MSPR Exam Miniproject

Analysis on UCI Wine dataset

Mattia Paterna

Sound and Music Computing
Aalborg University, Copenhagen

Outline

1. Problem formulation
   - General information
   - Questions
2. Preliminaries
   - Exploring data using variance measure
   - Exploring data using correlation measure
3. Supervised learning
   - Training and testing subsets
   - Cross-validation
   - Classification on several subsets
4. Unsupervised learning
   - Clustering
   - Principal Component Analysis
5. Feature Selection
   - Assessments
   - Filter
   - Wrapper

Outline

1. Problem formulation
   - General information
   - Questions
2. Preliminaries
   - Exploring data using variance measure
   - Exploring data using correlation measure
3. Supervised learning
   - Training and testing subsets
   - Cross-validation
   - Classification on several subsets
4. Unsupervised learning
   - Clustering
   - Principal Component Analysis
5. Feature Selection
   - Assessments
   - Filter
   - Wrapper

General information
UCI Wine Data Set

- UCI wine is a multivariate dataset.
- It contains results of a chemical analysis of wines derived from three different cultivars.
- 178 observation, each of one has 13 features.
- No missing values

Outline

1. Problem formulation
   - General information
   - Questions
2. Preliminaries
   - Exploring data using variance measure
   - Exploring data using correlation measure
3. Supervised learning
   - Training and testing subsets
   - Cross-validation
   - Classification on several subsets
4. Unsupervised learning
   - Clustering
   - Principal Component Analysis
5. Feature Selection
   - Assessments
   - Filter
   - Wrapper

Questions

- which features are most relevant to draw differences between cultivars?
- is there any correlation between any features?
- how do wines differ when deriving from different cultivars?
- is it possible to state precisely the belonging to a cultivars without any additional knowledge?

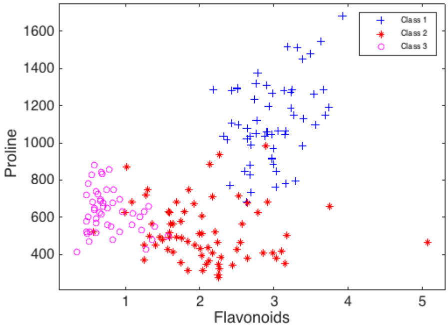
Outline

1. Problem formulation
   - General information
   - Questions
2. Preliminaries
   - Exploring data using variance measure
   - Exploring data using correlation measure
3. Supervised learning
   - Training and testing subsets
   - Cross-validation
   - Classification on several subsets
4. Unsupervised learning
   - Clustering
   - Principal Component Analysis
5. Feature Selection
   - Assessments
   - Filter
   - Wrapper

Assessments

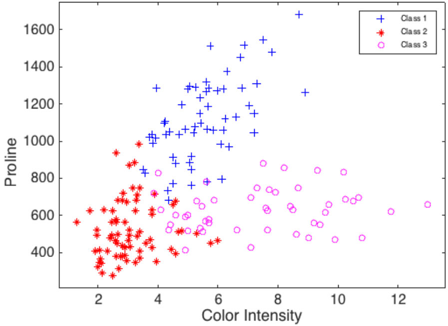
- get the variance measure for each feature
- plot a group scatter - feature with the highest variance against each features

Feature 13 over 7



- best separation among classes
- class 1 is well spaced and divided
- little overlap between class 2 and 3

Feature 13 over 10



- good separation among classes
- class 1 and 3 are well spaced and divided
- little overlap among all classes

Outline

1 Problem formulation
- General information
- Questions

2 Preliminaries
- Exploring data using variance measure
- Exploring data using correlation measure

3 Supervised learning
- Training and testing subsets
- Cross-validation
- Classification on several subsets

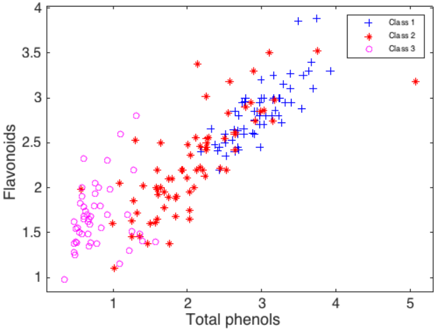
4 Unsupervised learning
- Clustering
- Principal Component Analysis

