Foundations of Machine Learning

Module 3: Instance Based Learning and Feature Reduction

Part B: Feature Selection

Sudeshna Sarkar
IIT Kharagpur

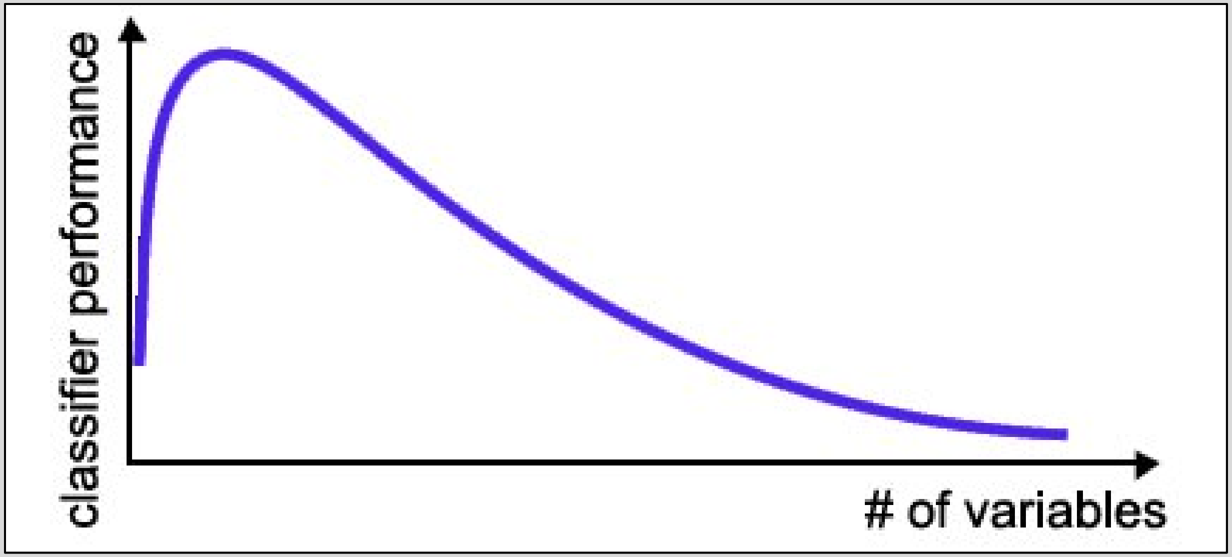
Feature Reduction in ML

- The information about the target class is **inherent in the variables**.
- Naïve view:
  More features
  => More information
  => More discrimination power.
- In practice:
  many reasons why this is not the case!

Curse of Dimensionality

- number of training examples is fixed

  => the classifier's performance usually will degrade for a large number of features!



Feature Reduction in ML

- Irrelevant and
- redundant features
  - can confuse learners.

- Limited training data.
- Limited computational resources.
- **Curse of dimensionality**.

Feature Selection

Problem of selecting some subset of features, while ignoring the rest

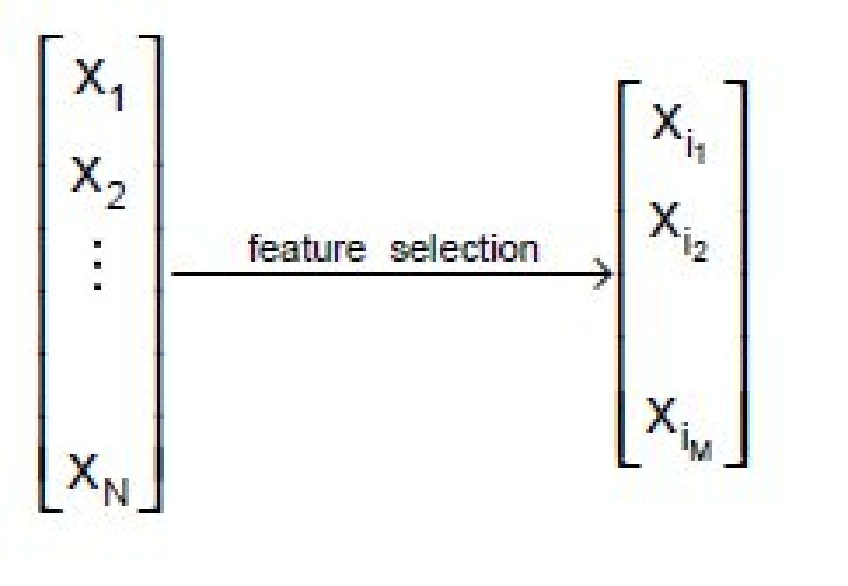
Feature Extraction

- Project the original x_i , i =1,...,d dimensions to new $k < d$ dimensions, z_j , j =1,...,k

Criteria for selection/extraction:
either improve or maintain the classification accuracy, simplify classifier complexity.

