Feature Selection Methods based on Mutual Information

M. ROSÁRIO OLIVEIRA
CEMAT AND DEPT. MATEMÁTICA
Everest and papers about Feature Selection

Pedro Larrañaga.
The team:

Francisco Macedo
António Pacheco
Rui Valadas
Cláudia Pascoal
Eunice Carrasquinha
THE DATA BOOM
“It’s easy to lie with statistics. It’s hard to tell the truth without statistics.”

Andrejs Dunkels, Swedish mathematics teacher, mathematician, and writer

“Data is not information, information is not knowledge, knowledge is not understanding, understanding is not wisdom.”

Clifford Stoll, American astronomer, author and teacher

“Big Data is like teenage sex: -everyone talks about it, -nobody really knows how to do it, -everyone thinks everyone else is doing it, -so everyone claims they are doing it.”

Dan Ariely, Duke University

According with *Internet live stats*:

- ≈40% of the world population has an Internet connection today
- In 1995, < 1%
- The number of Internet users has increased *tenfold* from 1999 to 2013
  - 2005: 1st billion
  - 2010: 2nd billion
  - 2014: 3rd billion

Source: [https://www.internetlivestats.com/](https://www.internetlivestats.com/)
Internet

- It **changed** the way we live and interact
- We are **generating data** according with our:
  - business, professional and social preferences
  - habits and activities

- Google
- Tweeter
- YouTube
- Instagram
- Skype
- Facebook
Data Seen as Value

- “Big Data ... is the revolutionary belief that data are valuable.”
- “Scientists have long known that data could create new knowledge, but now the rest of the world, including government and management in particular, has realized that data create value.”
- “Suddenly it makes economic sense to try to extract value from all this data out there.”
- “But People don’t want data! They want answers!”

S. P. Murphy, 2013
S. Owen, 2014

David Hand
New Ways to Collect Data

- **Online/web surveys**
- **Mobile phone surveys**
- **GPS tracking**
- **Web tracking technologies** (like cookies or meters)
- **Social media monitoring/listening**
- **Crowdsourcing**

- **IoT - Internet of things**
- **Chatbot** is an artificially intelligent software program that uses natural language processing to hold a conversation with its users
- **Web Scraping** from websites

- **Volunteer Monitoring/Citizen Science**
- **Satellite data**
- **Invented/fake data**

---


Trump's celebration of its election as USA president
FEATURE SELECTION
More data is not necessarily more information...

Feature Selection:
- Extract from the data useful and valuable knowledge for real problem solving
FEATURE SELECTION: THE RIGHT DATA

- Select a **small subset** of the original features
- Designed to remove **irrelevant** and **redundant** features
- Reduce **computational complexity**
- Improve model **accuracy**
- Increase model **interpretability**
Feature Selection Methods

Classifier dependent

Wrapper

Embedded

Classifier independent

Filter
• **Idea:** search for feature subsets, using the classifier accuracy as the measure of utility for a candidate subset

• **Disadvantages:**
  • computational cost
  • selected features are classifier specific

• **Example:**
  • Stepwise regression
• **Idea:** Classifier estimations and feature selection are not separated and interact

• **Disadvantages:**
  • Selected features are classifier specific
  • Regularized_OF=OF+\lambda \text{regularization\_penalty}

• **Example:**
  • Regularization methods
Idea: Classifier estimations and feature selection are separated and depend on a specific measure of benefit

Most popular ones: rely on Mutual Information and Entropy

Mutual Information: measures linear and non-linear associations among features

Example:
- Forward feature selection methods based on MI
ENTROPY, MUTUAL INFORMATION
Motivated by problems in the field of telecommunications

A Mathematical Theory of Communication
C. E. Shannon (1948)

- A measure of uncertainty
- One formula that changed the world...