5 Feature Selection
- Assessments
- Filter
- Wrapper

Assessments

- get the correlation matrix
- plot a group scatter - feature with the highest correlation

Feature 7 over 6



- classes are not well separated
- large overlap between class 1 and 2

Outline

1. Problem formulation
   - General information
   - Questions
2. Preliminaries
   - Exploring data using variance measure
   - Exploring data using correlation measure
3. Supervised learning
   - Training and testing subsets
   - Cross-validation
   - Classification on several subsets
4. Unsupervised learning
   - Clustering
   - Principal Component Analysis
5. Feature Selection
   - Assessments
   - Filter
   - Wrapper

Assessments

Criterion

Data set has been divided into two subsets

Percentages

- training set: 70 %
- training set: 30 %

Choice method

data has been split randomly selecting shuffled indexes

Outline

1. Problem formulation
   - General information
   - Questions
2. Preliminaries
   - Exploring data using variance measure
   - Exploring data using correlation measure
3. Supervised learning
   - Training and testing subsets
   - Cross-validation
   - Classification on several subsets
4. Unsupervised learning
   - Clustering
   - Principal Component Analysis
5. Feature Selection
   - Assessments
   - Filter
   - Wrapper

Cross-validation I

- using PRTools function
- *leave-one-out* in size of S

Cross-validation II

Classifier	Accuracy	Errors
NMSC	0.95	6
LDC	0.99	1
QDC	0.98	3
KNNC	0.65	44
SVM	0.45	69

Table: Accuracy and total errors number using cross-validation (run I)

- LDC and QDC best accuracy and low errors score
- non parametric classifiers work bad

Cross-validation III

Classifier	Accuracy	Errors
NMSC	0.98	3
LDC	1	0
QDC	0.984	2
KNNC	0.69	38
SVM	0.06	117

Table: Accuracy and total errors number using cross-validation (run II)

- LDC best accuracy
- SVM completely failed

Cross-validation IV

- To sum up:
  - parametric classifiers seem work properly
  - high accuracy values ($> 90\,\%$)
  - non-parametric classifiers not successful
  - SVM lowest accuracy in both trials

Outline

1. Problem formulation
   - General information
   - Questions
2. Preliminaries
   - Exploring data using variance measure
   - Exploring data using correlation measure
3. Supervised learning
   - Training and testing subsets
   - Cross-validation
   - Classification on several subsets
4. Unsupervised learning
   - Clustering
   - Principal Component Analysis
5. Feature Selection
   - Assessments
   - Filter
   - Wrapper

Subsets

1. two highest *variance* features
2. features with highest *correlation* value
3. PCA projection on two highest eigenvalues

Subset I (highest variance)

Classifier	Accuracy	Errors
NMSC	0.85	8
LDC	0.85	8
QDC	0.79	11
KNNC	0.62	20
SVM	0.81	10

Table: Accuracy and total errors number for subset I

- more errors compared to cross-validation
- not much difference between parametric classifiers and non-parametric ones

Subset II (highest correlation)

Classifier	Accuracy	Errors
NMSC	0.75	13
LDC	0.75	13
QDC	0.78	12
KNNC	0.81	10
SVM	0.79	11

Table: Accuracy and total errors number for subset II

- more errors compared to cross-validation
- non parametric classifiers work *better*
- support vector machine most accurate

Subset III (PCA projection)

Classifier	Accuracy	Errors
NMSC	0.6	21
LDC	0.62	20
QDC	0.6	21
KNNC	0.62	20
SVM	0.66	18

Table: Accuracy and total errors number for subset III

- LDC parametric classifier with best accuracy
- no relevant differences between parametric and non-parametric classifiers
- support vector machine most accurate