Feature Selection - Definition

- Given a set of features $F = \{x_1, \ldots, x_n\}$
  the Feature Selection problem is
  to find a subset $F' \subseteq F$ that maximizes the learners ability to classify patterns.
- Formally F' should maximize some scoring function



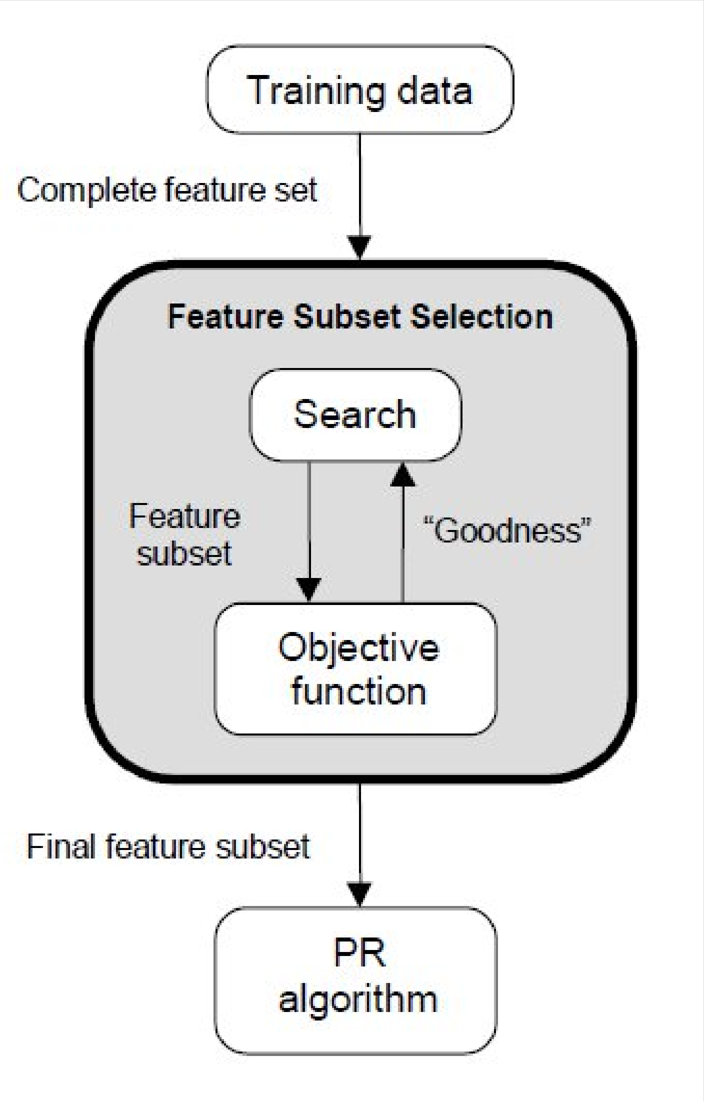
Subset selection

- d initial features
- There are 2^d possible subsets
- Criteria to decide which subset is the best:
  - classifier based on these m features has the lowest probability of error of all such classifiers
- Can't go over all 2^d possibilities
- Need some heuristics

Feature Selection Steps

Feature selection is an **optimization** problem.

