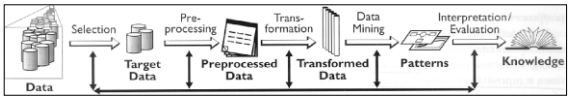
Data Mining and Knowledge Discovery Practice notes: Classification 2

Data Mining and Knowledge Discovery: Practice Notes

Petra Kralj Novak
Petra.Kralj.Novak@ijs.si
2013/01/08

Keywords



- Data
  - Attribute, example, attribute-value data, target variable, class, discretization
- Data mining
  - Heuristics vs. exhaustive search, decision tree induction, entropy, information gain, overfitting, Occam's razor, model pruning, naïve Bayes classifier, KNN, association rules, support, confidence, predictive vs. descriptive DM, numeric prediction, regression tree, model tree
- Evaluation
  - Train set, test set, accuracy, confusion matrix, cross validation, true positives, false positives, ROC space, error

Practice plan

- 2012/11/20: Predictive data mining 1
  - Decision trees
  - Evaluating classifiers 1: separate test set, confusion matrix, classification accuracy
  - Hands on Weka 1: Just a taste of Weka
- 2012/12/04: Predictive data mining 2
  - Discussion about decision trees
  - Naïve Bayes classifier
  - Evaluating classifiers 2: Cross validation
  - Numeric prediction
  - Hands on Weka 2: Classification and numeric prediction
- 2013/01/08: Descriptive data mining
  - Discussion on classification
  - Association rules
  - Hands on Weka 3: Descriptive data mining
  - Discussion about seminars and exam
- 2013/1/15: Written exam, seminar proposal discussion
- 2013/2/12: Data mining seminar presentations

Discussion

1. Compare naïve Bayes and decision trees (similarities and differences) .
2. Can KNN be used for classification tasks?
3. Compare KNN and Naïve Bayes.
4. Compare decision trees and regression trees.
5. Consider a dataset with a target variable with five possible values:
   1. non sufficient
   2. sufficient
   3. good
   4. very good
   5. excellent
   1. Is this a classification or a numeric prediction problem?
   2. What if such a variable is an attribute, is it nominal or numeric?
6. Compare cross validation and testing on a different test set.
7. Why do we prune decision trees?
8. List 3 numeric prediction methods.

Comparison of naïve Bayes and decision trees

- Similarities
  - Classification
  - Same evaluation
- Differences
  - Missing values
  - Numeric attributes
  - Interpretability of the model

Comparison of naïve Bayes and decision trees: Handling missing values

Will the spider catch these two ants?

- Color = white, Time = night ← **missing value for attribute Size**
- Color = black, Size = large, Time = day

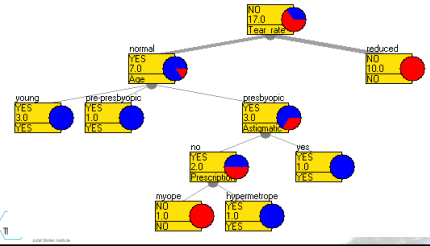
$$p(c_1|v_1,v_2) = p(Caught = YES|Color = white, Time = night) = p(Caught = YES) * \frac{p(Caught = YES|Color = white)}{p(Caught = YES)} * \frac{p(Caught = YES|Time = night)}{p(Caught = YES)} = \frac{1}{2} * \frac{\frac{1}{2}}{\frac{1}{2}} * \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{4}$$

Naïve Bayes uses all the available information.

Data Mining and Knowledge Discovery
Practice notes: Classification 2

Comparison of naïve Bayes and decision trees: Handling missing values

Age	Prescription	Astigmatic	Tear_Rate
?	hypermetrope	no	normal
pre-presbyopic	myope	?	normal



Comparison of naïve Bayes and decision trees: Handling missing values

Algorithm **ID3**: does not handle missing values
Algorithm **C4.5** (J48) deals with two problems:

- Missing values in **train** data:
  - Missing values are not used in gain and entropy calculations
- Missing values in **test** data:
  - A missing **continuous** value is replaced with the median of the training set
  - A missing **categorical** values is replaced with the most frequent value

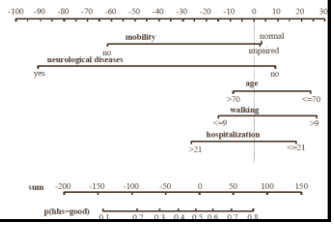
Comparison of naïve Bayes and decision trees: numeric attributes

- Decision trees **ID3** algorithm: does not handle continuous attributes → data need to be discretized
- Decision trees **C4.5** (J48 in Weka) algorithm: deals with continuous attributes as shown earlier
- **Naïve Bayes**: does not handle continuous attributes → data need to be discretized

(some implementations do handle)

Comparison of naïve Bayes and decision trees: Interpretability

- Decision trees are easy to understand and interpret (if they are of moderate size)
- Naïve bayes models are of the "black box type".
- Naïve bayes models have been visualized by nomograms.