Subset V

- To sum up:
  - subset I (highest variance features) has highest accuracy values
  - generally less difference between parametric and non-parametric classifiers
  - SVM highest accuracy classifier for subset II and III

Outline

1. Problem formulation
   - General information
   - Questions
2. Preliminaries
   - Exploring data using variance measure
   - Exploring data using correlation measure
3. Supervised learning
   - Training and testing subsets
   - Cross-validation
   - Classification on several subsets
4. Unsupervised learning
   - Clustering
   - Principal Component Analysis
5. Feature Selection
   - Assessments
   - Filter
   - Wrapper

Applied clustering method

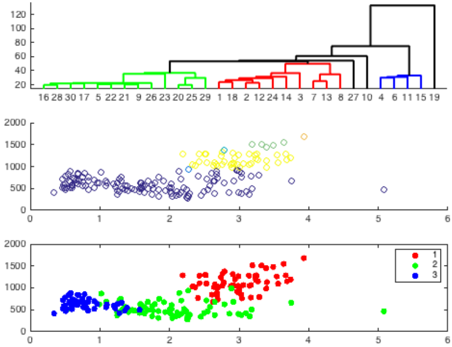
1. agglomerative clustering with single linkage
2. *k-means*
3. evaluation using *variance-ratio* criterion
4. Gaussian Mixture Model

Agglomerative clustering I

Clusters/Classes

Cluster	1	2	3
I	13	69	48
II	0	1	0
III	1	0	0
IV	5	0	0
V	1	0	0
VI	39	1	0

Table: Comparison between cluster assignment and true labels



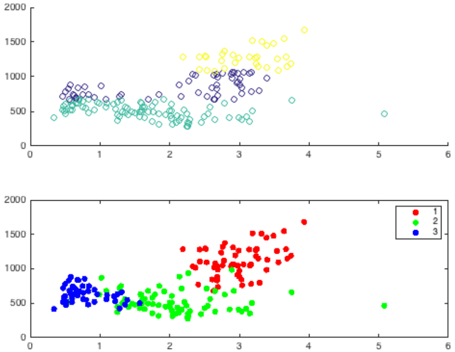
Agglomerative clustering II

- optimal treshold = 52
- classes 2 and 3 are not distinguished, class 1 is well defined though
- 5 elements don't belong to any cluster
- class 1 is split between clusters I and VI

K-means clustering I

Clusters/Classes			
Cluster	1	2	3
I	28	13	15
II	0	58	33
III	31	0	0

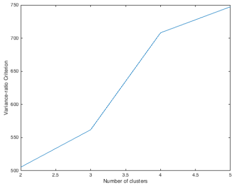
Table: Comparison between cluster assignment and true labels

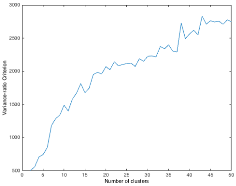


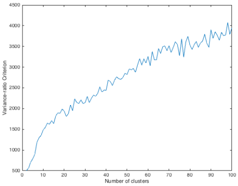
K-means clustering II

- number of clusters to be found = 3
- classes 2 and 3 are mostly combined in cluster II
- class 1 is split between clusters I and III
- none of the classes seem well separate

Variance-ration Criterion







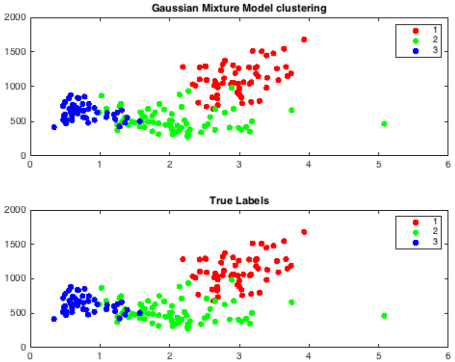
- optimal K increases when increasing overall inspected K
- elbow rule is not appliable
- it's not possible define an optimal number of cluster in the dataset

