Information Theoretic Feature Selection for Clustering

Satyavarta

EC517 Information Theory
Boston University

Term Project Presentation, Apr 30, 2008
The Problem of Classification

Hastie, Tibshirani, Friedman, Elements of Statistical Learning, fig 2.5
Good Features: What we need

\[ I(f_1, f_2) \sim \]
\[ I(f_1, f_2 | I) = \text{LOW} \]

\[ I(f_2, L) = \text{HIGH} \]

Hastie, Tibshirani, Friedman, Elements of Statistical Learning, fig 2.5
Good Features: What we got

\[ I(f_1, L) = \text{LOW} \]
\[ I(f_1, f_2) = \sim \]
\[ I(f_1, f_2|I) = \text{LOW} \]
\[ I(f_2, L) = \text{HIGH} \]

Hastie, Tibshirani, Friedman, Elements of Statistical Learning, fig 2.5
Benefits of Using Information Theory

- Captures dependence beyond second order statistics
  \[ I(f_1, f_2) \text{ is KL divergence} \]
- Invariant to monotonic transformations on variables
  \[ I(f_1, f_2) = I(\phi(f_1), \phi(f_2)) \]
- Independent of decision algorithm → limits on separability
  Perfect reconstruction: \[ I(f, L) = H(L) \]
Case: Both Cues are Good

\[ I(f_1, L) = 0.269 \]

\[ I(f_1, f_2) = 3.003 \]

\[ I(f_1, f_2|L) = 4.886 \]

\[ I(f_2, L) = 0.653 \]

Anechoic, 60', fm
Case: Both Cues are Degraded

\[ I(f_1, L) = 0.273 \]

\[ I(f_1, f_2) = 2.742 \]

\[ I(f_1, f_2 | l) = 5.128 \]

\[ I(f_2, L) = 0.209 \]
Case: Interaural Cues are Minimized

\[ I(f_1, L) = 0.289 \]

\[ I(f_1, f_2) = 2.453 \]
\[ I(f_1, f_2 | l) = 5.000 \]

Reverberant 0° fm

\[ I(f_2, L) = 0.060 \]
Case: Pitch Cues are Minimized

\[ I(f_1, L) = 0.232 \]

\[ I(f_1, f_2) = 2.358 \]

\[ I(f_1, f_2 | l) = 4.807 \]

Reverberant 60', mm

\[ I(f_2, L) = 0.213 \]

Satya

Info Theoretic Feature Selection
Case: Three sources, cues need to work together

\[ I(f_1, L) = 0.296 \]

\[ I(f_1, f_2) = 2.591 \]

\[ I(f_1, f_2 | l) = 4.189 \]

\[ I(f_2, L) = 0.419 \]
Conclusions

- Confirms what is known for this problem
- Computation of multi-variate density $I(f_1, f_2, ...)$ is a bear
- Difficult to compare across features
  - Mutual information is not a metric
  - Mutual information depends on # bins in discretized data
- Feature selection straightforward, feature weighting not so direct