INFO411 Lecture 5 - Feature Selection

Jeremiah Deng

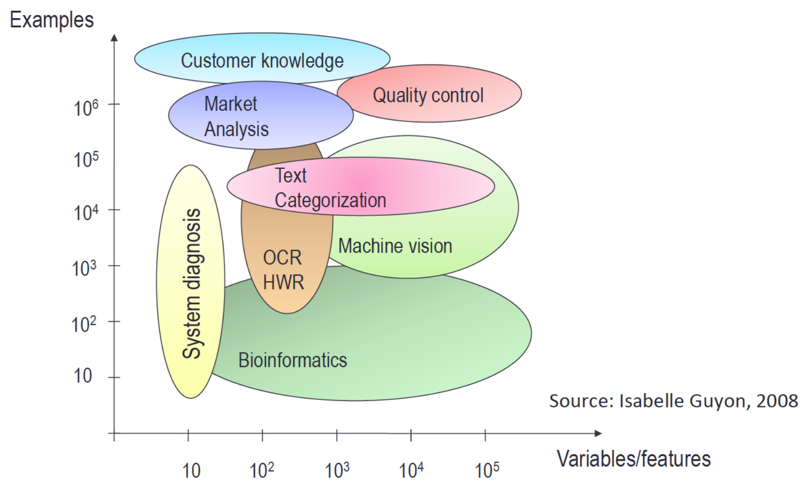
InfoSci, University of Otago

August 8, 2017

Overview

1. Introduction
2. Feature Evaluation
3. Search Strategies
4. Applications

Applications and Their Data



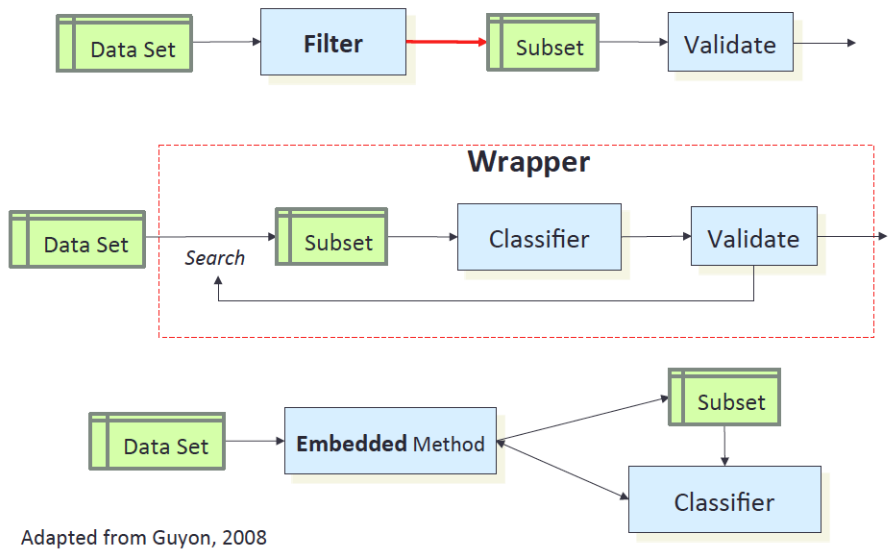
Feature Analysis

- Feature analysis serves to extract meaningful representation from data and maybe further reduce the dimensionality.
- We need certain criterion or mechanism to eventually select from those features and form a new data set for further processing.
- Select the most relevant one to build better, faster, and easy-to-understand computing models.

Major Approaches

- Domain knowledge.
- Statistical. E.g., use LDA to evaluate feature subsets.
- Machine learning methods
  - Filter: use statistical indexes to rank features and find good subsets by choosing the best features
  - Wrapper: use classifiers to validate the effectiveness of selected feature subsets with the help of stochastic search methods.
  - Embedded: guide search of features *during* classification/prediction

Diagrams



Adapted from Guyon, 2008

The Filter Approach

General steps:

- Construct a metric to evaluate each single feature; select those with the biggest metric values.
- Select individual features that contribute the most to the classification performance.
- Evaluate the selected feature subset

Question: How to evaluate features?