Gaussian Mixture Model I

Cluster	1	2	3
I	59	1	0
II	0	70	0
III	0	0	48

Clusters/Classes

Table: Comparison between cluster assignment and true labels



Gaussian Mixture Model II

- number of Gaussian to be found $= 3$
- always convergence
- total errors number $= 1$
- clusters fit classes separation at their best

Clustering

- To sum up:
  - generally clustering doesn't work well
  - variance-ratio criterion doesn't find optimal K
  - GMM however performs an excellent clustering

Outline

1 Problem formulation
- General information
- Questions

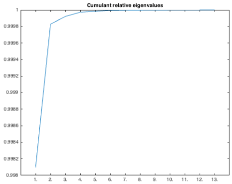
2 Preliminaries
- Exploring data using variance measure
- Exploring data using correlation measure

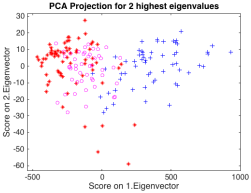
3 Supervised learning
- Training and testing subsets
- Cross-validation
- Classification on several subsets

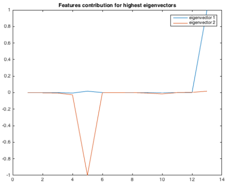
4 Unsupervised learning
- Clustering
- Principal Component Analysis

5 Feature Selection
- Assessments
- Filter
- Wrapper

PCA on covariance matrix I

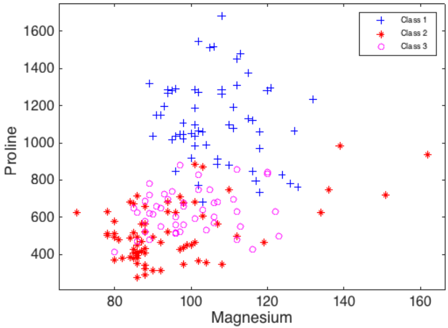






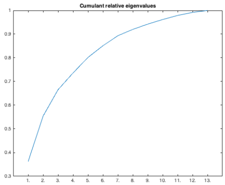
- reduction to two highest eigenvalues
- preserved variance: $> 99\,\%$
- good classes separation
- eigenvector contribution from a single feature

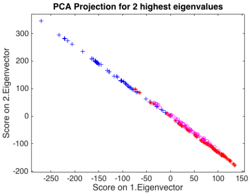
PCA on covariance matrix II

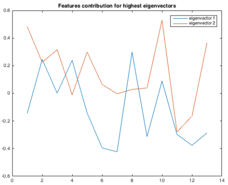


- plot using detected features from PCA
  - result is not clear
  - large overlap between class 2 and 3

PCA on correlation matrix I

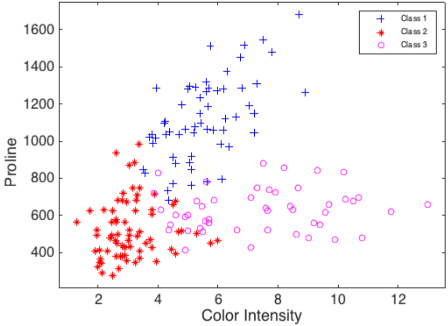






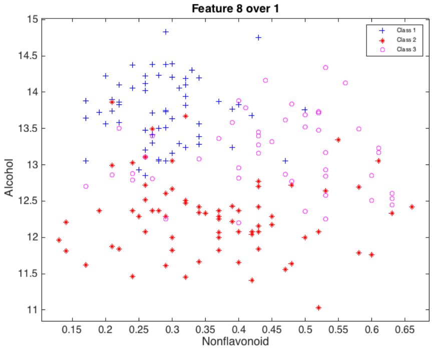
- reduction to two highest eigenvalues
- preserved variance: 55 %
- no useful classes separation
- main features contribution for first eigenvector: 7,8
- main features contribution for second eigenvector: 10,1

PCA on correlation matrix II



- plot using feature 7 over 10
  - same features as using variance measure
  - good separation between classes

