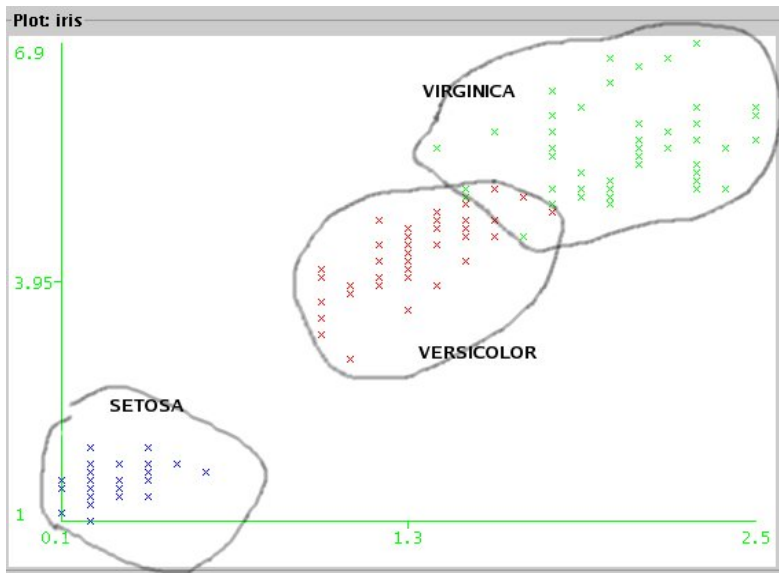


Iris

Differentiate among three species of flowers (Iris)

- 4 continuous attributes
- Attributes: Measures of characteristics of the flowers
- 150 instances
- 3 classes
- 96 % accuracy for supervised learning

Iris



Iris - Expectation/maximization

- We use the EM algorithm looking for 3 clusters
- Clusters are relatively clear, accuracy is a little bit lower

```

0  1  2  <-- assigned to cluster
0 50  0  | Iris-setosa
50  0  0  | Iris-versicolor
14  0 36  | Iris-virginica

```

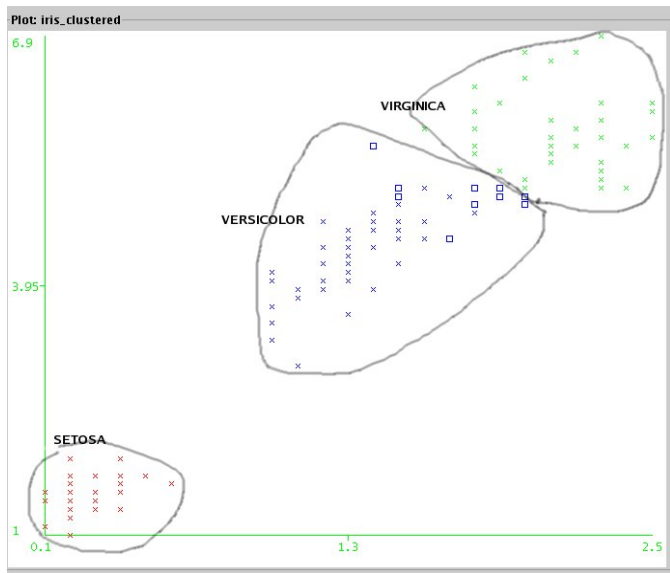
```

Cluster 0 <-- Iris-versicolor
Cluster 1 <-- Iris-setosa
Cluster 2 <-- Iris-virginica

```

Incorrectly clustered instances : 14.0 9.3333 %

Iris - Expectation/maximization



Iris - K-means

- K-means algorithm looking of 3 clusters
- Clusters are relatively clear, but cluster intersection affects prediction

```

0  1  2  <-- assigned to cluster
0 50  0  | Iris-setosa
47  0  3  | Iris-versicolor
14  0 36  | Iris-virginica

```

```

Cluster 0 <-- Iris-versicolor
Cluster 1 <-- Iris-setosa
Cluster 2 <-- Iris-virginica

