Problem Analysis and Preprocessing

Paul Prasse, Niels Landwehr, Tobias Scheffer
Overview

- Analysis of learning problems
  - Understanding requirements
  - Deriving a solution
  - Developing an evaluation protocol

- Data preprocessing
  - Data integration
  - Feature representation
  - Missing values
  - Feature selection
Problem Analysis

- Engineering approach to problem solving
- Understanding the requirements
  - Application goal, quality metric
  - Properties of the data, of the process that generates the data
  - Application-specific requirements
  - Locating the problem in the taxonomy of machine learning paradigms
  - Do underlying assumptions of methods match the problem requirements?
- Developing a solution
- Developing an evaluation strategy
Understanding the Problem Requirements

- Differing cultures in different industries
- E.g., automotive industry
  - 10- to 20-pages of written software requirements specifications are not uncommon
- Usual case:
  - User / customer has idea which property a good solution should have
  - Exact problem setting and requirements have to be determined in interviews.
Example: Email Service Provider

- Problem: email spam exhausts hard drives and processing capacity
- Server and storage are massive cost factors
- Legal requirements: messages that have been accepted for delivery must not be deleted
Example: Email Service Provider

- Individual mailing campaigns distributed by botnets create enormous volumes of data

This is a good way to make a right move and receive your due benefits... if you are qualified but are lacking that piece of paper. Get one from us in a fraction of the time.

If you want to get better - you must Contact us 24 hours a day to start improving your life!

~CALL FOR A FREE CONSULTATION~
1-407-245-7320 Please leave us a voice message with your phone number with country code if outside USA and name and we will get back to you as soon as possible.

This is a nice way to make a right move and receive your due benefits... if you are qualified but are lacking that piece of paper. Get one from us in a fraction of the time.

If you want to get better - you must Call us NOW to start improving your life!

~CONTACT US FOR A FREE CONSULTATION~
1-407-245-7320 You must leave us a voice message with your name and phone number with country code if outside USA and we'll call you back as soon as possible.

This is an excellent chance to make a right move and receive your due benefits... if you are qualified but are lacking that piece of paper. Get one from us in a short time.

Call Us to start improving your life!

~CONTACT US FOR A FREE CONSULTATION~
1-407-245-7320 You should leave us a message with your phone number with country code if outside USA and name and we'll contact you asap.

This is a good chance to make a right move and receive your due benefits... if you are qualified but are lacking that piece of paper. Get one from us in a fraction of the time.

If you want to get better - you must Contact us 24 hours a day and 7 days a week! to start improving your life!

~CALL US FOR A FREE CONSULTATION~
1-407-245-7320 You should leave us a message with your phone number with country code if outside USA and name and we'll contact you asap.
Example: Email Service Provider

- Administrators notice large campaigns, write regular expression that matches campaign
- Email server then refuses to accept messages that match these regular expressions
- Problems: campaigns have to be noticed in time, admin has to act (weekends? Holidays?)
- If legitimate messages do not arrive, number of complaint calls to call center increases
Fallbeispiel Email Service Provider

- Requirements for automated solution?
- Evaluation metric?
- Modeling as a learning problem?
  - Type of learning problem?
  - Model space?
  - Loss function? Regularizer?
Taxonomy of Learning Problems

- Supervised: Training data contain values for variable that model has to predict
  - Classification: categorical variable
  - Regression: continuous variable
  - Ordinal regression, finite, ordered set of values
  - Rankings: ordering of elements
  - Structured prediction: sequence, tree, graph, ...
  - Recommendation: Item-by-user matrix
Taxonomy of Learning Problems

- Unsupervised: discover structural properties of data
  - Clustering
  - Unsupervised feature learning: find attributes that can be used to describe the data well
  - Anomaly detection
- Control / reinforcement learning: learning to control a dynamical system
- Many further models
  - Semi-supervised learning
  - Supervised clustering
  - ...
Data Availability

- Batch learning: all data available
- Online learning: data come in one at a time; incremental model updates
Data Availability

- Number of data
  - Very few?
  - So many that they have to be stored and processed distributedly?