PCA on correlation matrix III



- plot using feature 8 over 1
  - result is much confused
  - great overlap among all classes

Principal Component Analysis

- To sum up:
  - variance is well preserved using covariance matrix
  - correlation matrix doesn't give meaningful results
  - features 7,10 and 13 seem the most relevant to describe the data set

Outline

1. Problem formulation
   - General information
   - Questions
2. Preliminaries
   - Exploring data using variance measure
   - Exploring data using correlation measure
3. Supervised learning
   - Training and testing subsets
   - Cross-validation
   - Classification on several subsets
4. Unsupervised learning
   - Clustering
   - Principal Component Analysis
5. Feature Selection
   - Assessments
   - Filter
   - Wrapper

Assessments

Goodness of Subset Criterion

Filter and wrapper

Data set split technique for wrapper

- training set: 70 %
- validation set: 20 % of training set
- training set: 30 %

Outline

1 Problem formulation
- General information
- Questions

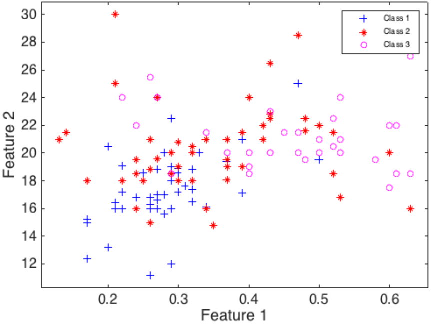
2 Preliminaries
- Exploring data using variance measure
- Exploring data using correlation measure

3 Supervised learning
- Training and testing subsets
- Cross-validation
- Classification on several subsets

4 Unsupervised learning
- Clustering
- Principal Component Analysis

5 Feature Selection
- Assessments
- Filter
- Wrapper

Filter I



- correlate each feature with true labels
  - highest correlation for features 8, 4
  - great overlap between classes

Filter II

Classifier	Accuracy	Errors
NMSC	0.61	49
LDC	0.61	49
QDC	0.58	52
KNNC	0.46	67
SVM	0.52	60

Table: Accuracy and total errors number (subset with feature 8,4)

- cross-validation using reduced data set
- all classifiers are not accurate
- far from accuracy using the whole data set

Outline

1. Problem formulation
   - General information
   - Questions
2. Preliminaries
   - Exploring data using variance measure
   - Exploring data using correlation measure
3. Supervised learning
   - Training and testing subsets
   - Cross-validation
   - Classification on several subsets
4. Unsupervised learning
   - Clustering
   - Principal Component Analysis
5. Feature Selection
   - Assessments
   - Filter
   - Wrapper

Wrapper I
General information

- feature selection using **forward selection** scheme
- features to be selected $= 2$
- predictor trained on training data
- features selection based on best accuracy when tested on validation set

Wrapper II

Classifier	Feat	Accuracy	Errors
NMSC	7,1	0,87	7
LDC	7,1	0,87	7
QDC	7,1	0,89	6
KNNC	13,11	0,66	18
SVM	13,5	0,83	9

Table: Features selected and accuracy for each classifier using forward selection

- good overall accuracy
- no relevant differences parametric and non-parametric classifiers
- exception: knnc

Feature Selection

- To sum up:
  - filter criterion gives not relevant features
  - classification over that subset performs quite bad
  - wrapper criterion is better
  - features selected are the same detected in PCA and in preliminaries

Conclusion I

- cross-validation gives the best accuracy for parametric classifiers, but works quite bad with non-parametric ones
- subsets doesn't provide better accuracy, but in most cases SVM is the most accurate classifier
- clustering seems not work properly, except Gaussian Mixture Model
- generally, unsupervised learning is not excellent in get the belonging to the exact class

Conclusion II

- PCA on covariance matrix preserves a high variance value, but classification on that score is however the worst one
- Features underlined from both PCA and features selection provide a good 2-D representation of the entire dataset
- but, we should guess that classification on the entire data set is the most successful way to analyze it - probably because of its small dimensions