Molecular Biomedical Informatics

分 子 生 醫 資 訊 實 驗 室

Machine Learning and Bioinformatics

機 器 學 習 與 生 物 資 訊 學

Feature seleciton

Related issues

- Feature selection
  - scheme independent/specific
- Feature discretization
- Feature transformations
  - Principal Component Analysis (PCA), text and time series
- Dirty data (data cleaning and outlier detection)
- Meta-learning
  - bagging (with costs), randomization, boosting, …
- Using unlabeled data
- Clustering for classification, co-training and EM

- Engineering the input and output

Just apply a learner?

- Please DON'T
- As scheme/parameter selection
  - treat selection process as part of the learning process
- Modifying the input
  - data engineering to make learning possible or easier
- Modifying the output
  - combining models to improve performance

Feature selection

- Adding a random (i.e. irrelevant) attribute can significantly degrade C4.5's performance
  - attribute selection based on smaller and smaller amounts of data

why

- Instance-based learning is very susceptible to irrelevant attributes
  - number of training instances required increases exponentially with number of irrelevant attributes
- Relevant attributes can also be harmful

“What’s the difference between theory and practice?”
an old question asks.
“There is no difference,”
the answer goes,
“—in theory. But in practice, there is.”

Scheme-independent selection

- Assess based on general characteristics (relevance) of the feature
- Find smallest subset of features that separates data
- Use different learning scheme
  - e.g. use attributes selected by a decision tree for KNN
- KNN can also select features
  - weight features according to “near hits” and “near misses”

Redundant (but relevant) features

- Correlation-based Feature Selection (CFS)
  - correlation between attributes measured by symmetric uncertainty

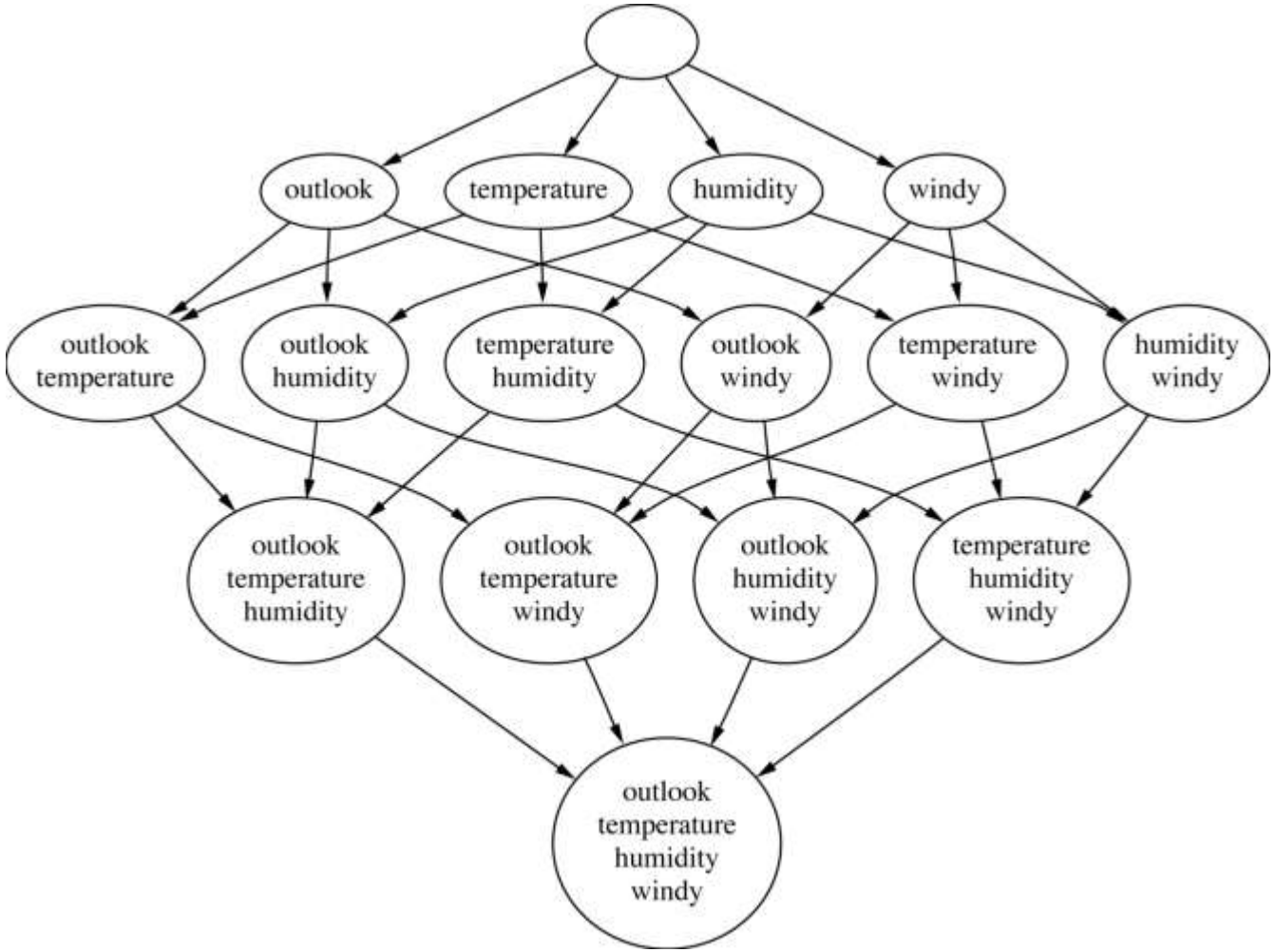
$$U(A,B) = 2\frac{H(A) + H(B) - H(A,B)}{H(A) + H(B)},$$

