Dimensionality Reduction

Jakramate Bootkrajang

10 Aug 2014

Outlines

- Principle Component Analysis (PCA)
- Feature Subset Selection
- Random Projection

Feature Subset Selection

FSS: Goal

- Goal: Find the optimal feature subset
- There are a number of methods
  - Wrappers
  - Filters
  - Embedded methods

FSS: Key ideas

- Need a measure for assessing the goodness of a feature subset (scoring function)
- A strategy to search the space of possible feature subsets
- Brute force is not applicable (NP-Hard)
- Involve ranking features based on some criteria.
- Commonly used as a preprocessing step for classification task

FSS: Widely used ranking criteria

- Signal-2-Noise Ratio
- Information gain
- Mutual Information

FSS: Signal-2-Noise Ratio

- Define how well a feature discriminates two classes.
- $S2N = \frac{\mu_+ - \mu_-}{\sigma_+ + \sigma_-}$
- μ_+ is the mean of the data points having positive class label.
- μ_- is the mean of the data points having negative class label.
- σ_+ is the standard deviation of the data points having positive class label.
- σ_- is the standard deviation of the data points having negative class label.

FSS: Information gain

- Measures the number of bits of information gained about the class label when knowing the feature.
- Measure the uncertainty about Y (the class label)
- Measure the uncertainty about Y given feature X (the class label with feature)
- The uncertainties can be measure using the entropy $H(Y)$ and $H(Y|X)$
- Information gain, $IG(X) = H(Y) - H(Y|X)$

FSS: Entropy

- A measure of uncertainties of information content.
- For example
  - An entropy of a fair coin toss is large since we cannot be certain about the outcome,
    $P(outcome = \text{`}head\text{'}) = P(outcome = \text{`}tail\text{'}) = 0.5$
  - However, an entropy of a biased coin toss is lower than that of the above, i.e., $P(outcome = \text{`}head\text{'})$ is much higher than $P(outcome = \text{`}tail\text{'})$.

FSS: Calculating an entropy

- Entropy of event Y

  $H(Y) = -\sum_{i=1}^{k} P(Y = y_i) \log_2(P(Y = y_i))$
- Conditional entropy of Y given X

  $H(Y|X) = \sum_{j=1}^{r} P(X = x_j) H(Y|X = x_j)$

  where,

  $H(Y|X = x_j)$ refers to an entropy of Y among only those records in which X has value x_j
- Specifically $H(Y|X = x_j)$ is

  $= -\sum_{i=1}^{k} P(Y = y_i|X = x_j) \log_2(P(Y = y_i|X = x_j))$

FSS: Information gain (an example)

X = College Major

Y = Likes "Gladiator"

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

- $H(Y) =$
- $H(Y|X) =$

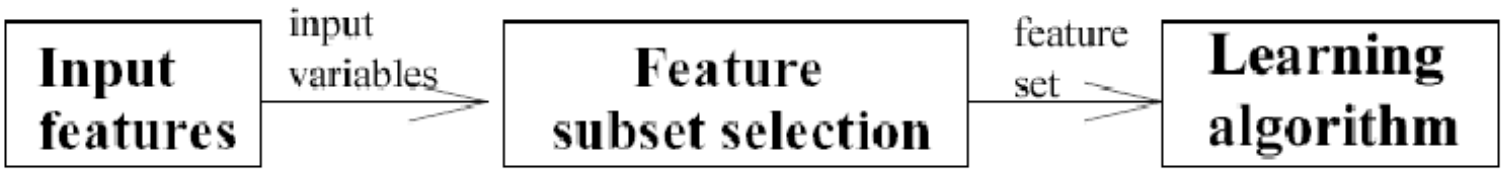
- $IG = H(Y) - H(Y|X)$

FSS: Mutual Information

- I(x,y) = how much information do x and y share.
- It measures how much knowing one of these variables reduces uncertainty about the other.
- High MI example: Person's name and Gender
- Low MI example: ??? and Gender
- Definition: $I(x,y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$
- Can you show that there is a connection between the conditional entropy $H(Y|X)$ and mutual information?

