Feature Selection

Razvan C. Bunescu
School of Electrical Engineering and Computer Science
bunescu@ohio.edu
Feature Selection

- Datasets with thousands of features are common:
  - text documents
  - gene expression data
- Processing thousands of features during training & testing can be computationally infeasible.
- Many irrelevant features can lead to overfitting.

=> select most relevant features in order to obtain faster, better and easier to understand learning models.
Feature Selection: Methods

- **Wrapper method:**
  - uses a classifier to assess features or feature subsets.

- **Filter method:**
  - ranks features or feature subsets independently of the classifier.

- **Univariate method:**
  - considers one feature at a time.

- **Multivariate method:**
  - considers subsets of features together.
The Wrapper Method

**Greedy Forward Selection:**

- $F$ is the set of all features.
- $S \subseteq F$ is the subset of selected features.

1. Start with no features in $S = \{\}$
2. For each feature $f$ in $F - S$, train model with $S + \{f\}$
3. Add to $S$ the best performing feature(s).
4. Repeat from 2 until:
   - (a) performance does not improve, or
   - (b) performance good enough.
The Wrapper Method

**Greedy Backward Elimination:**
- \( F \) is the set of all features.
- \( S \subseteq F \) is the subset of selected features.

1. Start with all features in \( S = F \)
2. For each feature in \( S \), train model without that feature.
3. Remove from \( S \) feature corresponding to best model.
4. Repeat from 2 until:
   - (a) performance does not improve, or
   - (b) performance good enough.
The Wrapper Method

- **Forward**: Greedily add features one (more) at a time. 
  
  *Efficiently Inducing Features of Conditional Random Fields*
  
  [McCallum, UAI’03]

- **Backward**: Greedily remove features one (more) at a time.
  
  *Multiclass cancer diagnosis using tumor gene expression signatures*
  
  [Ramaswamy et al., PNAS’01]

- **Combined**: Two steps forward, one step back.

- Train multiple times ⇒ can be very time consuming!
  
  - Alternative: use external criteria to decide feature relevance ⇒ the Filter Method.
Recursive Feature Elimination with SVM

[Guyon et al., ML’03]

- An instance of Greedy Backward Elimination.

1. Let $F = \{1, 2, ..., K\}$ be the set of features.
2. Let $S = []$ be the ranked set of features.
3. Repeat until $F - S$ is empty:
   I. Train weight vector $w$ using a linear SVM and $F - S$.
   II. Find feature $f$ in $F - S$ with minimum $|w_f|$.
   III. Append $f$ to $S$.
4. Return $S$. 
The Filter Method

1. Rank all features using a measure of correlation with the label.
2. Select top $k$ features to use in the model.

- Measures of correlation between feature $X$ and label $Y$:
  - Mutual Information
  - Chi-square Statistic
  - Pearson Correlation Coefficient
  - Signal-to-Noise Ratio
  - T-test

$\textit{nominal features & label}$
Mutual Information

- Independence:
  \[ P(X, Y) = P(X)P(Y) \]

- Measure of dependence:
  \[
  MI(X, Y) = \sum_{X \in \mathcal{X}} \sum_{Y \in \mathcal{Y}} p(X, Y) \log \frac{p(X, Y)}{p(X)p(Y)}
  = KL(p(X, Y) \parallel p(X)p(Y))
  \]
  - It is 0 when X and Y are independent.
  - It is maximum when X=Y.
Mutual Information

• Problems:
  – Works only with nominal features & labels ⇒ discretization.
  – Biased toward high arity features ⇒ normalization.
  – May choose redundant features.
  – Features may become relevant in the context of other ⇒ use conditional MI [Fleuret, JMLR ‘04].

• Other measures:
  – Chi square ($\chi^2$).
  – Log-likelihood Ratio (LLR).

• Comparison between MI, $\chi^2$, and LLR in [Dunning, CL’98]
  “Accurate methods for the statistics of surprise and coincidence”
Chi Square ($\chi^2$) Test of Independence

- $N$ training examples (observations).
- $X$ is a discrete feature with $k$ possible values.
- $Y$ is a label with $l$ possible values.
- Create $k$-by-$l$ contingency table with cells for every feature-label combination.

$$O_{ij} \quad N_{X=i}$$

$N_{Y=j}$

$\text{need } O_{ij} > 5$
Chi Square ($\chi^2$) Test of Independence

- $O_{ij}$ is the observed count for $X=i$ & $Y=j$.
- $E_{ij}$ is the expected value for $X=i$ & $Y=j$, assuming $X,Y$ are independent.

$$E_{ij} = \frac{N_{X=i} \times N_{Y=j}}{N} = \left( \sum_{c=1}^{l} O_{ic} \right) \times \left( \sum_{r=1}^{k} O_{rj} \right) \frac{1}{N}$$
Chi Square ($\chi^2$) Test of Independence

$X^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ 

asymptotically distributed as $\chi^2$ with $(k-1)(l-1)$ degrees of freedom if $X, Y$ are independent.

Use $X^2$ test value to rank features $X$ with respect to label $Y$. 
Pearson Correlation Coefficient

- Feature X and label Y are two random variables.
- Population correlation coefficient (*linear dependence*):
  \[ \rho(X,Y) = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \]
- Sample correlation coefficient:
  \[ \rho(X,Y) = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}} \]
- Values always between \([-1,+1]\]
  - when linearly dependent +1, −1, when independent 0.
Pearson Correlation Coefficient
Signal-to-Noise Ratio (S2N)

- Feature X and label Y are two random variables:
  - Y is binary, \( Y \in \{ y_+, y_- \} \)
- Let \( \mu_+, \sigma_+ \) be the sample \( \mu, \sigma \) of X for which \( Y = y_+ \).
- Let \( \mu_-, \sigma_- \) be the sample \( \mu, \sigma \) of X for which \( Y = y_- \).

\[
\mu(X, Y) = \frac{|\mu_+ - \mu_-|}{\sigma_+ + \sigma_-}
\]

related to Fisher’s criterion
Ranking Features with the T-test

- Let \( m_+ \) be the number of samples in class \( y_+ \).
- Let \( m_- \) be the number of samples in class \( y_- \).

\[
T(X,Y) = \frac{|\mu_+ - \mu_-|}{\sqrt{\sigma^2_+/m_+ + \sigma^2_-/m_-}}
\]

\( \mu_- \quad \mu_+ \)

\( \sigma_- \quad \sigma_+ \)

\( P(X|Y=1) \)

\( P(X|Y=-1) \)