    where H is the entropy function
  - goodness of subset of features measured by

$$\sum_j U(A_j, C) / \sqrt{\sum_i \sum_j U(A_i, A_j)},$$

    where C is the class

Feature subsets for weather data



Searching feature space

- Number of feature subsets is
  - exponential in number of features
- Common greedy approaches
  - forward selection
  - backward elimination
- More sophisticated strategies
  - bidirectional search
  - best-first search ← can find optimum solution
  - beam search ← approximation to best-first search
  - genetic algorithms

Scheme-specific selection

- Wrapper approach to attribute selection
  - implement "wrapper" around learning scheme
  - evaluate by cross-validation performance
- Time consuming
  - prior ranking of attributes
- Can use significance test to stop branches if it is unlikely to "win" (race search)
  - can be used with forward, backward selection, prior ranking or special purpose schemata search

Feature selection

itself is a research topic in machine learning

Random forest

Random forest

- Breiman (1996,1999)
- Classification and regression Algorithm
- Bootstrap aggregation of classification trees
- Attempt to reduce bias of single tree
- Cross-validation to assess misclassification rates
  - out-of-bag (OOB) error rate
- Permutation to determine feature importance
- Assumes all trees are independent draws from an identical distribution, minimizing loss function at each node in a given tree– randomly drawing data for each tree and features for each node

The algorithm

- Bootstrap sample of data
- Using 2/3 of the sample, fit a tree to its greatest depth determining the split at each node through minimizing the loss function considering a random sample of covariates (size is user specified)
- For each tree
  - predict classification of the leftover 1/3 using the tree, and calculate the misclassification rate ← OOB error rate
  - for each feature in the tree, permute the feature values and compute the OOB error, compare to the original OOB error, the increase is a indication of the feature's importance
- Aggregate OOB error and importance measures from all trees to determine overall OOB error rate and feature Importance measure

Today’s exercise

Feature selection

Uses feature selection tricks to refine your feature program. Upload and test them in our simulation system. Finally, commit your best version and send TA Jang a report before 23:59 11/19 (Mon).