```

Incorrectly clustered instances : 17.0 11.3333 %

Voting Records

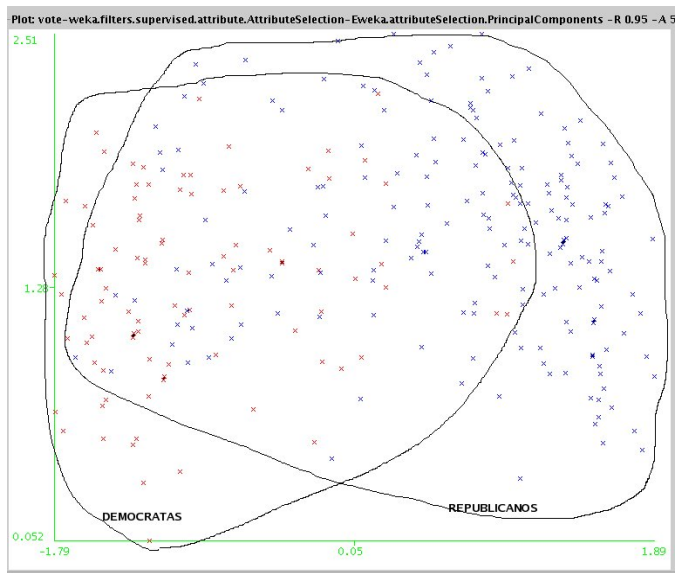
Classify US senators by their voting

- 16 binary attributes
- Attributes: Vote of the senator to different proposals (budget, immigration, taxes, military aid, ...)
- 435 instances
- 2 classes
- 96.3% accuracy for supervised learning
- Visualization of the data set is very difficult (binary attributes!)

Voting Records - PCA

- PCA is used to obtain a new set of attributes
- The data set does not hold the conditions to apply PCA (non gaussian data)
- The 3 first components explain the 60 % of the variance (the first one explains 45 %, All are needed to reach 95 % of variance)

Voting records - PCA



Voting Records - Expectation-maximization

- EM algorithm is applied looking for 2 clusters
- Clusters are not very clear, the error is large

```
0 1 <-- assigned to cluster
44 223 | democrat
159 9 | republican
```

```
Cluster 0 <-- republican
```

```
Cluster 1 <-- democrat
```

```
Incorrectly clustered instances : 53.0 12.1839 %
```

Voting Records - K-means

- K-means algorithm is applied looking for 2 clusters
- The error is larger because of the intersection among clusters

```
0 1 <-- assigned to cluster
50 217 | democrat
157 11 | republican
```

```
Cluster 0 <-- republican
```

```
Cluster 1 <-- democrat
```

```
Incorrectly clustered instances : 61.0 14.023 %
```

Mushroom

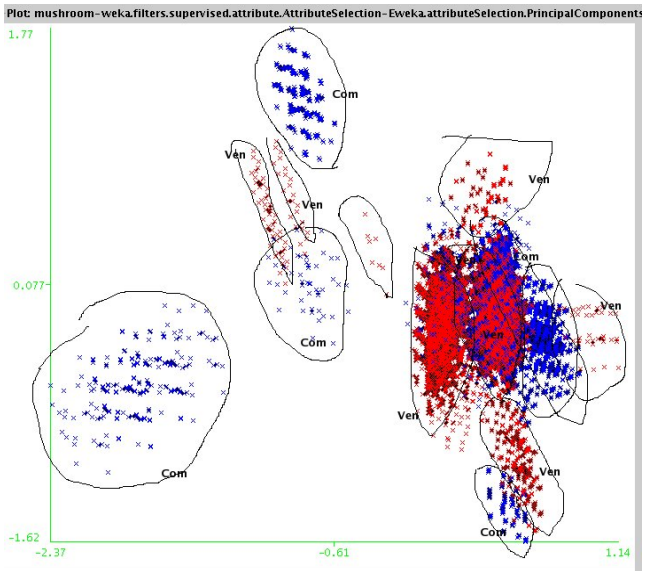
Distinguish between poisonous and edible mushrooms

- 22 Attributes binary and nominal
- Attributes: Visible characteristics of the mushrooms
- About 8000 instances
- 2 classes
- 100 % accuracy for supervised learning
- Visualization using the original attributes is difficult (binary and nominal attributes!)

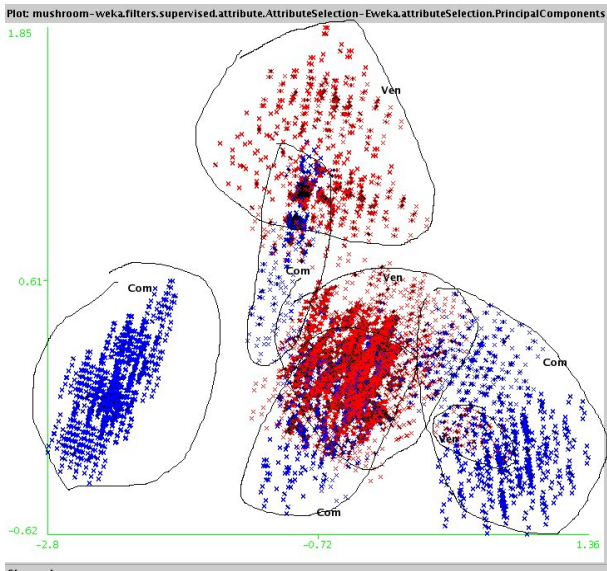
Mushroom - PCA

- PCA is used to obtain a new set of attributes
- The data set does not hold the conditions to apply PCA (non gaussian data)
- The first 10 components explain only 50% of the variance. Are necessary all to explain 95% of the variance (PCA has 59 components).