- Number of attributes
  - Too few?
  - Too many to process?
  - Sparse (most entries zero)?

- Quality of data
  - Missing values?
  - Erroneous values? Measurement errors?
Representational Properties of the Data

- Balanced class ratio? Rare classes?
- Class ratio representative?
- Marginal distribution $p(x)$ in the data equal to distribution at application time? (If not, learning under covariate shift)
- Values of the target attribute from the real target distribution or from an auxiliary distribution (laboratory experiments, simulation data)
- Recent data? Does the process change over time?
Data Properties

- One or several data sources?
- Credibility? Quality? Consistency?
- Availability
  - Fixed, given data set?
  - Does a data collection protocol have to be developed?
Data Dependencies

- Independent observations

- Sequences

- Interdependent data
Example: Email Service Provider

- Modeled as two subsequent learning problems
  1. Campaign discovery
Example: Email Service Provider

- Modeled as two subsequent learning problems
  1. Campaign discovery
  2. Creating a regular expression for each campaign

This is a \[ a-z]+ to make a right move and receive your due benefits… if you are qualified but are lacking that piece of paper. Get one from us in a \S+( \S+){0,2} time.

\S+( \S+){1,19} to start improving your life!

~C[A-Z]*FOR A FREE CONSULTATION~

(1-407-245-7320 You must |1-407-245-7320 You should |1-407-245-7320 Please )leave us a (|voice )message with your (name and |)phone number with country code if outside USA and \S+( \S+){0,4}ll (|contact |get back to )you \S+( \S+){0,4}.}
Example: Discovering Campaigns