Discussion

1. Compare naïve Bayes and decision trees (similarities and differences) .
2. Can KNN be used for classification tasks?
3. Compare KNN and Naïve Bayes.
4. Compare decision trees and regression trees.
5. Consider a dataset with a target variable with five possible values:
   1. non sufficient
   2. sufficient
   3. good
   4. very good
   5. excellent
   1. Is this a classification or a numeric prediction problem?
   2. What if such a variable is an attribute, is it nominal or numeric?
6. Compare cross validation and testing on a different test set.
7. Why do we prune decision trees?
8. List 3 numeric prediction methods.

KNN for classification?

- Yes.
- A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor.

Data Mining and Knowledge Discovery
Practice notes: Classification 2

Discussion

1. Compare naïve Bayes and decision trees (similarities and differences) .
2. Can KNN be used for classification tasks?
3. Compare KNN and Naïve Bayes.
4. Compare decision trees and regression trees.
5. Consider a dataset with a target variable with five possible values:
    1. non sufficient
    2. sufficient
    3. good
    4. very good
    5. excellent
    1. Is this a classification or a numeric prediction problem?
    2. What if such a variable is an attribute, is it nominal or numeric?
6. Compare cross validation and testing on a different test set.
7. Why do we prune decision trees?
8. List 3 numeric prediction methods.

Comparison of KNN and naïve Bayes

	Naïve Bayes	KNN
Used for	Classification	Classification and numeric prediction
Handle categorical data	Yes	Proper distance function needed
Handle numeric data	Discretization needed	Yes
Model interpretability	Limited	No
Lazy classification	Partial	Yes
Evaluation	Cross validation,...	Cross validation,...
Parameter tuning	No	No

Discussion

1. Compare naïve Bayes and decision trees (similarities and differences) .
2. Can KNN be used for classification tasks?
3. Compare KNN and Naïve Bayes.
4. Compare decision trees and regression trees.
5. Consider a dataset with a target variable with five possible values:
    1. non sufficient
    2. sufficient
    3. good
    4. very good
    5. excellent
    1. Is this a classification or a numeric prediction problem?
    2. What if such a variable is an attribute, is it nominal or numeric?
6. Compare cross validation and testing on a different test set.
7. Why do we prune decision trees?
8. List 3 numeric prediction methods.

Comparison of regression and decision trees

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)

Discussion

1. Compare naïve Bayes and decision trees (similarities and differences) .
2. Can KNN be used for classification tasks?
3. Compare KNN and Naïve Bayes.
4. Compare decision trees and regression trees.
5. Consider a dataset with a target variable with five possible values:
    1. non sufficient
    2. sufficient
    3. good
    4. very good
    5. excellent
    1. Is this a classification or a numeric prediction problem?
    2. What if such a variable is an attribute, is it nominal or numeric?
6. Compare cross validation and testing on a different test set.
7. Why do we prune decision trees?
8. List 3 numeric prediction methods.

Classification or a numeric prediction problem?

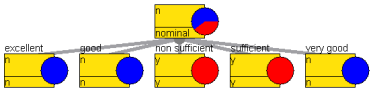
- Target variable with five possible values:
    1. non sufficient
    2. sufficient
    3. good
    4. very good
    5. excellent
- Classification: the **misclassification cost** is the same if "non sufficient" is classified as "sufficient" or if it is classified as "very good"
- Numeric prediction: 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.

Data Mining and Knowledge Discovery
Practice notes: Classification 2

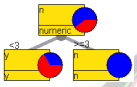
Nominal or numeric attribute?

- A variable with five possible values:
  1. non sufficient
  2. sufficient
  3. good
  4. very good
  5. Excellent

Nominal:



Numeric:



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

Discussion

1. Compare naïve Bayes and decision trees (similarities and differences) .
2. Can KNN be used for classification tasks?
3. Compare KNN and Naïve Bayes.
4. Compare decision trees and regression trees.
5. Consider a dataset with a target variable with five possible values:
   1. non sufficient
   2. sufficient
   3. good
   4. very good
   5. excellent
   1. Is this a classification or a numeric prediction problem?
   2. What if such a variable is an attribute, is it nominal or numeric?
6. Compare cross validation and testing on a different test set.
7. Why do we prune decision trees?
8. List 3 numeric prediction methods.

Comparison of cross validation and testing on a separate test set

- Both are methods for evaluating predictive models.
- Testing on a separate test set is simpler since we split the data into two sets: one for training and one for testing. We evaluate the model on the test data.
- Cross validation is more complex: It repeats testing on a separate test *n* times, each time taking 1/n of different data examples as test data. The evaluation measures are averaged over all testing sets therefore the results are more reliable.

Discussion

1. Compare naïve Bayes and decision trees (similarities and differences) .
2. Can KNN be used for classification tasks?
3. Compare KNN and Naïve Bayes.
4. Compare decision trees and regression trees.
5. Consider a dataset with a target variable with five possible values:
   1. non sufficient
   2. sufficient
   3. good
   4. very good
   5. excellent
   1. Is this a classification or a numeric prediction problem?
   2. What if such a variable is an attribute, is it nominal or numeric?
6. Compare cross validation and testing on a different test set.
7. Why do we prune decision trees?
8. List 3 numeric prediction methods.