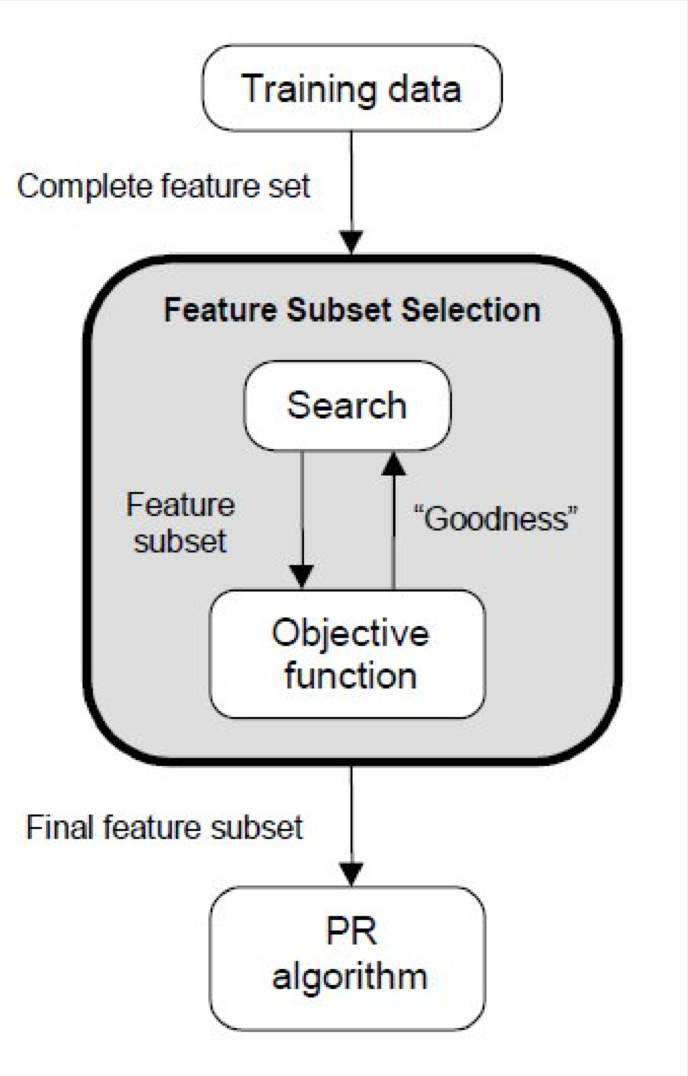
- Step 1: Search the space of possible feature subsets.
- Step 2: Pick the subset that is optimal or near-optimal with respect to some objective function.



Feature Selection Steps (cont'd)

