Performance Measures in Data Mining
Common Performance Measures used in Data Mining and Machine Learning Approaches

L. Richter  J.M. Cejuela

Department of Computer Science
Technical University of Munich

Master Lab Course Data Mining, SoSe 2016, Jun 29th
Outline

Item Set and Association Rule Weights
  Simple Measures
  Complex Measures

Classification
  Basic Performance Measures
  Complex Measures – Performance Curves

Regression

Assessment Strategies
Various algorithms can yield frequent item sets. From frequent item sets $c$ and $c \cup \{i\}$ you can derive if $c$ then $\{i\}$. Typically there is only one item in the RHS (right hand side of the rule).

- **Support (of an item set):**
  \[
  \text{sup}(X \rightarrow Y) = \text{sup}(Y \rightarrow X) = P(X \land Y)
  \]
  how many times the item set is found in the database

- **Confidence (of a rule):**
  \[
  \text{conf}(X \rightarrow Y) = P(Y|X) = P(X \text{ and } Y)/P(X) = \text{sup}(X \rightarrow Y)/\text{sup}(X)
  \]
Measures for Item Sets

Various algorithms can yield frequent item sets. From frequent item sets \( c \) and \( c \cup \{i\} \) you can derive \( if \ c \ then \ \{i\} \). Typically there is only one item in the RHS (right hand side of the rule).

- **Support (of an item set):**
  
  \[
  sup(X \rightarrow Y) = sup(Y \rightarrow X) = P(X \land Y)
  \]

  how many times the item set is found in the database

- **Confidence (of a rule):** \( conf(X \rightarrow Y) = P(Y|X) = \frac{P(X \text{ and } Y)}{P(X)} = \frac{sup(X \rightarrow Y)}{sup(X)} \)
Measures for Item Sets

Various algorithms can yield frequent item sets. From frequent item sets $c$ and $c \cup \{i\}$ you can derive $if \ c \ then \ \{i\}$. Typically there is only one item in the RHS (right hand side of the rule).

- **Support (of an item set):**
  \[
sup(X \rightarrow Y) = sup(Y \rightarrow X) = P(X \land Y)
  \]
  how many times the item set is found in the database

- **Confidence (of a rule):**
  \[
  conf(X \rightarrow Y) = P(Y|X) = P(X \ and \ Y)/P(X) = sup(X \rightarrow Y)/sup(X)
  \]
Measures for Item Sets cont.’d

Measures how frequent an item set / how interesting a rule is in comparison to the expected occurrence (interesting):

- **Leverage (of an item set):**
  \[
  \text{lev}(X \rightarrow Y) = P(X \text{ and } Y) - (P(X)P(Y))
  \]

- **Lift (of a rule):**
  \[
  \text{lift}(X \rightarrow Y) = \text{lift}(Y \rightarrow X) = \frac{P(X \land Y)}{P(X)P(Y)} = \frac{\text{conf}(X \rightarrow Y)/\text{sup}(Y)}{\text{conf}(Y \rightarrow X)/\text{sup}(X)}
  \]

- **Conviction (of a rule):** Similar to Lift, but directed.
  Compares the probability that X appears without Y, if they were independent with the observed frequency of X and Y.
  \[
  \text{conviction}(X \rightarrow Y) = \frac{P(X)P(\neg Y)}{P(X \land \neg Y)} = \frac{(1 - \text{sup}(Y))/(1 - \text{conf}(X \rightarrow Y))}
  \]
Measures for Item Sets cont.’d

Measures how frequent an item set / how interesting a rule is in comparison to the expected occurrence (interesting):

- Leverage (of an item set):
  \[
  lev(X \rightarrow Y) = P(X \text{ and } Y) - (P(X)P(Y))
  \]

- Lift (of a rule):
  \[
  \text{lift}(X \rightarrow Y) = \text{lift}(Y \rightarrow X) = \frac{P(X \land Y)}{(P(X)P(Y))} = \frac{\text{conf}(X \rightarrow Y)}{\sup(Y)} = \frac{\text{conf}(Y \rightarrow X)}{\sup(X)}
  \]