- Unsupervised learning: cluster analysis
- Online processing of the data stream
- Optimization criterium:
  - Most likely partitioning of stream into clusters
- Instances: header and word-occurrence attributes

```
<table>
<thead>
<tr>
<th>Alternative</th>
<th>Beneficiary</th>
<th>Friend</th>
<th>Sterling</th>
<th>Zoo</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Dear Beneficiary,

your Email address has been picked online in this year's MICROSOFT CONSUMER AWARD as a Winner of One Hundred and Fifty Five Thousand Pounds Sterling…
**Example: Discovering Campaigns**

- **Offline evaluation**
  - Save all emails within a limited period of time
  - Manually partition into clusters
  - Metrics: agreement between manually and automatically generated clustering
  - False-positive rate, False-negative rate

- **Online evaluation, testing**
  - Find clusters during regular business operation
  - Show to admins in charge of blocking campaigns
  - Admin feedback: incomplete? Multiple campaigns? Ok to block?
Example: Find Regular Expression

- **Instances** $x$: sets of emails (set of strings)
- **Target attribute** $y$: Regular expression

$$y = \text{a [a-z]+ russian (girl|lady). I am 2[123] years old, weigh } \backslash d^+ \text{ kilograms and am } 1\backslash d\{2\} \text{ centimeters tall.}$$
Example: Find Regular Expression

- Training data \{(x_i, y_i)\}: Sets of strings and corresponding regular expression written by admin.

<table>
<thead>
<tr>
<th>This is a nice way to make a right move and receive your due</th>
<th>This is a best chance to make a right move and receive your due benefits... if you are qualified but are lacking that piece of paper. Get one from us in a fraction of the time.</th>
</tr>
</thead>
<tbody>
<tr>
<td>This is a good chance to make a right move and receive your due</td>
<td>If you want to get better - you must Contact us 24 hours a day and 7 days a week! to start improving your life!</td>
</tr>
<tr>
<td><del>CALL US FOR A FREE CONSULTATION</del></td>
<td>\textbackslash S+( \textbackslash S+){0,2} time.</td>
</tr>
<tr>
<td>1-407-245-7320 You should leave us a message with your phone number with country code if outside USA and name and we'll contact you asap.</td>
<td>\textbackslash S+( \textbackslash S+){1,19} to start improving your life!</td>
</tr>
<tr>
<td><del>C[ A-Z]* FOR A FREE CONSULTATION</del></td>
<td></td>
</tr>
</tbody>
</table>

This is a \([ a-z]+\) to make a right move and receive your due benefits... if you are qualified but are lacking that piece of paper. Get one from us in a \(S+( S+)\{0,2\}\) time.

\(S+( S+)\{1,19\}\) to start improving your life!

~C[ A-Z]* FOR A FREE CONSULTATION~
Example: Find Regular Expression

- Type of learning problem?
Example: Find Regular Expression

- Type of learning problem:
  - Training data contain correct regular expressions: ⇒ supervised learning
  - Target variable is a regular expression: discrete, structured ⇒ Structured prediction (learning with structured output spaces)
Example: Find Regular Expression

- Loss function $\ell(y_\theta(x_i), y_i)$: should measure how different the expressions are
Example: Find Regular Expression

- Loss function $\ell(\gamma_\theta(x_i), y_i)$: should measure how different the expressions are:
  - Proportion of non-identical nodes in the syntax tree.
- Regularization: $L_2$
Example: Evaluation and Testing
Example: Evaluation and Testing

- Online evaluation
  - Discovered campaigns and generated regular expressions presented to admins who have to blacklist campaigns
  - Rates of acceptance, acceptance + editing, rejection of the generated expressions
  - Rate of complaint calls to the call center
Overview

- Analysis of learning problems
  - Understanding requirements
  - Deriving a solution
  - Developing an evaluation protocol

- Data preprocessing
  - Data integration
  - Feature representation
  - Missing values
  - Feature selection
Data Integration

- Multiple data sources: integrate consistently; e.g., in data warehouse.
- Integration of multiple data formats.
- Schema integration: identify same / related attributes in different sources.
- Data conflicts (e.g., conversion of differing units).
- Discover redundant information (duplicate detection).
Feature Representation

- Transform attributes, depending on model structure
- For instance, linear model compute inner product of attributes and model parameters.
  - All attributes have to be numeric.
  - Larger attribute values: larger value of inner product
  - Categorial attributes, attributes without ordering, textual attributes have to be converted.

<table>
<thead>
<tr>
<th>Farbe</th>
<th>Größe</th>
<th>Produktkategorie</th>
</tr>
</thead>
<tbody>
<tr>
<td>rot</td>
<td>23</td>
<td>173</td>
</tr>
<tr>
<td>blau</td>
<td>23</td>
<td>173</td>
</tr>
<tr>
<td>grün</td>
<td>45</td>
<td>36</td>
</tr>
</tbody>
</table>

→

<table>
<thead>
<tr>
<th>Farbe rot</th>
<th>Farbe blau</th>
<th>Farbe grün</th>
<th>Größe</th>
<th>Produktkategorie 1</th>
<th>...</th>
<th>Produktkategorie 173</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>23</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>23</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>45</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Feature Representation

- Texts: TF- or TFIDF-representation