Search strategies
- Optimum
- Heuristic
- Randomized

Evaluation strategies
- Filter methods
- Wrapper methods

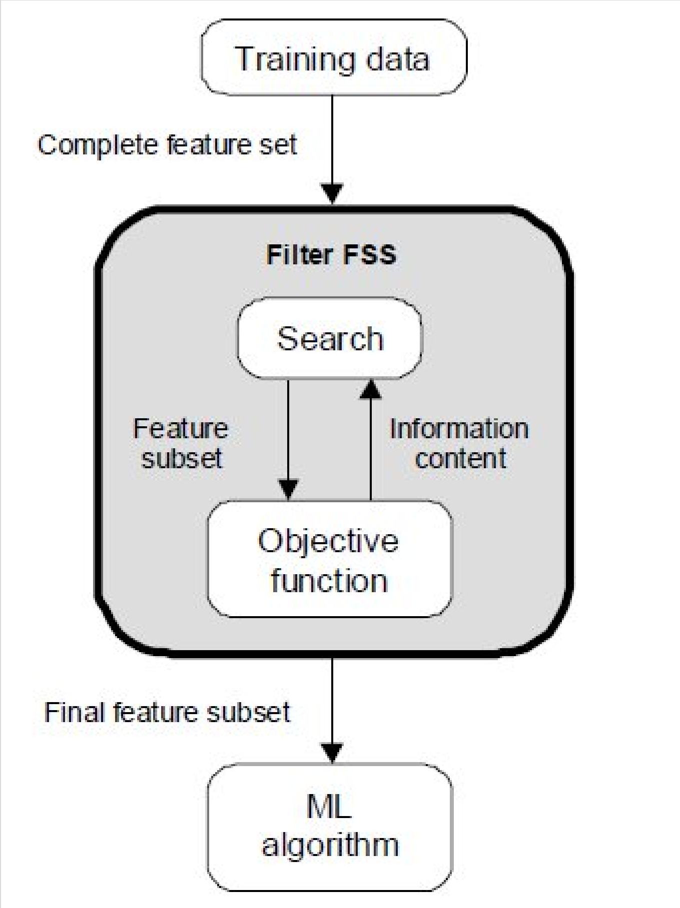


Evaluating feature subset

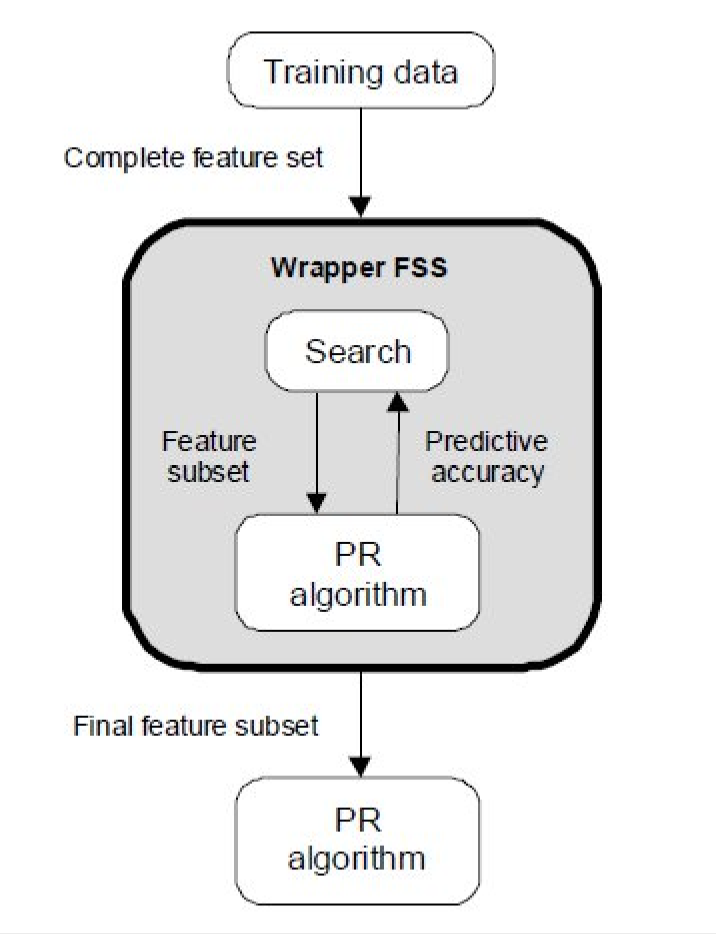
- Supervised (wrapper method)
  - Train using selected subset
  - Estimate error on validation dataset
- Unsupervised (filter method)
  - Look at input only
  - Select the subset that has the most information

Evaluation Strategies

Filter Methods



Wrapper Methods



Subset selection

- Select uncorrelated features
- Forward search
  - Start from empty set of features
  - Try each of remaining features
  - Estimate classification/regression error for adding specific feature
  - Select feature that gives maximum improvement in validation error
  - Stop when no significant improvement
- Backward search
  - Start with original set of size d
  - Drop features with smallest impact on error

Feature selection

Univariate (looks at each feature independently of others)

- Pearson correlation coefficient
- F-score
- Chi-square
- Signal to noise ratio
- mutual information
- Etc.

Univariate methods measure some type of correlation between two random variables

- the label (y_i) and a fixed feature (x_{ij} for fixed j)