Assessing Features - I

We can calculate Pearson correlation between each attribute X_i and the target variable (Y)

$$R(i) = \frac{\mathrm{cov}(X_i, Y)}{\sqrt{\mathrm{var}(X_i)\mathrm{var}(Y)}}$$

Remove insignicant variables.

- For binary classification problem $Y = \pm 1$.
- For regression tasks, Y takes continuous values.
- Pearson correlation assesses linear correlation only
- Nonlinear transform on the data sometimes necessary

Assessing Features - II

Question: How much information does a feature f_i contribute to the class label c?

- Define mutual information (between X and Y):

$$I(X;Y) = H(Y) - H(Y|X)$$

  or

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

- To rank f_i, we need to calculate $I(f_i, c)$
- Often referred to as 'information gain' (IG)
- Rank $I(f_i, c)$; Choose top-ranked features.
- *Does this make sense?*

Other information theory metrics

- Gain ratio (GR)

$$\mathrm{GR} = \frac{\mathrm{IG}}{\mathrm{H(X)}}$$

- symmetric uncertainty (SU)

$$\mathrm{GR} = \frac{2 \times \mathrm{IG}}{\mathrm{H(X)} + \mathrm{H(Y)}}$$

Best+2nd Best + ... = The Best Subset?

- Often the highest-ranking features get selected into the subset (S)
- This is *often* suboptimal!
  - Redundancy exists in the selections
  - Less important features might be more useful!
- Max dependency: maximal mutual information between attribute selection and class label c:

$$\max D(S,c), D = I(\{x_1, x_2, \cdots, x_m\}; c)$$

☹ Hard to implement!

mRMR

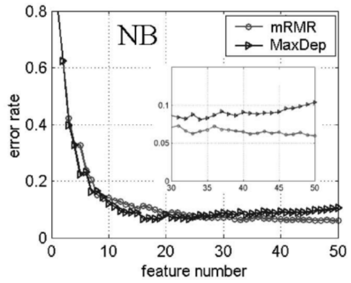
- minimal-redundancy-Maximal-relevance (mRMR)
- Peng et al., PAMI, 2005
- Maximize relevance

$$D = \frac{1}{|S|} \sum_{f_i \in S} I(f_i, c)$$

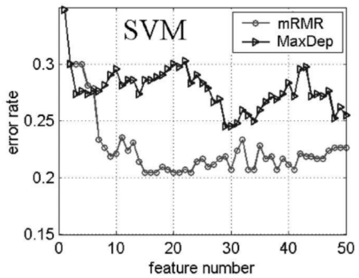
- Minimize redundancy

$$R = \frac{1}{|S|^2} \sum_{f_i, f_j \in S} I(f_i, f_j)$$

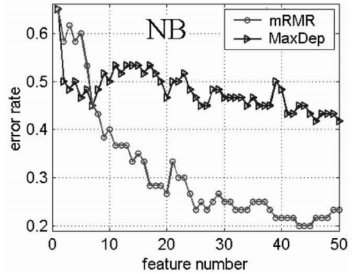
- mRMR: Find optimal S that maximizes $D - R$ or D/R



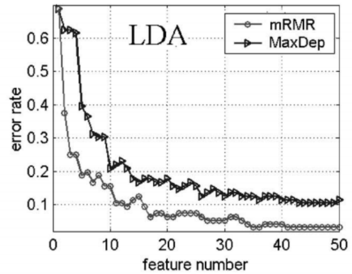
(a)



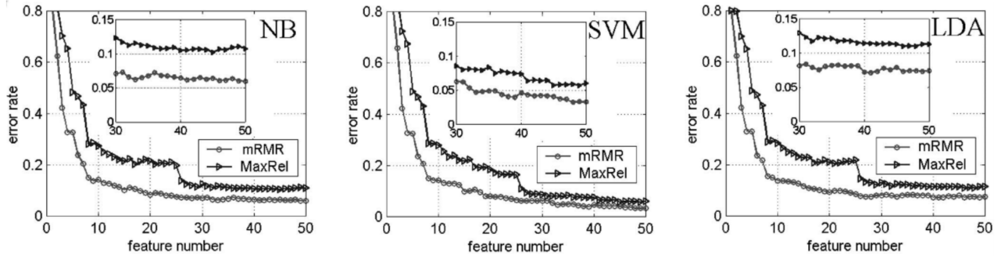
(b)



(c)



(d)



mRMR outperforms MaxDep and MaxRel in cross-validation evaluation on a number of UCI datasets.