**Entropy**

\[
H(X) = - \sum_{x \in X} P(X = x) \ln P(X = x).
\]

- Does not depend on the values of \( X \), only on its prob.
- \( H(a) = 0 \)
- \( H(X) \geq 0 \), Non-negative
- \( H(X) = \ln(n), \ X \sim \text{Unif}\{a_1, \ldots, a_n\}, \) maximum

**Entropy Discrete rv**

\[
h(X) = - \int_{x \in \mathcal{X}} f_X(x) \ln f_X(x) \, dx.
\]

- Does not depend on the values of \( X \), only on its prob.
- Can be negative
- \( h(X) = \ln(a), \ X \sim \text{Unif}(0,a), \)
  - \( a = 1, \ h(X) = 0 \)
  - \( a < 1, \ h(X) < 0 \)

**Differential Entropy**

**Continuous rv**

- \( h(X) = \ln(a), \ X \sim \text{Unif}(0,a), \)
  - \( a = 1, \ h(X) = 0 \)
  - \( a < 1, \ h(X) < 0 \)
Mutual Information

\[ \text{Mutual Information} \]

**Discrete rv**

\[
\begin{align*}
\text{MI}(X, Y) &= \sum_{x \in X} \sum_{y \in Y} P(X = x, Y = y) \ln \frac{P(X = x, Y = y)}{P(X = x)P(Y = y)}. \\
\end{align*}
\]

- Measures linear and non-linear associations between \( X \) and \( Y \)
- \( \text{MI}(X, Y) \geq 0 \)
- Symmetric
- \( \text{MI}(X, Y) = 0 \) iff \( X \perp \!
\!
\!
\!
\perp Y \)
- \( \text{MI}(X, X) = H(X) \)

**Continuous rv**

\[
\begin{align*}
\text{MI}(X, Y) &= \int_{y \in Y} \int_{x \in X} f_{X,Y}(x,y) \ln \frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)} \, dx \, dy.
\end{align*}
\]

- Measures linear and non-linear associations between \( X \) and \( Y \)
- All properties hold, except
- \( \text{MI}(X, X) = +\infty \)
### Generalizations of MI:

<table>
<thead>
<tr>
<th>Triple Mutual Information</th>
<th>Conditional Mutual Information</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TMI</strong>(X,Y,Z)</td>
<td><strong>MI</strong>(X,Y</td>
</tr>
</tbody>
</table>

- The generalization to more than 2 rv -- not unique
- Measures association among X, Y, and Z
- Not necessarily non-negative

\[
\text{TMI}(X,Y,Z) = \text{MI}(X,Y) - \text{MI}(X,Y|Z).
\]

- Measures association between X and Y given Z
- **MI**(X,Y|Z)=0 iff X ⊥⊥ Y | Z, conditional independence
FORWARD FEATURE SELECTION
FORWARD FEATURE SELECTION

Goal: Select a small subset of the original features, excluding irrelevant and redundant features

\[ X_j = \arg \max_{X_i \in F} \text{MI}(C, S \cup \{X_i\}) \]

F= Candidate features

S= Selected features

C= Class-variable
FORWARD FEATURE SELECTION

Problem:

\[ X_j = \arg \max_{X_i} \text{MI}(C, S \cup \{X_i\}). \]

\[ \text{MI}(C, S \cup \{X_i\}) = \text{MI}(C, S) + \text{MI}(C, X_i | S) \]

- Calculation / Estimation
- Method
First Group of Methods

• **Goal:** Obtain subset of features leading to:
  • *maximum relevance* between the candidate feature and the class

<table>
<thead>
<tr>
<th>Method</th>
<th>Objective function evaluated at $X_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIM</td>
<td>MI($\mathcal{C}, X_i$)</td>
</tr>
</tbody>
</table>

• Maximizes *association with the class*
Second Group of Methods

• **Goal:** Obtain subset of features leading to:
  • *maximum relevance* between the candidate feature and the class
  • *minimum redundancy* of the candidate feature with respect to the already selected ones
Second Group of Methods

- **Inter-feature redundancy**: association between the candidate feature and the selected ones
- Avoids **collinearity** in the classifier estimation

<table>
<thead>
<tr>
<th>Method</th>
<th>Objective function evaluated at $X_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIFS</td>
<td>$MI(C, X_i) - \beta \sum_{X_j \in S} MI(X_i, X_j)$</td>
</tr>
<tr>
<td>mRMR</td>
<td>$MI(C, X_i) - \frac{1}{</td>
</tr>
<tr>
<td>maxMIFS</td>
<td>$MI(C, X_i) - \max_{X_j \in S} MI(X_i, X_j)$</td>
</tr>
</tbody>
</table>
Third Group of Methods

- Adds a third term:
  - Complementary accommodates possible dependencies among the features given the class
FORWARD FEATURE SELECTION
3RD GROUP

Third Group of Methods

- **Class-relevant redundancy**: contribution of a candidate feature to the explanation of the class, when taken together with already selected features.
FEATURE SELECTION METHODS

A THEORETICAL COMPARISON
How comparisons are usually done:

Source: Botelho (2020), Study project about feature selection.
Theoretical Setup:

Class-Variable: \( C = \text{Sgn}(X + kY) \)

Candidate Features: \( X, X - k'Y, \text{Sgn}(X), Z \)
THEORETICAL COMPARISON:
USING A DISTRIBUTIONAL SETTING

Features Order: Objective functions were calculated theoretically assuming $X, Y, \text{ and } Z$ are $N(0,1)$

Performance Measure:
Minimum Bayes Risk
= Minimum Probability of Misclassifiction
Advantages and drawbacks:

**FORWARD FEATURE SELECTION METHODS**

<table>
<thead>
<tr>
<th>Advantage</th>
<th>MIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class association</td>
<td>A</td>
</tr>
<tr>
<td>Feature redundancy</td>
<td></td>
</tr>
<tr>
<td>Feature complementarity</td>
<td></td>
</tr>
<tr>
<td>Drawback</td>
<td></td>
</tr>
<tr>
<td>Redundancy undervalued</td>
<td></td>
</tr>
<tr>
<td>Redundancy overscaled</td>
<td></td>
</tr>
<tr>
<td>Complementarity penalized</td>
<td></td>
</tr>
<tr>
<td>Unimportant term approximated</td>
<td></td>
</tr>
</tbody>
</table>

Ignores Redundancy
Advantages and drawbacks:

Comparisons:

Forward Feature Selection Methods

Cannot guarantee that relevant are selected before redundant and irrelevant features
Advantages and drawbacks:

Recommended, but there are room for improvements!

Comparisons:

Forward Feature Selection Methods
FEATURE SELECTION METHODS

SOLVING REAL PROBLEMS - ESTIMATION
Empirical estimators:

- Discretize continuous rv
- Estimate entropy and MI (discrete case)

\[ H(X^\Delta) + \ln \Delta \rightarrow h(X), \quad \text{as} \quad \Delta \rightarrow 0 \]
Empirical estimators:

\[ \text{MI}\left( X^\Delta, Y^\Delta \right) \rightarrow \text{MI}\left( X, Y \right) \]

\[
\text{MI}\left( X^\Delta, Y^\Delta \right) \\
- \ln \Delta_1 - \ln \Delta_2 + \ln \Delta_3 + \ln \Delta_4 \\
\rightarrow \text{MI}\left( X, Y \right)
\]
Example:

<table>
<thead>
<tr>
<th></th>
<th>MI(C,X)</th>
<th>MI(C,X-Y)</th>
<th>MI(X,X-Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kwak, Choi (2002)</td>
<td>0.8459</td>
<td>0.2621</td>
<td>0.6168</td>
</tr>
<tr>
<td>Huang et al. (2008)</td>
<td>0.8438</td>
<td>0.2807</td>
<td>0.6099</td>
</tr>
<tr>
<td>Pascoal (2014)</td>
<td>0.5932</td>
<td>0.1779</td>
<td>0.5004</td>
</tr>
<tr>
<td>TRUE Value</td>
<td>0.5932</td>
<td>0.1785</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Challenges:

- How to discretize
- Number of classes
- Continuity corrections
FEATURE SELECTION

THE FUTURE

Principal Component Analysis

Estimation: $\Sigma$

Eigen Problem

Statistics
THE TEAM:

Giulia Ferrandi
Igor Kravchenko
Michiel Hochstenbach
Main References

Theoretical foundations of forward feature selection methods based on mutual information

Francisco Macedo\textsuperscript{a,b}, M. Rosário Oliveira\textsuperscript{a,c}, António Pacheco\textsuperscript{a}, Rui Valadas\textsuperscript{a}

\textsuperscript{a}CEMAT and Department of Mathematics, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, Lisboa 1049-001, Portugal
\textsuperscript{b}ETH Zurich, 8092 Zurich, Switzerland
\textsuperscript{c}UT and Department of Electrical and Computer Engineering, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, Lisboa 1049-001, Portugal

Theoretical evaluation of feature selection methods based on mutual information

Cláudia Pasea\textsuperscript{a}, M. Rosário Oliveira\textsuperscript{a,b}, António Pacheco\textsuperscript{a}, Rui Valadas\textsuperscript{a}

\textsuperscript{a}CEMAT and Department of Mathematics, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1049-001 Lisboa, Portugal
\textsuperscript{b}ETH Zurich, 8092 Zurich, Switzerland