Lecture 2:
Introduction to Feature Selection

Isabelle Guyon
isabelle@clopinet.com
Notations and Examples
Feature Selection

- Thousands to millions of low level features: select the most relevant one to build better, faster, and easier to understand learning machines.
Leukemia Diagnosis

\[ \begin{align*}
\{y_i\}, & \quad i=1:m \\
\{-y_i\}, & \quad m
\end{align*} \]
**Prostate Cancer Genes**


Application to prostate cancer. Elisseeff-Weston, 2001
RFEM SVM for cancer diagnosis

Differenciation of 14 tumors. Ramaswamy et al, PNAS, 2001
Binding to Thrombin (DuPont Pharmaceuticals)
- 2543 compounds tested for their ability to bind to a target site on thrombin, a key receptor in blood clotting; 192 “active” (bind well); the rest “inactive”. Training set (1909 compounds) more depleted in active compounds.
- 139,351 binary features, which describe three-dimensional properties of the molecule.

*Weston et al, Bioinformatics, 2002*
Text Filtering

Top 3 words of some categories:

- **Alt.atheism**: atheism, atheists, morality
- **Comp.graphics**: image, jpeg, graphics
- **Sci.space**: space, nasa, orbit
- **Soc.religion.christian**: god, church, sin
- **Talk.politics.mideast**: israel, armenian, turkish
- **Talk.religion.misc**: jesus, god, jehovah

**Reuters**: 21578 news wire, 114 semantic categories.

**20 newsgroups**: 19997 articles, 20 categories.

**WebKB**: 8282 web pages, 7 categories.

**Bag-of-words**: >100000 features.

_Bekkerman et al, JMLR, 2003_
Face Recognition

- Male/female classification
- 1450 images (1000 train, 450 test), 5100 features (images 60x85 pixels)

Relief:

Simba:

Navot-Bachrach-Tishby, ICML 2004
Nomenclature

- **Univariate method**: considers one variable (feature) at a time.
- **Multivariate method**: considers subsets of variables (features) together.
- **Filter method**: ranks features or feature subsets independently of the predictor (classifier).
- **Wrapper method**: uses a classifier to assess features or feature subsets.
Univariate Filter Methods
Individual Feature Irrelevance

\[
P(X_i, Y) = P(X_i) P(Y) \\
P(X_i | Y) = P(X_i) \\
P(X_i | Y=1) = P(X_i | Y=-1)
\]
Individual Feature Relevance

- $\mu - \mu^+$
- $\sigma^-$
- $\sigma^+$

ROC curve

Sensitivity

1 - Specificity

AUC

$\mathbf{R_i}$
\[ S_{2N} = \frac{|\mu^+ - \mu^-|}{\sigma^+ + \sigma^-} \]

\[ S_{2N} \cong R \sim x \cdot y \]

after “standardization” \( x \leftarrow (x - \mu_x)/\sigma_x \)
Univariate Dependence

• Independence:
  \[ P(X, Y) = P(X) \cdot P(Y) \]

• Measure of dependence:
  \[
  \text{MI}(X, Y) = \int P(X,Y) \log \frac{P(X,Y)}{P(X)P(Y)} \, dX \, dY
  = KL( P(X,Y) \parallel P(X)P(Y) )
  \]
Correlation and MI

\[ R = 0.02 \]
\[ \text{MI} = 1.03 \text{ nat} \]

\[ R = 0.002 \]
\[ \text{MI} = 1.65 \text{ nat} \]
Gaussian Distribution

\[
\text{MI}(X, Y) = -\frac{1}{2} \log(1 - R^2)
\]
### Other criteria (chap. 3)