- Conviction (of a rule): Similar to Lift, but directed
  Compares the probability that X appears without Y, if they were independent with the observed frequency of X and Y.
  \[
  \text{conviction}(X \rightarrow Y) = P(X)P(\neg Y)/P(X \land \neg Y) = \frac{(1 - \sup(Y))}{(1 - \text{conf}(X \rightarrow Y))}
  \]
Measures for Item Sets cont.’d

Measures how frequent an item set / how interesting a rule is in comparison to the expected occurrence (interesting):

- **Leverage (of an item set):**
  \[
  \text{lev}(X \rightarrow Y) = P(X \text{ and } Y) - (P(X)P(Y))
  \]

- **Lift (of a rule):**
  \[
  \text{lift}(X \rightarrow Y) = \text{lift}(Y \rightarrow X) = \frac{P(X \land Y)}{P(X)P(Y)} = \frac{\text{conf}(X \rightarrow Y)}{\text{sup}(Y)} = \frac{\text{conf}(Y \rightarrow X)}{\text{sup}(X)}
  \]

- **Conviction (of a rule):** Similar to Lift, but directed
  Compares the probability that X appears without Y, if they were independent with the observed frequency of X and Y.
  \[
  \text{conviction}(X \rightarrow Y) = P(X)P(\neg Y)/P(X \land \neg Y) = (1 - \text{sup}(Y))/(1 - \text{conf}(X \rightarrow Y))
  \]
J-Measure

- empirically observed accuracy of rule
- Cross-entropy (measuring how good a distribution approximates another distribution) between the binary variables $\phi$ and $\theta$ with vs. without conditioning on event $\theta$

$$J(\theta \Rightarrow \phi) = p(\theta) \left( p(\phi|\theta) \log \frac{p(\phi|\theta)}{p(\phi)} + (1 - p(\phi|\theta)) \log \frac{1 - p(\phi|\theta)}{1 - p(\phi)} \right)$$
### Basic Building Blocks for Performance Measures

- **True Positive (TP):** positive instances predicted as positive
- **True Negative (TN):** negative instances predicted as negative
- **False Positive (FP):** negative instances predicted as positive
- **False Negative (FN):** positive instances predicted as negative

#### Confusion Matrix:

<table>
<thead>
<tr>
<th></th>
<th>Predicted a</th>
<th>Predicted b</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Real a</strong></td>
<td>TP</td>
<td>FN</td>
</tr>
<tr>
<td><strong>Real b</strong></td>
<td>FP</td>
<td>TN</td>
</tr>
</tbody>
</table>
Performance Measures

Accuracy, \( acc = \frac{TP + TN}{TP + FN + FP + TN} \)

\[ = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \]

Error rate, \( err = \frac{FN + FP}{TP + FN + FP + TN} = 1 - acc \)

\[ = \frac{\text{Number of wrong predictions}}{\text{Total number of predictions}} \]
Performance Measures cont’d

True Positive Rate, TPR, Sensitivity = \( \frac{TP}{TP + FN} \)

True Negative Rate, TNR, Specificity = \( \frac{TN}{TN + FP} \)

False Positive Rate, FPR = \( \frac{FP}{TN + FP} \)

False Negative Rate, FNR = \( \frac{FN}{TP + FN} \)
Performance Measures cont’d

True Positive Rate, $TPR$, Sensitivity $= \frac{TP}{TP + FN}$

True Negative Rate, $TNR$, Specificity $= \frac{TN}{TN + FP}$

False Positive Rate, $FPR = \frac{FP}{TN + FP}$

False Negative Rate, $FNR = \frac{FN}{TP + FN}$
Performance Measures cont’d

True Positive Rate, TPR, Sensitivity = \( \frac{TP}{TP + FN} \)

True Negative Rate, TNR, Specificity = \( \frac{TN}{TN + FP} \)

False Positive Rate, FPR = \( \frac{FP}{TN + FP} \)

