Goals for the lecture

you should understand the following concepts

• filtering-based feature selection
• information gain filtering
• Markov blanket filtering
• frequency pruning
• wrapper-based feature selection
• forward selection
• backward elimination
• dimensionality reduction
Irrelevant and redundant features can lead to incomprehensible models and poor performance.

**Florida To Experiment With New 600-Lever Voting Machines**

Motivation for feature selection

1. We want models that we can interpret. We’re specifically interested in which features are relevant for some task.

2. We’re interested in getting models with better predictive accuracy, and feature selection may help.

3. We are concerned with efficiency. We want models that can be learned in a reasonable amount of time, and/or are compact and efficient to use.
Motivation for feature selection

• some learning methods are sensitive to irrelevant or redundant features
  • \textit{k}-NN
  • naïve Bayes
  • etc.

• other learning methods are ostensibly insensitive to irrelevant features (e.g. Weighted Majority) and/or redundant features (e.g. decision tree learners)

• empirically, feature selection is sometimes useful even with the latter class of methods [Kohavi & John, \textit{Artificial Intelligence} 1997]

Feature selection approaches

filtering-based
feature selection

all features

feature selection

subset of features

learning method

model

wrapper-based
feature selection

all features

feature selection

calls learning method \textit{many times}, uses it to help select features

model

learning method
Information gain filtering

• select only those features that have significant information gain (mutual information with the class variable)

\[
\text{InfoGain}(Y, X_i) = H(Y) - H(Y \mid X_i)
\]

- entropy of class variable (in training set)
- entropy of class variable given feature \( X_i \)

• unlikely to select features that are highly predictive only when combined with other features
• may select many redundant features

Markov blanket filtering

[Koller & Sahami, ICML 1996]

• a Markov blanket \( M_i \) for a variable \( X_i \) is a set of variables such that all other variables are conditionally independent of \( X_i \) given \( M_i \)
• we can try to find and remove features that minimize the criterion:

\[
\Delta(X_i, M_i) = \sum_{x_{M_i} \cdot x_i} \left[ P(M_i = x_{M_i}, X_i = x_i) \times D_{KL}(P(Y \mid M_i = x_{M_i}, X_i = x_i) \parallel P(Y \mid M_i = x_{M_i})) \right]
\]

- Kullback-Leibler divergence (distance between 2 distributions)
• if \( Y \) is conditionally independent of feature \( X_i \) given a subset of other features, we should be able to omit \( X_i \)
Bayes net view of a Markov blanket

\[ P(X_i \mid M_i, Z) = P(X_i \mid M_i) \]

- the Markov blanket \( M_i \) for variable \( X_i \) consists of its parents, its children, and its children's parents

- but we know that finding the best Bayes net structure is NP-hard; can we find approximate Markov blankets efficiently?

Heuristic method to find an approximate Markov blanket

\[
\Delta(X_i, M_i) = \sum_{x_{M_i}, x_i} P(M_i = x_{M_i}, X_i = x_i) \times D_{KL} \left( P(Y \mid M_i = x_{M_i}, X_i = x_i) \mid \mid P(Y \mid M_i = x_{M_i}) \right)
\]

// initialize feature set to include all features
\( F = X \)

iterate

for each feature \( X_i \) in \( F \)

let \( M_i \) be set of \( k \) features most correlated with \( X_i \)

compute \( \Delta(X_i, M_i) \)

choose the \( X_r \) that minimizes \( \Delta(X_r, M_r) \)

\( F = F - \{ X_r \} \)

return \( F \)
Another filtering-based method:  
*frequency pruning*

- remove features whose value distributions are highly skewed
- common to remove very high-frequency and low-frequency words in text-classification tasks such as spam filtering

Some words occur so frequently that they are not informative about a document's class

- the
- be
- to
- of
- ...

Some words occur so infrequently that they are not useful for classification

- accubation
- cacodaemonomania
- echopraxia
- ichneutic
- zoosemiotics
- ...

Example: feature selection for cancer classification

- classification task is to distinguish two types of leukemia: AML, ALL
- 7130 features represent expression levels of genes in tumor samples
- 72 instances (patients)
- three-stage filtering approach which includes information gain and Markov blanket [Xing et al., *ICML 2001*]
Wrapper-based feature selection

- frame the feature-selection task as a search problem
- evaluate each feature set by using the learning method to score it (how accurate of a model can be learned with it?)

Feature selection as a search problem

state = set of features
start state = empty (forward selection)
    or full (backward elimination)

operators
add/subtract a feature

scoring function
training or tuning-set or CV accuracy using learning method on a given state’s feature set
Forward selection

Given: feature set \( \{X_1, \ldots, X_n\} \), training set \( D \), learning method \( L \)

\[
F \leftarrow \{ \}
\]
while score of \( F \) is improving
for \( i \leftarrow 1 \) to \( n \) do
if \( X_i \notin F \)
\[
G_i \leftarrow F \cup \{X_i\}
\]
\[
\text{Score}_i = \text{Evaluate}(G_i, L, D)
\]
\[
F \leftarrow G_b \text{ with best } \text{Score}_b
\]
return feature set \( F \)

scores feature set \( G \) by learning model(s) with \( L \) and assessing its (their) accuracy

Forward selection

\[
\{ \} \quad 50\%
\]
\[
\{X_1\} \quad 50\%
\]
\[
\{X_2\} \quad 51\%
\]
\[
\{X_7\} \quad 68\%
\]
\[
\{X_n\} \quad 62\%
\]
\[
\{X_7, X_1\} \quad 72\%
\]
\[
\{X_7, X_2\} \quad 68\%
\]
\[
\{X_7, X_n\} \quad 69\%
\]
Backward elimination

\[ X = \{ X_1 \ldots X_n \} \]

68%

\[ X - \{ X_j \} \]

65%

\[ X - \{ X_2 \} \]

71%

\[ \ldots \]

\[ X - \{ X_9 \} \]

72%

\[ \ldots \]

\[ X - \{ X_n \} \]

62%

Forward selection vs. backward elimination

- both use a hill-climbing search

**forward selection**

- efficient for choosing a small subset of the features
- misses features whose usefulness requires other features (feature synergy)

**backward elimination**

- efficient for discarding a small subset of the features
- preserves features whose usefulness requires other features
Dimensionality reduction

- *feature selection*: equivalent to projecting feature space to a lower dimensional subspace perpendicular to removed feature

- *dimensionality reduction*: allow other kinds of projection (e.g. PCA re-represents data using linear combinations of original features)

### Dimensionality reduction example

We can represent a face using all of the pixels in a given image (# features = # pixels)

More effective method: represent each face as a linear combination of *eigenfaces* (# features = 25)
Comments on feature selection

- filtering-based methods are generally more efficient
- wrapper-based methods use the inductive bias of the learning method to select features
- forward selection and backward elimination are most common search methods in the wrapper approach, but others can be used [Kohavi & John, *Artificial Intelligence* 1997]
- feature-selection methods may sometimes be beneficial to get
  - more comprehensible models
  - more accurate models
- alternative approach: incorporate feature selection into the learning process (e.g. L₁ regularization)
- dimensionality reduction methods may sometimes lead to more accurate models, but often lower comprehensibility