<table>
<thead>
<tr>
<th>Method</th>
<th>Formula</th>
<th>$X$</th>
<th>$Y$</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayesian accuracy</td>
<td>Eq. 3.1</td>
<td>$+$</td>
<td>$+$</td>
<td>Theoretically the golden standard, rescaled Bayesian relevance Eq. 3.2.</td>
</tr>
<tr>
<td>Balanced accuracy</td>
<td>Eq. 3.4</td>
<td>$+$</td>
<td>$+$</td>
<td>Average of sensitivity and specificity; used for unbalanced dataset, same as AUC for binary targets.</td>
</tr>
<tr>
<td>Bi-normal separation</td>
<td>Eq. 3.5</td>
<td>$+$</td>
<td>$+$</td>
<td>Used in information retrieval.</td>
</tr>
<tr>
<td>F-measure</td>
<td>Eq. 3.7</td>
<td>$+$</td>
<td>$+$</td>
<td>Harmonic of recall and precision, popular in information retrieval.</td>
</tr>
<tr>
<td>Odds ratio</td>
<td>Eq. 3.5</td>
<td>$+$</td>
<td>$+$</td>
<td>Popular in information retrieval.</td>
</tr>
<tr>
<td>Means separation</td>
<td>Eq. 3.10</td>
<td>$+$</td>
<td>$+$</td>
<td>Based on two class means, related to Fisher’s criterion.</td>
</tr>
<tr>
<td>T-statistics</td>
<td>Eq. 3.11</td>
<td>$+$</td>
<td>$+$</td>
<td>Based also on the means separation.</td>
</tr>
<tr>
<td>Pearson correlation</td>
<td>Eq. 3.9</td>
<td>$+$</td>
<td>$+$</td>
<td>Linear correlation, significance test Eq. 3.12, or a permutation test.</td>
</tr>
<tr>
<td>Group correlation</td>
<td>Eq. 3.13</td>
<td>$+$</td>
<td>$+$</td>
<td>Pearson’s coefficient for subset of features.</td>
</tr>
<tr>
<td>$\chi^2$</td>
<td>Eq. 3.8</td>
<td>$+$</td>
<td>$+$</td>
<td>Results depend on the number of samples $m$.</td>
</tr>
<tr>
<td>Relief</td>
<td>Eq. 3.15</td>
<td>$+$</td>
<td>$+$</td>
<td>Family of methods, the formula is for a simplified version ReliefX, captures local correlations and feature interactions.</td>
</tr>
<tr>
<td>Separability Split Value</td>
<td>Eq. 3.41</td>
<td>$+$</td>
<td>$+$</td>
<td>Decision tree index.</td>
</tr>
<tr>
<td>Kolmogorov distance</td>
<td>Eq. 3.16</td>
<td>$+$</td>
<td>$+$</td>
<td>Difference between joint and product probabilities.</td>
</tr>
<tr>
<td>Bayesian measure</td>
<td>Eq. 3.16</td>
<td>$+$</td>
<td>$+$</td>
<td>Same as Vajda entropy Eq. 3.23 and Gini Eq. 3.39.</td>
</tr>
<tr>
<td>Kullback-Leibler divergence</td>
<td>Eq. 3.20</td>
<td>$+$</td>
<td>$+$</td>
<td>Equivalent to mutual information.</td>
</tr>
<tr>
<td>Jeffreys-Matusita distance</td>
<td>Eq. 3.22</td>
<td>$+$</td>
<td>$+$</td>
<td>Rarely used but worth trying.</td>
</tr>
<tr>
<td>Value Difference Metric</td>
<td>Eq. 3.22</td>
<td>$+$</td>
<td>$+$</td>
<td>Used for symbolic data in similarity-based methods, and symbolic feature-feature correlations.</td>
</tr>
<tr>
<td>Mutual Information</td>
<td>Eq. 3.29</td>
<td>$+$</td>
<td>$+$</td>
<td>Equivalent to information gain Eq. 3.30.</td>
</tr>
<tr>
<td>Information Gain Ratio</td>
<td>Eq. 3.32</td>
<td>$+$</td>
<td>$+$</td>
<td>Information gain divided by feature entropy, stable evaluation.</td>
</tr>
<tr>
<td>Symmetrical Uncertainty</td>
<td>Eq. 3.35</td>
<td>$+$</td>
<td>$+$</td>
<td>Low bias for multivalued features.</td>
</tr>
<tr>
<td>J-measure</td>
<td>Eq. 3.36</td>
<td>$+$</td>
<td>$+$</td>
<td>Measures information provided by a logical rule.</td>
</tr>
<tr>
<td>Weight of evidence</td>
<td>Eq. 3.37</td>
<td>$+$</td>
<td>$+$</td>
<td>So far rarely used.</td>
</tr>
<tr>
<td>MDL</td>
<td>Eq. 3.38</td>
<td>$+$</td>
<td>$+$</td>
<td>Low bias for multivalued features.</td>
</tr>
</tbody>
</table>
T-test

• Normally distributed classes, equal variance $\sigma^2$ unknown; estimated from data as $\sigma^2_{\text{within}}$.