FSS: Filter methods

- Select subsets of variables as a pre-processing step,
- Independently of the choice of classifier



FSS: Pros/Cons of filter methods

PROS

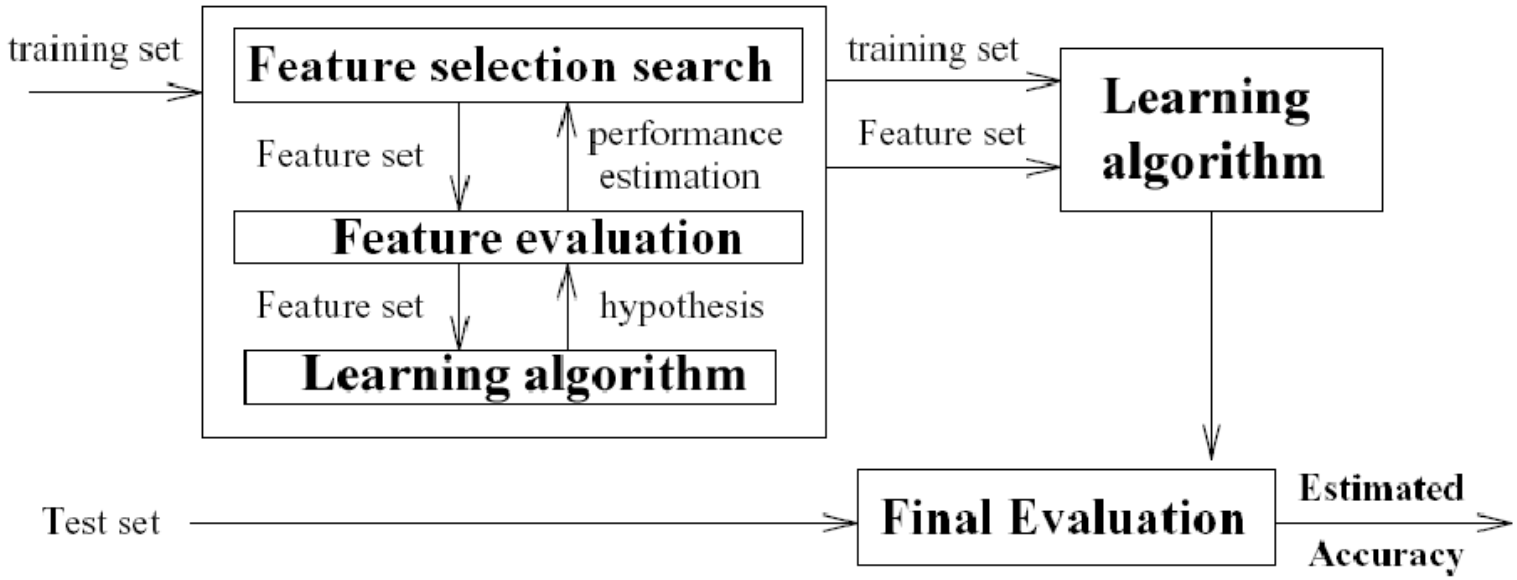
- Usually fast
- Provide generic selection of features, not tuned by given learner
- Can be used as a preprocessing step for other methods

CONS

- Feature not optimised for the choice of classifier

FSS: Wrapper methods

- Learner is considered a black-box
- Final subsets vary for different learners



- Need to define
  - How to search the space of all possible subsets?
  - How to assess the performance of a learner?

FSS: Wrapper methods

- How to search the space?
  - Brute force: Not a good idea, there are 2^m subsets to evaluate.
  - Forward selection (start with empty feature set and add features at each step)
  - Backward elimination (start with full feature set and discard features at each step)
- How to evaluate the learner?
  - For classification: Measure accuracy on hold-out validation set.

FSS: Embedded methods

- Specific to a given learning machine
- Performs variable selection in the process of training
- Optimise the regularised objective: *obj* + *regularisation*
- For example,
  $argmin_w \sum_{i=1}^{N} y_i(\mathbf{w}^T\mathbf{x}_i + b) - \lambda \sum_{j=1}^{M} |w_j|$