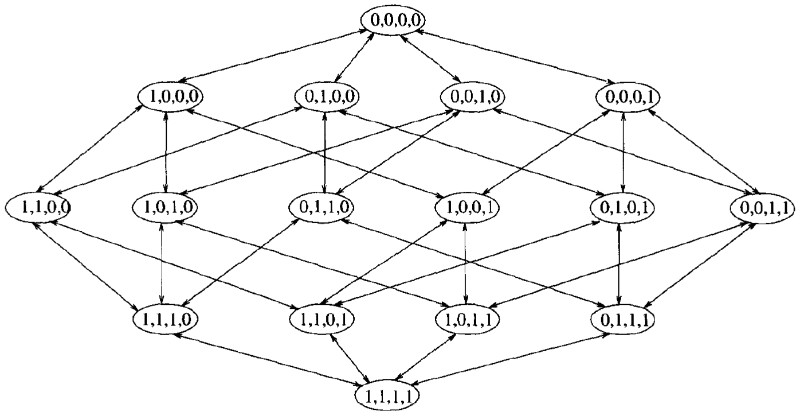
The Wrapper Approach

- Multivariate Feature Selection implies a search in the space of all possible combinations of features.
- For n features, there are 2^n possible subsets of features, yielding both to a high computational and statistical complexity.
- Wrappers use the performance of a learning machine to evaluate each subset.
- Training 2^n learning machines is infeasible for large n, so most wrapper algorithms resort to greedy or heuristic search.

Search Strategies

- ☹ Exhaustive search.
- Stochastic search: simulated annealing, genetic algorithm
- Greedy search
  - Sequential forward selection (SFS)
  - Sequential backward elimination (SBS)
  - $PTA(l, r)$: plus l , take away r, i.e. at each step, run SFS l times then SBS r times.

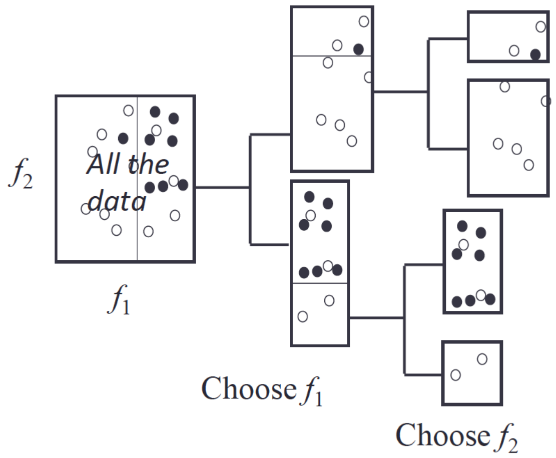
Search Strategies ...



Kohavi & John, Artificial Intelligence 97, 1997

- Traversing the state space

Decision Tree as an example



Decision trees use SFS: at each step, choose the feature that "reduces entropy" most. Work towards "node purity".

Discriminant Analysis

- We can also adopt the mechanism in Fisher's Linear Discriminant Analysis (LDA)
- Calculates within-class scatter and between-class scatter.
  - For classification tasks it is to find a linear projection that produces big between-class scatter and small within-class scatter
  - In feature selection, it is only to find a subset of features that generates the biggest between-class scatter and the smallest within-class scatter.

