Topics In Feature Selection

CSI 5388
Theme Presentation

Joe Burpee
2005/2/16

Feature Selection (FS)

- aka “Attribute Selection”
- Witten and Frank book
  - Section 7.1
- Liu site http://athena.csee.umbc.edu/IDM02/
  - “Toward a Unifying Taxonomy for Feature Selection”
  - efforts at an integrated framework with some standard nomenclature (ca. 2002)

Feature Preprocessing

- FS: Feature Selection reduces the feature set $A=\{A_1,\ldots,A_m\}$ to a proper subset $A'\subset A$.
- FE: Feature Extraction reduces the feature set A to a derived set $B=F(A)$, where usually $|B|<|A|$.
  - e.g. principal component analysis (PCA)
- FC: Feature Construction augments A, e.g.
  - numeric: polynomials
  - nominal: interactions

FS Methods

Used separately or in combination:

- Manual: it is often very important for the user to control the selection process, e.g. to specify an equation consistent with theory
- Implicit: learning schemes typically perform some degree of FS; e.g. feature weighting in Instance-based learning, tree pruning, etc.
- Algorithm: for reasons of very high dimensionality, exploratory data analysis, etc.

FS Performance

- supervised
  - typically *accuracy*
  - estimation requires separate data
- unsupervised
  - other criteria

Dimensionality Reduction

Drop attributes that are ***redundant*** (related to other attributes)

Keep attributes that are most ***relevant*** (related to the class variable)

Reasons:

- efficiency of learning (search)
- performance of learned classifier
  - optimality of search
  - avoidance of overfitting to meaningless random effects (cf PAC rationale)
  - stability, given available data
- interpretability of learned model, meaningful theory

Data Subsets

To counteract overfitting:

- Three instance sets are needed
  - **training** set
  - **validation** set (e.g. for FS)
  - **test** set (final model with selected features)
- or alternatively, nested cross-validations
  - e.g. 10 x 10 fold

General- or Special-Purpose FS

- ***Wrapper***: scheme-specific
  - typically supervised-learning performance is measured just by trying the chosen scheme on each subset
    - may be expensive -- many cross-validations
- ***Filter***: scheme-independent
  - one way is to search for smallest feature subset that separates the classes
    - expensive
    - noise causes overfitting
  - or use different (inexpensive) learning scheme to select features for scheme of interest

Feature Space Search for Wrapper

- m features => 2^m subsets
- heuristics to reduce cost
  - forward selection: greedy, $m(m$-1)/2 subsets
  - backward elimination
  - bidirectional, stepwise
  - best-first (can retry any previous subsets)
  - beam search (limited list of subsets)
  - genetic algorithm

Filter Examples

<u>Instance-based</u> learner as a filter:

- Weighted distance function

  $[\, w_1(x_1 - y_1)^2 + \ldots + w_m(x_m - y_m)^2 \,]^{1/2}$

- Relevance = magnitude of weight
- Can use this for FS for a scheme of interest
- But it cannot detect redundant attributes

Filter Examples (2)

Decision Tree builder as filter for Instance-based Learning:

- Nearest-Neighbour (unweighted) metric

  $[\ (x_1 - y_1)^2 + \ldots + (x_m - y_m)^2\ ]^{1/2}$

- NN sensitive to irrelevant features; DT is less so
- Filtered NN may perform better than DT alone
- Similarly One-Rule may be used as filter for DT

Wrapper Examples

- Naïve Bayes is quite sensitive to redundancy (which violates the conditional independence assumption)

$$\mathrm{P}(c|\mathbf{A}) \propto \mathrm{P}(c) \cdot \mathrm{P}(\mathbf{A}|c) \simeq \mathrm{P}(c) \cdot \mathrm{P}(A_1|c) \cdot \cdot \cdot \mathrm{P}(A_m|c)$$

- E.g. inserting the same attribute twice, squares the probability
- “Selective Naïve Bayes” uses Forward Selection to avoid adding a variable that contributes nothing to performance on the training set

Wrapper Examples (2)

- Linear Regression: not strictly ML, but heuristics like Forward Selection, etc. are common
- Sensitive to redundancy (multicollinearity)
- Performance on the training data typically measured by R^2 which is numeric version of

  1 - (*squared-error rate*)
- Statistical significance determined using F- or t-tests (functions of R) in analysis of variance (ANOVA)
- Nonlinear models use asymptotic analogues