- Term-frequency vector: one dimension per term in a dictionary.
- Use all terms that appear at least 3 times in data.
- Value: number of occurrences of the term in the document.

\[
\begin{pmatrix}
0 \\
1 \\
0 \\
... \\
1 \\
0 \\
\end{pmatrix}
\]

Dear Beneficiary,
your Email address has been picked online in this year's MICROSOFT CONSUMER AWARD as a Winner of One Hundred and Fifty Five Thousand Pounds Sterling…
Feature Representation

- Texts: TF- or TFIDF-representation
- Frequently occurring terms (and, or, not, is) carry little semantic meaning.
- Idea: assign lower weight to frequent terms.
- Inverse document frequency

\[
IDF(\text{term}_i) = \log \frac{\#\text{documents}}{\#\text{documents that contain } \text{term}_i}
\]
Feature Representation

- **TFIDF vector**

\[
TFIDF(x) = \frac{1}{|x|} \left( \begin{array}{c}
TF(\text{term}_1) \cdot IDF(\text{term}_1) \\
\vdots \\
TF(\text{term}_n) \cdot IDF(\text{term}_n)
\end{array} \right)
\]
Feature Representation

- Texts: N-Gram vectors
- In TFIDS representation, information about the ordering of the terms is lost
- N-Gram features: one attribute for each k-tuple of subsequent terms (for all $k \leq N$)

$$
\begin{bmatrix}
1 \\
1 \\
1 \\
1 \\
... \\
1 \\
1 \\
... \\
1
\end{bmatrix}
$$

Dear Beneficiary,
your Email address has been picked online in this years MICROSOFT CONSUMER AWARD as a Winner of One Hundred and Fifty Five Thousand Pounds Sterling…

Email

Dear Beneficiary,
your Email address has been picked online in this years MICROSOFT CONSUMER AWARD as a Winner of One Hundred and Fifty Five Thousand Pounds Sterling…
Feature Representation

- Attributes can assume greatly varying ranges of values.
- Larger absolute values will have a larger impact on the decision function.
- It can be beneficial to normalize these ranges.
- This is equivalent to applying stronger regularization to the corresponding weights.
Feature Representation

- **Feature normalization**

  - Min/Max normalization:
    \[ x^{\text{new}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \left( x^{\text{new}}_{\text{max}} - x^{\text{new}}_{\text{min}} \right) + x^{\text{new}}_{\text{min}} \]

  - Z-Score normalisierung:
    \[ x^{\text{new}} = \frac{x - \mu_x}{\sigma_x} \]

  - Decimal scaling:
    \[ x^{\text{new}} = |x| \cdot 10^a \quad a = \max_{x} \left\{ i \in \mathbb{Z} \mid |x| \cdot 10^i < 1 \right\} \]

  - Logarithmic scaling:
    \[ x^{\text{new}} = \log_a x \]
Feature Representation

- It can be beneficial to construct complex features if such combinations cannot be found by the model.
  - With polynomial kernels, the model space contains all polynomials of the attributes, but not with linear kernels.

- Feature construction
  - Combinations of elementary features, e.g.,
    \[ (x_i, x_j) \rightarrow \left( x_j, \sqrt{x_i x_j}, x_i + x_j \right) \]
  - Mapping of elementary features, e.g.,
    \[ x_i \rightarrow \left( x_i, \log x_i, x_i^2 \right) \]
Attributes with Missing Values

- **Cause of missing values**
  - Missing at random (e.g., memory error, measuring fault).
  - Some values may be missing systematically, some classes may be more likely to have missing values.
  - Data integration: values may have been deleted due to inconsistencies.
  - Data aggregation: may have been aggregated or deleted for privacy.
  - ...
Attributes with Missing Values

- Delete all affected instances (attributes)
  - Makes only sense if very few instances have missing values (attribute is almost always missing)
- Extend range of values to special value “missing”.
- Introduce new binary attributes “Attribute_XY_missing”.
- Estimate missing values
  - (Class-specific) mean / median imputation
  - Infer most likely values (e.g., using EM algorithm).
- Do not handle missing values (if learning algorithm can intrinsically work with missing values).
Attributes with Erroneous Values

- Identifying incorrect values
  - Binning: equidistant discretization into bins
  - Bins with few instances may be outliers.
  - Clustering ⇒ Clusters with one or few instances may be outliers.
  - Active learning/labeling: inconsistencies between data and model
  - query human for correct label.

- Handling of erroneous values
  - Smoothing of numeric values (e.g., regression, moving average).
  - Treat as missing value / delete instance
  - ...
Feature Selection

- Selecting a subset of attributes can lead to better result.
- Dimensionality reduction.
- Plenty of approaches to feature selection.
  - For instance, principal component analysis
  - Forward-/Backward-selection
- Evaluation with training/test or n-fold cross validation
  - Train models with varying feature subsets on training set
  - Evaluate on testing set.
Feature Selection

- Feature selection for linear models; e.g.,
  - Train linear model,
  - Delete attributes with smallest weights,
  - Train linear model again,
  - Evaluate on test set,
  - Reiterate until optimum is reached.
Problem Analysis, Data Preprocessing

- Machine learning is an engineering science.
  - Analyze problem, understand requirements,
  - Map to known paradigms, refer to state of the art.
  - Derive a solution, and evaluation protocol.

- Data preprocessing can make difference between a model that works great and one that works not at all
  - Data integration,
  - Engineering of good features,
  - Handling of missing, erroneous values,
  - Feature selection