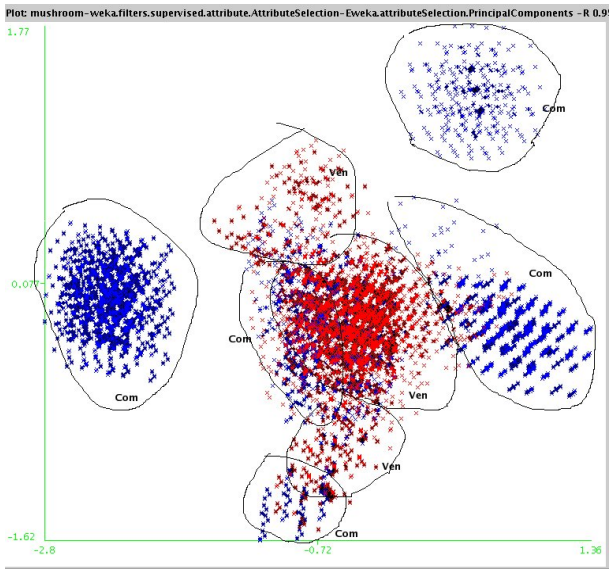
Mushroom - PCA



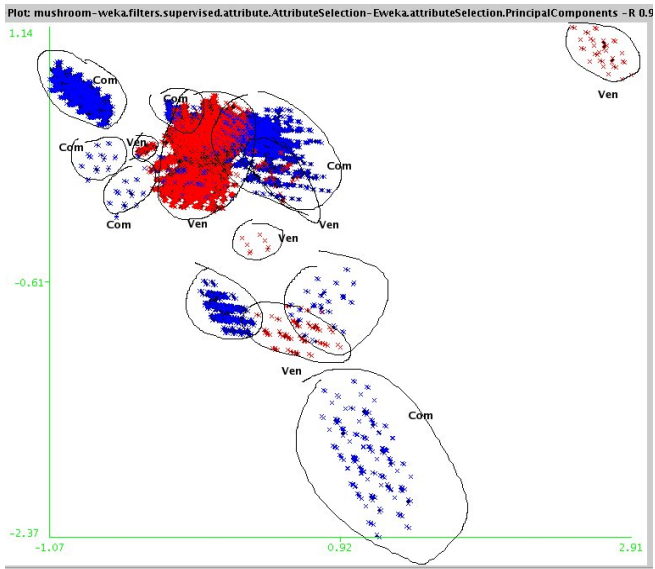
Mushroom - PCA



Mushroom - PCA



Mushroom - PCA



Mushroom - Expectation/maximization

- EM algorithm is applied looking for 2 clusters
- Clusters are not very clear, the error is large
- Probably it is more interesting to look for more clusters and analyze them (the data set has more structure than the supervised labels show)

```

0      1  <-- assigned to cluster
4208   0 | e
836 3080 | p

```

Cluster 0 <-- e

Cluster 1 <-- p

Incorrectly clustered instances : 836.0 10.2905 %

Mushroom - Expectation/maximization + attribute selection

- We are cheating :-)
- A wrapper using decision trees is used to find the relevant attributes (5 relevant attributes)
- EM algorithm is applied looking for 2 clusters

```
0    1  <-- assigned to cluster
4000 208 | e
528 3388 | p
```

```
Cluster 0 <-- e
```

```
Cluster 1 <-- p
```

```
Incorrectly clustered instances : 736.0    9.0596 %
```

Mushroom - K-means

- K-means algorithm is applied looking for 2 clusters
- The result is awful, intersection among classes is large, there is no good partition of the data

```
0    1  <-- assigned to cluster
```

```
1234 2974 | e
```

```
2093 1823 | p
```

```
Cluster 0 <-- p
```

```
Cluster 1 <-- e
```

```
Incorrectly clustered instances: 3057.0  37.6292 %
```

Clustering in image processing



Clustering in image processing