False Negative Rate, FNR = \( \frac{FN}{TP + FN} \)
Performance Measures cont’d

\[
\text{True Positive Rate, } TPR, \text{ Sensitivity} = \frac{TP}{TP + FN}
\]

\[
\text{True Negative Rate, } TNR, \text{ Specificity} = \frac{TN}{TN + FP}
\]

\[
\text{False Positive Rate, } FPR = \frac{FP}{TN + FP}
\]

\[
\text{False Negative Rate, } FNR = \frac{FN}{TP + FN}
\]
Performance Measures cont’d

Precision, \( p = \frac{TP}{TP + FP} \)

Recall, \( r = \frac{TP}{TP + FN} \)

\( F_1 \) measure = \( \frac{2rp}{r + p} = \frac{2 \times TP}{2 \times TP + FP + FN} \)
Performance Measures cont’d

**Precision**, \( p = \frac{TP}{TP + FP} \)

**Recall**, \( r = \frac{TP}{TP + FN} \)

\( F_1 \text{measure} = \frac{2rp}{r + p} = \frac{2 \times TP}{2 \times TP + FP + FN} \)
Performance Measures cont’d

\[
\text{Precision, } p = \frac{TP}{TP + FP}
\]

\[
\text{Recall, } r = \frac{TP}{TP + FN}
\]

\[
F_1 \text{ measure} = \frac{2rp}{r + p} = \frac{2 \times TP}{2 \times TP + FP + FN}
\]
Performance Curves

- "costs" of different error types are different
- prediction behaviour changes over the test set
- performance display in 2D
- different domains prefer different chart types
Lift Charts

- from marketing area to evaluate mailing success
- y-axis: number or percentage of responders
- x-axis: sample
- red diagonal: random lift
- green line: optimum lift

Taken from
http://www.dmg.org/rfc/pmml-3.2/ModelExplanation.html
ROC Curves

- Receiver Operator Characteristics
- $y$-axis: TPR
- $x$-axis: FPR
- diagonal: random guessing (TPR=FPR)

Taken from http://en.wikipedia.org/wiki/File:Roccurves.png
Sensitivity vs. Specificity

- preferred in medicine
- $y$-axis: TPR
- $x$-axis: TNR (specificity)
- also frequently as ROC curve with 1 - specificity

Taken from http://i.stack.imgur.com/fnUd2.png
Recall Precision Curves

- preferred in information retrieval
- positives are the documents retrieved in response to a query
- true positives are documents really relevant to the query
- $y$-axis: precision
- $x$-axis: recall

Taken from http://scikit-learn.github.io/scikit-learn.org/
Error Measures for Regression

Mean-squared error, $MSE = \frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{n}$

Root mean-squared error, $RMSE = \sqrt{\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{n}}$

Mean-absolute error, $MAE = \frac{|p_1 - a_1| + \cdots + |p_n - a_n|}{n}$
Error Measures for Regression

Mean squared error, \( MSE = \frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{n} \)

Root mean squared error, \( RMSE = \sqrt{\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{n}} \)

Mean absolute error, \( MAE = \frac{|p_1 - a_1| + \cdots + |p_n - a_n|}{n} \)
Error Measures for Regression

Mean−squared error, \(MSE = \frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{n}\)

Root mean−squared error, \(RMSE = \sqrt{\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{n}}\)

Mean−absolute error, \(MAE = \frac{|p_1 - a_1| + \cdots + |p_n - a_n|}{n}\)
Relative Error Measures

\[
\text{Relative squared error} = \frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{(a_1 - \bar{a})^2 + \cdots + (a_n - \bar{a})^2}
\]

\[
\text{Root relative squared error} = \sqrt{\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{(a_1 - \bar{a})^2 + \cdots + (a_n - \bar{a})^2}}
\]

\[
\text{Relative absolute error} = \frac{|p_1 - a_1| + \cdots + |p_n - a_n|}{|a_1 - \bar{a}| + \cdots + |a_n - \bar{a}|}
\]
General Problem

- each algorithm abstracts from observations (instances)
- the aspects kept and the aspect discarded differ between the learning scheme (inductive bias)
- this means also: information about individual instances are contained in the model, too
- individual instance information leads to overfitting
Solution Strategies

- use fresh data, i.e. instances not used for the training
- for very large numbers of instances: simple split in test and training set
- most common: 10-fold cross validation
- LOOCV: Leave one out cross validation