval" function found att53 as the vital feature
2. Omitting 5% of values for the vital features
3. Using linear regression to predict missing values
   - Weka/explorer/classify/choose/functions/LinearRegression

Learning a Regressor

This is the learned formula. Since we have 68 features, we used 67 features for learning attr53

attr53 = 0.4475 * attr1 + -0.6931 * attr2 + 0.0732 * attr4 + -0.0372 * attr6 + -0.0344 * attr7 + -0.019 * attr11 + -1.3553 * attr18 + 1.3247 * attr19 + -0.0396 * attr21 + 0.0266 * attr26 + -0.0335 * attr27 + 0.0328 * attr29 + -0.0468 * attr30 + -0.0264 * attr31 + -0.249 * attr35 + 0.7137 * attr36 + -0.1287 * attr37 + -0.0644 * attr38 + -0.0625 * attr39 + -0.0587 * attr40 + -0.0452 * attr42 + 0.0254 * attr46 + 0.0269 * attr47 + -0.0316 * attr49 + 0.7299 * attr52 + 0.0511 * attr54 + 0.0232 * attr55 + 0.0855 * attr56 + 0.0516 * attr57 + -0.0397 * attr60 + 0.0221 * attr61 + 0.035 * attr65 + 0.0401 * attr68 + -0.0063

NaiveBayes Results

Correct classified	F-Measure	Recall	Precision	FP Rate	TP Rate	
33.33	0.324	0.333	0.351	0.024	0.33	**Main Set**
33.14	0.322	0.331	0.349	0.024	0.33	**5%_miss**
15.96	0.142	0.16	0.149	0.029	0.16	**10%_miss**
16.05	0.143	0.161	0.15	0.029	0.161	**5%_miss1**
16.05	0.143	0.161	0.15	0.029	0.161	**20%_miss**
16.05	0.143	0.161	0.15	0.029	0.161	**25%_miss**

Thank You