- Rank features by importance
- Ranking cut-off is determined by user

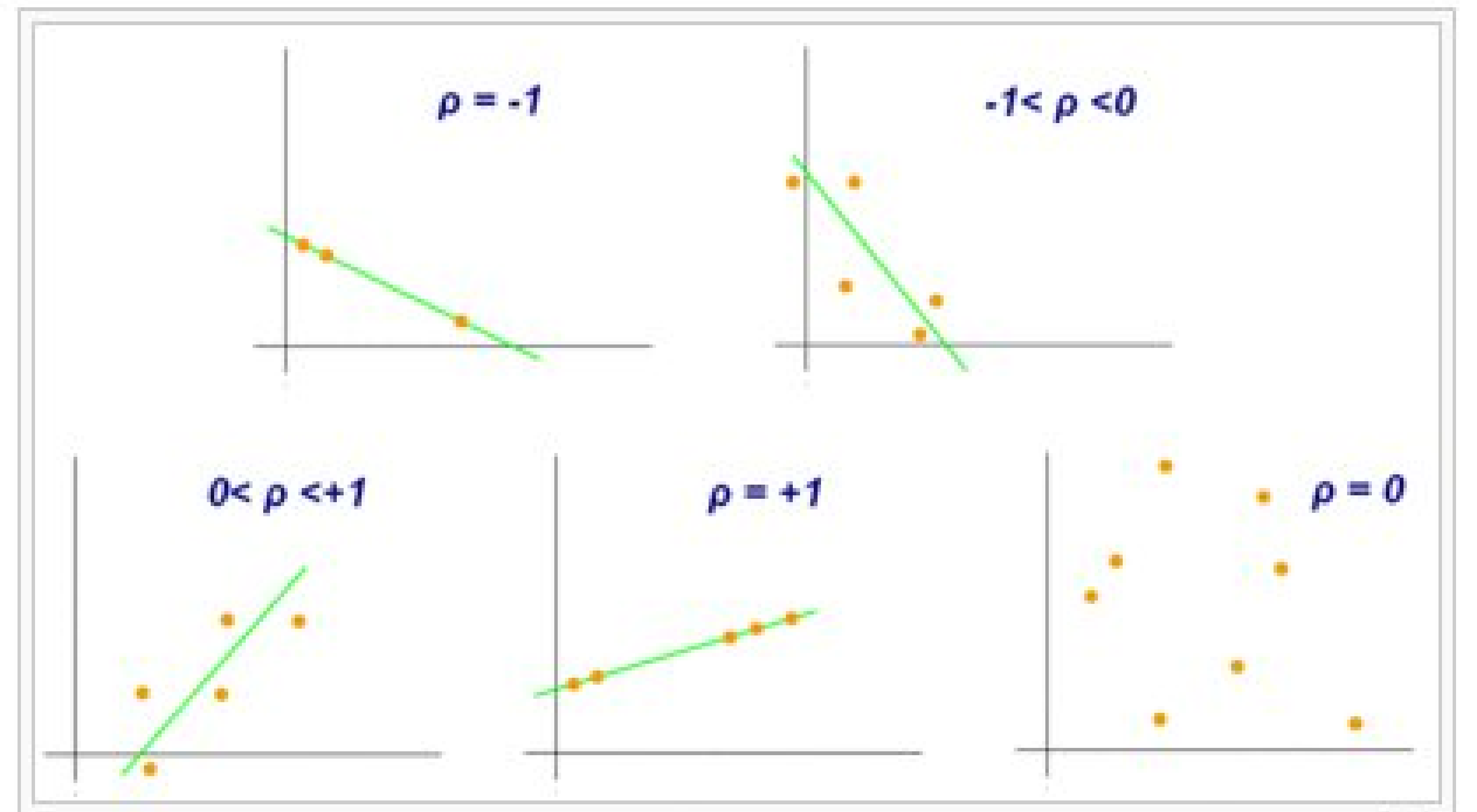
Pearson correlation coefficient

- Measures the correlation between two variables
- Formula for Pearson correlation =

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

- The correlation r is between $+1$ and -1.
  - $+1$ means perfect positive correlation
  - -1 in the other direction

Pearson correlation coefficient



From Wikipedia

Signal to noise ratio

- Difference in means divided by difference in standard deviation between the two classes

$$S2N(X,Y) = (\mu_X - \mu_Y)/(\sigma_X - \sigma_Y)$$

- Large values indicate a strong correlation

Multivariate feature selection

- Multivariate (considers all features simultaneously)
- Consider the vector w for any linear classifier.
- Classification of a point x is given by $w^T x + w_0$.
- Small entries of w will have little effect on the dot product and therefore those features are less relevant.
- For example if $w = (10, .01, -9)$ then features 0 and 2 are contributing more to the dot product than feature 1.
  - A ranking of features given by this w is 0, 2, 1.

Multivariate feature selection

- The w can be obtained by any of linear classifiers
- A variant of this approach is called recursive feature elimination:
  - Compute w on all features
  - Remove feature with smallest w_i
  - Recompute w on reduced data
  - If stopping criterion not met then go to step 2