Decision tree pruning

- To avoid overfitting
- Reduce size of a model and therefore increase understandability.

Discussion

1. Compare naïve Bayes and decision trees (similarities and differences) .
2. Can KNN be used for classification tasks?
3. Compare KNN and Naïve Bayes.
4. Compare decision trees and regression trees.
5. Consider a dataset with a target variable with five possible values:
   1. non sufficient
   2. sufficient
   3. good
   4. very good
   5. excellent
   1. Is this a classification or a numeric prediction problem?
   2. What if such a variable is an attribute, is it nominal or numeric?
6. Compare cross validation and testing on a different test set.
7. Why do we prune decision trees?
8. List 3 numeric prediction methods.

Data Mining and Knowledge Discovery
Practice notes: Classification 2

Numeric prediction methods

- Linear regression
- Regression trees
- Model trees
- KNN

Association Rules

Association rules

- Rules **X → Y**, X, Y conjunction of items
- Task: Find **all** association rules that satisfy minimum support and minimum confidence constraints

- **Support**:
  Sup(X → Y) = #XY/#D ≅ p(XY)

- **Confidence**:
  Conf(X → Y) = #XY/#X ≅ p(XY)/p(X) = p(Y|X)

Association rules - algorithm

1. generate frequent itemsets with a minimum support constraint
2. generate rules from frequent itemsets with a minimum confidence constraint

* Data are in a transaction database

Association rules – transaction database

Items: **A**=apple, **B**=banana,
C=coca-cola, **D**=doughnut

- Client 1 bought: A, B, C, D
- Client 2 bought: B, C
- Client 3 bought: B, D
- Client 4 bought: A, C
- Client 5 bought: A, B, D
- Client 6 bought: A, B, C

Frequent itemsets

- Generate frequent itemsets with support at least 2/6

A	B	C	D
1	1	1	1
	1	1	
	1		1
1		1	
1	1		1
1	1	1	

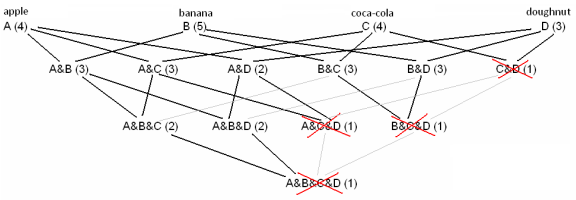
Data Mining and Knowledge Discovery
Practice notes: Classification 2

Frequent itemsets algorithm

Items in an itemset should be sorted alphabetically.

- Generate all 1-itemsets with the given minimum support.
- Use 1-itemsets to generate 2-itemsets with the given minimum support.
- From 2-itemsets generate 3-itemsets with the given minimum support as unions of 2-itemsets with the same item at the beginning.
- …
- From n-itemsets generate (n+1)-itemsets as unions of n-itemsets with the same (n-1) items at the beginning.

Frequent itemsets lattice



Frequent itemsets:

- A&B, A&C, A&D, B&C, B&D
- A&B&C, A&B&D

Rules from itemsets

- A&B is a frequent itemset with support 3/6
- Two possible rules
  - A→B confidence = #(A&B)/#A = 3/4
  - B→A confidence = #(A&B)/#B = 3/5
- All the counts are in the itemset lattice!

Quality of association rules

Support(X) = #X / #D ..... P(X)
Support(X→Y) = Support (XY) = #XY / #D ..... P(XY)
Confidence(X→Y) = #XY / #X ..... P(Y|X)

Lift(X→Y) = Support(X→Y) / (Support (X)*Support(Y))

Leverage(X→Y) = Support(X→Y) – Support(X)*Support(Y)

Conviction(X → Y) = 1-Support(Y)/(1-Confidence(X→Y))

Quality of association rules

Support(X) = #X / #D ..... P(X)
Support(X→Y) = Support (XY) = #XY / #D ..... P(XY)
Confidence(X→Y) = #XY / #X ..... P(Y|X)

Lift(X→Y) = Support(X→Y) / (Support (X)*Support(Y))
How many more times the items in X and Y occur together then it would be expected if the itemsets were statistically independent.

Leverage(X→Y) = Support(X→Y) – Support(X)*Support(Y)
Similar to lift, difference instead of ratio.

Conviction(X → Y) = 1-Support(Y)/(1-Confidence(X→Y))
Degree of implication of a rule.
Sensitive to rule direction.

Discussion

- Transformation of an attribute-value dataset to a transaction dataset.
- What would be the association rules for a dataset with two items A and B, each of them with support 80% and appearing in the same transactions as rarely as possible?
  - minSupport = 50%, min conf = 70%
  - minSupport = 20%, min conf = 70%
- What if we had 4 items: A, ¬A, B, ¬ B
- Compare decision trees and association rules regarding handling an attribute like "PersonID". What about attributes that have many values (eg. Month of year)