Data Mining and Knowledge Discovery: Practice Notes

Petra Kralj Novak
Petra.Kralj.Novak@ijs.si
2013/12/09
Practice plan

• 2013/11/11: 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

• 2013/11/18: Predictive data mining 2
  – Discussion on decision trees
  – Naïve Bayes classifier
  – Evaluating classifiers 2: Cross validation
  – Numeric prediction
  – Hands on Weka 2: Classification and numeric prediction

• 2013/12/9: Descriptive data mining
  – Discussion on classification
  – Association rules
  – Hands on Weka 3: Descriptive data mining
  – Discussion about seminars and exam

• 2013/12/16: Written exam, seminar proposal discussion
• 2014/1/8: Data mining seminar presentations
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
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.
9. What is discretization?
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  \textcolor{red}{\text{missing value for attribute Size}}
- Color = black, Size = large, Time = day

Naïve Bayes uses all the available information.
Comparison of naïve Bayes and decision trees: Handling missing values

<table>
<thead>
<tr>
<th>Age</th>
<th>Prescription</th>
<th>Astigmatic</th>
<th>Tear_Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>?</td>
<td>hypermetrope</td>
<td>no</td>
<td>normal</td>
</tr>
<tr>
<td>pre-presbyopic</td>
<td>myope</td>
<td>?</td>
<td>normal</td>
</tr>
</tbody>
</table>

Decision tree diagram showing the classification process for age and prescription with handling of missing values.
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.
9. What is discretization?
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.
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.
9. What is discretization?
Comparison of KNN and naïve Bayes

<table>
<thead>
<tr>
<th></th>
<th>Naïve Bayes</th>
<th>KNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Used for</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Handle categorical data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Handle numeric data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model interpretability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lazy classification</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parameter tuning</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Comparison of KNN and naïve Bayes

<table>
<thead>
<tr>
<th></th>
<th>Naïve Bayes</th>
<th>KNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Used for</td>
<td>Classification</td>
<td>Classification and numeric prediction</td>
</tr>
<tr>
<td>Handle categorical data</td>
<td>Yes</td>
<td>Proper distance function needed</td>
</tr>
<tr>
<td>Handle numeric data</td>
<td>Discretization needed</td>
<td>Yes</td>
</tr>
<tr>
<td>Model interpretability</td>
<td>Limited</td>
<td>No</td>
</tr>
<tr>
<td>Lazy classification</td>
<td>Partial</td>
<td>Yes</td>
</tr>
<tr>
<td>Evaluation</td>
<td>Cross validation,...</td>
<td>Cross validation,...</td>
</tr>
<tr>
<td>Parameter tuning</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>
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.
9. What is discretization?
Comparison of regression and decision trees

1. Data
2. Target variable
3. Evaluation
4. Error
5. Algorithm
6. Heuristic
7. Stopping criterion
Comparison of regression and decision trees

<table>
<thead>
<tr>
<th>Regression trees</th>
<th>Decision trees</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data</strong>: attribute-value description</td>
<td></td>
</tr>
<tr>
<td><strong>Target variable</strong>: Continuous</td>
<td><strong>Target variable</strong>: Categorical (nominal)</td>
</tr>
<tr>
<td><strong>Evaluation</strong>: cross validation, separate test set, ...</td>
<td></td>
</tr>
<tr>
<td><strong>Error</strong>: MSE, MAE, RMSE, ...</td>
<td><strong>Error</strong>: 1-accuracy</td>
</tr>
<tr>
<td><strong>Algorithm</strong>: Top down induction, shortsighted method</td>
<td></td>
</tr>
<tr>
<td><strong>Heuristic</strong>: Standard deviation</td>
<td><strong>Heuristic</strong>: Information gain</td>
</tr>
<tr>
<td><strong>Stopping criterion</strong>: Standard deviation&lt; threshold</td>
<td><strong>Stopping criterion</strong>: Pure leafs (entropy=0)</td>
</tr>
</tbody>
</table>
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.
9. What is discretization?
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.
Nominal or numeric attribute?

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

- 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.
9. What is discretization?
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.
9. What is discretization?
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.
9. What is discretization?
Numeric prediction methods

- Linear regression
- Regression trees
- Model trees
- KNN
Association Rules
Association rules

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

**Support:**
\[
\text{Sup}(X \rightarrow Y) = \frac{\#XY}{\#D} \approx p(XY)
\]

**Confidence:**
\[
\text{Conf}(X \rightarrow Y) = \frac{\#XY}{\#X} \approx \frac{p(XY)}{p(X)} = p(Y|X)
\]
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

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
Frequent itemsets algorithm

Items in an itemset should be sorted alphabetically.
1. Generate all 1-itemsets with the given minimum support.
2. Use 1-itemsets to generate 2-itemsets with the given minimum support.
3. From 2-itemsets generate 3-itemsets with the given minimum support as unions of 2-itemsets with the same item at the beginning.
4. ...
5. From n-itemsets generate (n+1)-itemsets as unions of n-itemsets with the same (n-1) items at the beginning.

- To generate itemsets at level n+1 items from level n are used with a constraint: itemsets have to start with the same n-1 items.
Frequent itemsets:

- 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  \hspace{1cm} \text{Support} (X) \hspace{1cm} \text{P}(X)

Support(X \rightarrow Y) = \text{Support} (XY) = \#XY / \#D \hspace{1cm} \text{Support} (XY) \hspace{1cm} \text{P}(XY)

Confidence(X \rightarrow Y) = \#XY / \#X \hspace{1cm} \text{Confidence} (X \rightarrow Y) \hspace{1cm} \text{P}(Y|X)

Lift(X \rightarrow Y) = \frac{\text{Support}(X \rightarrow Y)}{\text{Support} (X) \times \text{Support} (Y)}

Leverage(X \rightarrow Y) = \text{Support}(X \rightarrow Y) - \text{Support}(X) \times \text{Support}(Y)

Conviction(X \rightarrow Y) = 1 - \frac{\text{Support}(Y)}{1 - \text{Confidence}(X \rightarrow Y)}
Quality of association rules

Support(X) = \#X / \#D \quad \text{P(X)}

Support(X \rightarrow Y) = \text{Support}(XY) = \#XY / \#D \quad \text{P(XY)}

Confidence(X \rightarrow Y) = \#XY / \#X \quad \text{P(Y|X)}

\hline

\text{Lift}(X \rightarrow Y) = \frac{\text{Support}(X \rightarrow Y)}{(\text{Support}(X) \times \text{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.

\text{Leverage}(X \rightarrow Y) = \text{Support}(X \rightarrow Y) - \text{Support}(X) \times \text{Support}(Y)

Similar to lift, difference instead of ratio.

\text{Conviction}(X \rightarrow Y) = 1 - \frac{\text{Support}(Y)}{(1 - \text{Confidence}(X \rightarrow 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 (e.g. Month of year)