Relief

- Kira and Rendell (1992); Kononenko (1994)
- For each attribute, compares attribute value with those of the near-hit (nearest neighbour of the same class) and near-miss (NN of the other class), and tunes the weight on the attribute
- Iterate through dataset randomly for T iterations

Relief algorithm

Require: Dataset X
  return weighting of attributes $\mathbf{w}$
  $\mathbf{w} \leftarrow 0$
  for $t = 1$ **to** T **do**
    Pick $\mathbf{x}$ randomly from X
    Find near-hit $\mathbf{h}$ and near-miss $\mathbf{m}$
    $\Delta w_i \leftarrow (x_i - m_i)^2/T - (x_i - h_i)^2/T$, $i = 1, \cdots, D$
  end for

Laplacian Scores

- He et al., NIPS 2005
- Based on Laplacian Eigenmaps and Locality Preserving Projection.
- Basic idea: evaluate the features according to their locality preserving power.

Locality Preserving Projection

- He & Niyogi, NIPS 2003
- Given data $\mathbf{x}_1, \cdots, \mathbf{x}_m \in \mathbb{R}^n$, find a mapping matrix A that maps these m points to $\mathbf{y}_1, \cdots, \mathbf{y}_m \in \mathbb{R}^d (d < n)$, such that $\mathbf{y}_i$ represents $\mathbf{x}_i$, where $\mathbf{y}_i = A^T \mathbf{x}_i$.

1. Construct adjacency graph of m nodes, e.g. finding k-NNs
2. Choose connection weights, e.g., Gaussian, or binary
3. Solve the generalized eigen-vector problem:

$$XLX^T\mathbf{a} = \lambda XDX^T\mathbf{a}$$

where D is a diagonal matrix with $D_{ii} = \sum_j W_{ji}$. Laplacian $L = D - W$, X is formed by $\mathbf{x}_i$ (as column vector).

4. $A = (\mathbf{a}_1, \cdots, \mathbf{a}_d), \mathbf{a}_i$ sorted by eigenvalues $\lambda_1 < \cdots < \lambda_d$.

AI and Music

Asimo conducts DSO

Deng, Simmermacher & Cranefield: A study on feature analysis for musical instrument classification. IEEE Trans. Systems, Man, and Cybernetics, Part B (2008)

Audio Feature Extraction

- MFCC- Auditory Toolbox
  - Based on the Mel-scale, typically used for speech detection
  - First 13 linear values
- Auditory Model- IPEM Toolbox
  - Auditory Nerve Image, simulation of ear filtering
  - Centroid, Bandwidth, Flux, Zero-Crossings, etc.
- MPEG-7 Harmonic Instrument Timbre DS
  - Scheme from out of 18 standardised features
  - Harmonic Centroid, -Deviation, -Spread, -Variation, Temporal and Spectral Centroid, Log-Attack-Time

Features Considered

Feature #	Category	Description
1	Perception-based	Zero Crossings (ZC)
2-3		Mean and standard deviation of Zero Crossing Ratio (ZCR)
4-5		Mean and standard deviation of RMS (volume root mean square)
6-7		Mean and standard deviation of Centroid
8-9		Mean and standard deviation of Bandwidth
10-11		Mean and standard deviation of Flux
12	MPEG-7	Harmonic Centroid (HC)
13		Harmonic Deviation (HD)
14		Harmonic Spread (HS)
15		Harmonic Variation (HV)
16		Spectral Centroid (SC)
17		Temporal Centroid (TC)
18		Log-Attack-Time (LAT)
19-**44**	MFCC	Mean and standard deviation of the first 13 linear MFCC

