


Data Mining and Knowledge Discovery

Practice notes – 12.11.2008

Discussion on numeric prediction


Data Mining and Knowledge Discovery
Knowledge Discovery and Knowledge Management in e-Science

Petra Kralj Novak
Petra.Kralj.Novak@ijs.si
 Practice, 2008/12/11




Discussion

- Consider a dataset with a target variable with five possible values:
 - non sufficient
 - sufficient
 - good
 - very good
 - excellent
 - Is this a classification or a numeric prediction problem?
 - What if such a variable is an attribute, is it nominal or numeric?
- Can KNN be used for classification tasks?
- Similarities between KNN and Naïve Bayes.
- Similarities and differences between decision trees and regression trees.



Classification or a numeric prediction problem?

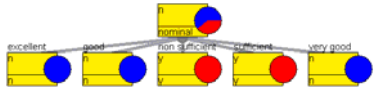
- Target variable with five possible values:
 - non sufficient
 - sufficient
 - good
 - very good
 - excellent
- Classification: the **misclassification cost** is the same if “non sufficient” is classified as “sufficient” or if it is classified as “very good”
- Numeric prediction: The error of predicting “2” when it should be “1” is 1, while the error of predicting “5” instead of “1” is 4.
- If we have a variable with ordered values, it should be considered numeric.



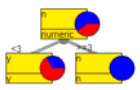
Nominal or numeric attribute?

- A variable with five possible values:
 - non sufficient
 - sufficient
 - good
 - very good
 - excellent


Nominal:



Numeric:




- If we have a variable with ordered values, it should be considered numeric.




Can KNN be used for classification tasks?

- YES.**
- In numeric prediction tasks, the average of the neighborhood is computed
- In classification tasks, the distribution of the classes in the neighborhood is computed



Similarities between KNN and Naïve Bayes.

- Both are **“black box”** models, which do not give the insight into the data.
- Both are **“lazy classifiers”**: they do not build a model in the training phase and use it for predicting, but they need the data when predicting the value for a new example (partially true for Naïve Bayes)



Data Mining and Knowledge Discovery

Practice notes – 12.11.2008

Discussion on numeric prediction

Regression trees	Decision trees
Data: attribute-value description	
Target variable: Continuous	Target variable: Categorical (nominal)
Evaluation: cross validation, separate test set, ...	
Error: MSE, MAE, RMSE, ...	Error: 1-accuracy
Algorithm: Top down induction, shortsighted method	
Heuristic: Standard deviation	Heuristic : Information gain
Stopping criterion: Standard deviation < threshold	Stopping criterion: Pure leafs (entropy=0)