• Null hypothesis $H_0$: $\mu^+ = \mu^-$

• T statistic: If $H_0$ is true,

\[ t = \frac{(\mu^+ - \mu^-)}{(\sigma_{\text{within}} \sqrt{1/m^+ + 1/m^-})} \sim \text{Student}(m^+ + m^- - 2 \text{ d.f.}) \]
Statistical tests (chap. 2)

- $H_0$: $X$ and $Y$ are independent.
- Relevance index $\Leftrightarrow$ test statistic.
- $P$-value $\Leftrightarrow$ false positive rate $\text{FPR} = \frac{n_{fp}}{n_{irr}}$
- Multiple testing problem: use Bonferroni correction $\text{pval} \leq \frac{n_{pval}}{n_{p}}$
- False discovery rate: $\text{FDR} = \frac{n_{fp}}{n_{sc}} \leq \frac{\text{FPR} n}{n_{sc}}$
- Probe method: $\text{FPR} \approx \frac{n_{sp}}{n_{p}}$
Multivariate Methods
Univariate selection may fail

Guyon-Elisseeff, JMLR 2004; Springer 2006
Filters vs. Wrappers

- **Main goal**: rank subsets of useful features.

- **Danger of over-fitting** with intensive search!
Search Strategies (chap. 4)

• **Forward selection** or **backward elimination**.
• **Beam search**: keep k best path at each step.
• **GSFS**: generalized sequential forward selection – when \((n-k)\) features are left try all subsets of \(g\) features i.e. \(\binom{n-k}{g}\) trainings. More trainings at each step, but fewer steps.
• **PTA(l,r)**: plus \(l\), take away \(r\) – at each step, run SFS \(l\) times then SBS \(r\) times.
• **Floating search** (SFFS and SBFS): One step of SFS (resp. SBS), then SBS (resp. SFS) as long as we find better subsets than those of the same size obtained so far. Any time, if a better subset of the same size was already found, switch abruptly.
Multivariate FS is complex

Kohavi-John, 1997

N features, $2^N$ possible feature subsets!
Embedded methods

Recursive Feature Elimination (RFE) SVM. *Guyon-Weston, 2000. US patent 7,117,188*
Embedded methods

**Feature subset assessment**

Split data into 3 sets: training, validation, and test set.

1) For each feature subset, train predictor on **training data**.

2) Select the feature subset, which performs best on **validation data**.
   - Repeat and average if you want to reduce variance (cross-validation).

3) Test on **test data**.
**Complexity of Feature Selection**

With high probability:

Generalization_error \( \leq \) Validation_error + \( \varepsilon (C/m_2) \)

<table>
<thead>
<tr>
<th>Method</th>
<th>Number of subsets tried</th>
<th>Complexity C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exhaustive search wrapper</td>
<td>( 2^N )</td>
<td>N</td>
</tr>
<tr>
<td>Nested subsets</td>
<td>( N(N+1)/2 ) or ( N )</td>
<td>( \log N )</td>
</tr>
<tr>
<td>Feature ranking</td>
<td>N</td>
<td></td>
</tr>
</tbody>
</table>

\( m_2 \): number of *validation* examples,  
\( N \): total number of features,  
\( n \): feature subset size.

Try to keep \( C \) of the order of \( m_2 \).
### Examples of FS algorithms

<table>
<thead>
<tr>
<th>Linear</th>
<th>Univariate</th>
<th>Multivariate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T-test, AUC, feature ranking</td>
<td>RFE with linear SVM or LDA</td>
</tr>
<tr>
<td>Non-linear</td>
<td>Mutual information feature ranking</td>
<td>Nearest Neighbors Neural Nets Trees, SVM</td>
</tr>
</tbody>
</table>
In practice...

• No method is universally better:
  – wide variety of types of variables, data distributions, learning machines, and objectives.

• Match the method complexity to the ratio M/N:
  – univariate feature selection may work better than multivariate feature selection; non-linear classifiers are not always better.

• Feature selection is not always necessary to achieve good performance.

Book of the NIPS 2003 challenge

Feature Extraction, Foundations and Applications
I. Guyon et al, Eds.
http://clopinet.com/fextract-book