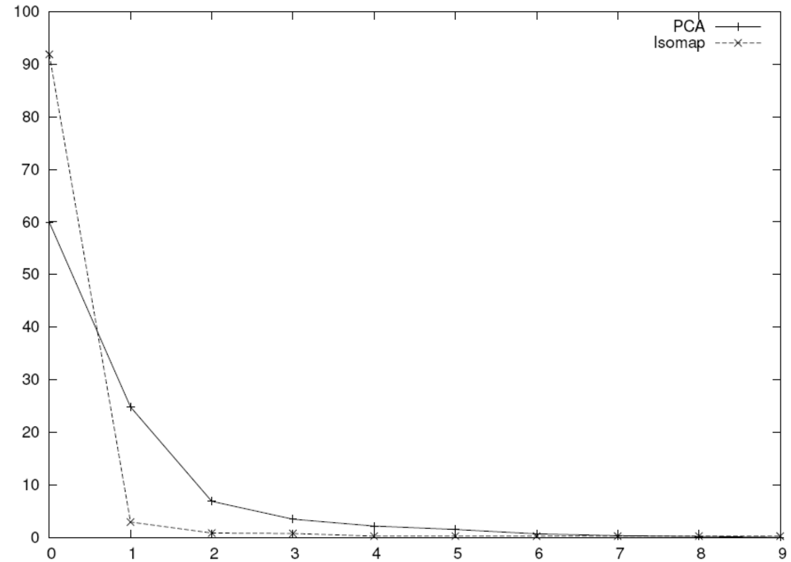
Questions Asked

- How useful are these features in instrument recognition?
  - Evaluation criteria
  - Ranking
  - Subset selection
- Can a subset be selected to produce satisfactory classification rates?
- Does it matter using a different classifier?

Methods

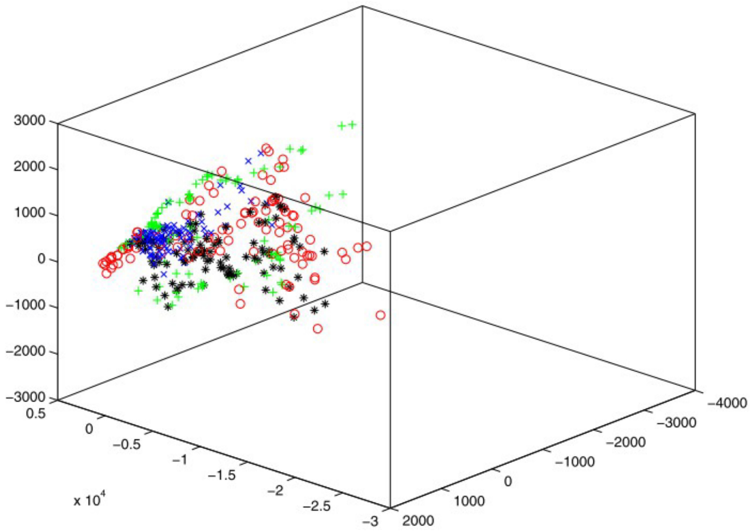
- Dimension reduction
  - Principal component analysis
  - Isomap
- Feature selection/ranking
  - Entropy-based: Information Gain, Gain Ratio, Symmetric Uncertainty
  - SVM attribute evaluation
- Classification
  - k-Nearest Neighbours
  - Naïve Bayes
  - Support Vector Machine
  - Radial Basis Functions

Dimension Reduction: Residuals



The effective dimensionality could be as small as 3.

Visualization: 3-D Isomap



Feature Ranking Results

Rank	IG		GR		SU		SVM
	Feature	*Value*	*Feature*	*Value*	*Feature*	*Value*	*Feature*
1	LAT	0.8154	LAT	0.5310	LAT	0.4613	HD
2	HD	0.6153	HD	0.5270	HD	0.3884	FluxD
3	FluxD	0.4190	MFCC2M	0.3230	BandwidthM	0.2267	LAT
4	BandwidthM	0.3945	MFCC12D	0.2970	FluxD	0.2190	MFCC3D
5	MFCC1D	0.3903	MFCC4D	0.2700	RMSM	0.2153	MFCC4M
6	MFCC3D	0.381	BandwidthM	0.2660	MFCC1D	0.2084	ZCRD
7	RMSM	0.3637	RMSM	0.2640	MFCC4M	0.1924	MFCC1M
8	BandwidthD	0.3503	MFCC13D	0.2580	MFCC11D	0.1893	HC
9	MFCC4M	0.3420	MFCC2D	0.2450	MFCC3D	0.1864	MFCC9D
10	MFCC11D	0.3125	MFCC11D	0.2400	BandwidthD	0.1799	ZC
11	ZCRD	0.3109	MFCC7D	0.2350	MFCC2M	0.1784	RMSM
12	CentroidD	0.2744	FluxD	0.2290	MFCC4D	0.1756	CentroidD
13	MFCC8D	0.2734	MFCC1D	0.2240	MFCC7D	0.1710	MFCC9M
14	MFCC6D	0.2702	MFCC4M	0.2200	MFCC12D	0.1699	BandwidthM
15	MFCC7D	0.2688	CentroidM	0.2150	ZCRD	0.1697	MFCC5D
16	ZC	0.2675	SC	0.2110	CentroidD	0.1653	SC
17	MFCC4D	0.2604	MFCC5M	0.2090	CentroidM	0.1610	MFCC12D
18	CentroidM	0.2578	CentroidD	0.2080	MFCC13D	0.1567	MFCC7M
19	MFCC10M	0.2568	HC	0.1950	SC	0.1563	MFCC2M
20	MFCC10D	0.2519	MFCC1M	0.1910	MFCC8D	0.1532	MFCC6M

