

# Heart disease

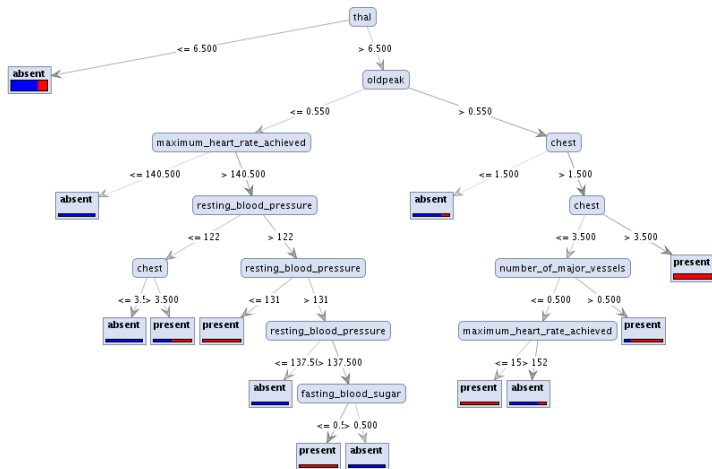
## Diagnose of heart disease

- 13 Attributes (6 continuous, 1 discrete, 3 binaries, 4 qualitative)
- Attributes: age, sex, chest pain type, resting blood pressure, serum cholestorol, fasting blood sugar, ...
- 270 instances
- 2 classes
- Validation: 10 fold cross validation

# Heart disease: Decision tree

- Parameters: Post pruning
- Size of the tree: 25 nodes (13 leaves)
- Accuracy: 70.3 %
- Unpruned: 70 nodes (44 leaves), accuracy 73.3 %
- Bagging: accuracy 81.1 %
- Boosting: accuracy 76.6 %

## Heart disease: Decision tree



## Heart disease: K-nearest neighbour

- Parameters: 1, 3, 5, 7, 9, 11, 13 neighbours
- Model: All 270 instances
- Accuracy: 55.5 %, 63.3 %, 66.3 %, 70 %, 65.5 %, 65.5 %, 65.2 %, 65.2 %

# Heart Disease: Naive Bayes

- Parameters: Kernel density probability estimation for continuous attributes
- Model: Density probability estimation for each class
- Accuracy: 84.0 %

# Conclusions

- The model from decision trees allow to see the relevant features and to know how the new examples are classified
- Other methods have better accuracy
- k-NN needs trial and error to find the good value for k. Retrieving similar cases could be use to justify (understand) the classification results
- Naive Bayes model is difficult to understand

# Letter Recognition

## Capital Letter Recognition

- 14 Attributes (All continuous)
- Attributes: horizontal position of box, vertical position of box, width of box, height of box, total num on pixels, mean x of on pixels in box, ...
- 20000 instances
- 26 classes (A-Z)
- Validation: 10 fold cross validation

# Letter Recognition: Decision trees

- Parameters: Post pruning (Binomial 0.25)
- Size of the tree: 2451 nodes (1226 leaves)
- Accuracy: 87.88 %
- Training time ( $\sim 30$  seconds), test time ( $\sim 10$  minutes)
- Without pruning: 2673 nodes (1337 leaves), accuracy 87.88 %
- Bagging: accuracy 92.7 %
- Boosting: accuracy 95.52 %

# Letter Recognition: K-nearest neighbour

- Parameters: 3 neighbours
- Model: the 20000 instances
- Accuracy: 96.0%
- Training time (0), test ( $>1$  hour)

# Letter Recognition: Naive Bayes

- Parameters: Kernel estimation for estimating continuous probability distributions
- Model: Class attributes probability distributions
- Accuracy: 73.0%
- Training time ( $\sim 2$  seconds), test ( $\sim 5$  minutes)

# Conclusions

- It is difficult to interpret the decision tree because of its size, accuracy is relatively high, the cost of obtaining the model and testing is moderate. The use of ensembles improves the accuracy.
- The cost of finding the right value for the parameter  $k$  for  $k$ -NN is high but its accuracy is higher
- The accuracy of naive bayes is the lower, the dependence among variables is obvious and this affects the results

# General Conclusions

- No method is systematically better
- Decision trees give more information about the data (class description, relevant attributes), they can be specially adequate for qualitative data. Lower accuracy can be compensated by the use of ensembles
- k-NN can approximate better concepts difficult to learn, but the cost of parameter fitting is usually higher. They can be specially good for datasets with quantitative attributes
- Naive Bayes does not need parameter fitting, its accuracy depends on the dependence of the attributes in the dataset. The number of examples can affect the result because of the estimation of the probability distributions.