Performance

CLASSIFIER PERFORMANCE (IN PERCENTAGE) OF THE INSTRUMENT FAMILIES

Feature Scheme	*k*-NN	Naive Bayes	SVM	MLP	RBF
All 44	95.75	86.5	97.0	95.25	95.0
Best 20	94.25	86.25	95.5	93.25	95.5
Best 10	90.25	86.25	94.25	91.0	87.0
Best 5	89.5	81.0	91.75	86.75	84.5

PERFORMANCE (IN PERCENTAGE) IN CLASSIFYING THE FOUR CLASSES (TENFOLD CROSS-VALIDATION)

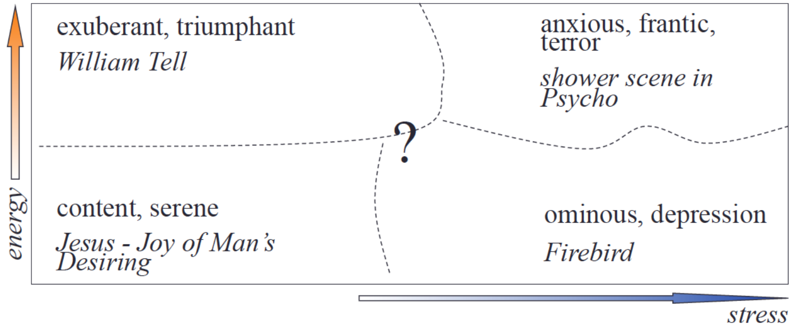
Feature Sets	Brass	Woodwind	String	Piano	Overall
MFCC (26)	99	90	89	95	93.25
MPEG-7 (7)	90	62	76	99	81.75
IPEM (11)	93	63	81	100	84.25
MFCC+MPEG-7 (33)	98	92	91	100	95.25
MFCC+IPEM (37)	98	89	94	98	94.75
IPEM+MPEG-7(18)	93	76	85	100	88.5
Top 50% mix (21)	95	89	88	100	93
Best 20	97	88	92	100	94.25
Selected 17	97	94	95	100	96.5

Questions Answered?

- Classification: random forests, ensembles, boosting ...? *tbc*
- Dictionary learning?
- Mairal et al., Online dictionary learning for sparse coding (ICML'09)
- Search for sparse representation α based on dictionary $\mathbf{D}$

$$\min_{D,\alpha} \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{2} \|\mathbf{x}_i - \mathbf{D}\alpha_i\|_2^2 + \lambda \|\alpha_i\|_1 \right)$$

More Music-related ML topics



- Music mood recognition
- Automatic music transcription
  - Pitch, notes
  - Expressions
- Music composition
- Interpreter recognition

Alternative Feature Extraction?

Or, *how to win a Kaggle competition?*
Feature Engineering Without Domain Expertise
https://www.youtube.com/watch?v=bL4b1sGnILU

Recap

- Feature selection vs feature extraction
- Main approaches: filter, wrapper, embedded search
- Feature evaluation criteria
- Search strategies

- Coming next: Presentation, Project ...