Feature selection

M. Vazirgiannis

December 2014

Feature selection

Data contain redundant or irrelevant to the learning problem features.

- Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context

Alternative terms: variable selection, attribute selection or variable subset selection

Feature selection vs. feature extraction.

- FE creates new features from functions of the original features (i.e. dimensionality reduction)
- FS returns a subset of the features.
- FS techniques are often used when there are many features and few data points.
- Example: analysing DNA microarrays, thousands of features, few tens to hundreds of samples.

Feature selection

Main benefits

- improved model interpretability
- shorter training times
- enhanced generalization by reducing over fitting.
- useful in the data analysis process
  - features important for prediction
  - how features are related

Feature selection

Select the “best” features (subset of the original set)

- Filter methods:
  rank the features individually according to some criteria (information gain, χ^2, etc.) and take the top-k or eliminate redundant features (correlation)
- Wrapper methods:
  evaluate each subset using some data mining algorithm; use heuristics for the exploration of the subset space (forward/backward search, etc.)
- Embedded methods:
  feature selection is part of the data mining algorithm. For example the LASSO method constructing linear model penalizing the regression coefficients, shrinking many of them to zero.

Filter methods - Information Gain (IG)

- For a random variable X (class) its entropy

$$H = -\sum_{i=1}^{c} P(x_i) \times \log(P(x_i)) \quad \text{, c classes}$$

  - "High Entropy": X is from a uniform distribution – lack on information
  - "Low Entropy": X is from varied (peaks and valleys) distribution – rich in information content
- Let variable A (feature), IG(X, A) represents the reduction in entropy (~ gain in Information) of X achieved by learning the state of A: IG(X,A)=H(X)–H(X|A)

Filter methods - Chi-squared test ($\chi 2$)

- Test of independence between a class X and a feature A
- $\chi^2(A) = \sum_{i=1}^{v} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$, v values, c classes

  O_{ij}: observed frequency of class j for feature A (value i)

  E_{ij}: the expected frequency

$$E_{ij} = \frac{\text{(\# of samples with value i) x (\# of samples with class j)}}{\text{\# of samples in total}}$$

Wrapper methods – Subset selection

- Find the subset of k variables that predicts best:
  - This is a generic problem when p is large
    (arises with all types of models, not just linear regression)

- Models with different complexity..
  - *p* models with a single variable
  - *p(p-1)/2* models with 2 variables, etc…
  - 2^p possible models in total
  - Exhaustive search is intractable

- What does “best” mean here?

Search Problem

- How can we search over all 2^p possible models?
  - exhaustive search is clearly infeasible

- Heuristic search is used to search over model space:
  - Forward search (greedy)
  - Backward search (greedy)
  - Branch and bound techniques

- variable selection problem in several data mining algorithms
  - Outer loop that searches over variable combinations
  - Inner loop that evaluates each combination

Forward selection

- Assume a regression problem
- Start with the feature the lowest p-value (i.e. the highest evidence for rejecting the null hypothesis, i.e. the variables are correlated to the class)
- add in each repetition the variable with the *highest* F-test value
- *Assume two models* p_2,p_1 with $|p_2|>|p_1|$, apparently $RSS_2<RSS_1$

$$F = \frac{\left(\frac{\mathrm{RSS}_1 - \mathrm{RSS}_2}{p_2 - p_1}\right)}{\left(\frac{\mathrm{RSS}_2}{n - p_2}\right)}$$

- Repeat until F-value < $threshold_f$ (or p-value > $threshold_p$)
- *RSSi* the residual sum of squares - the error induced by the model:

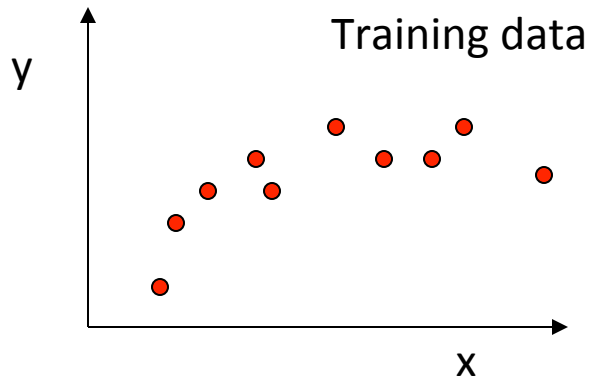
$$RSS = \sum_{i=1}^{n} (y_i - f(x_i))^2,$$

with y_i real value and $f(x_i)$ predicted by models containing p_i .

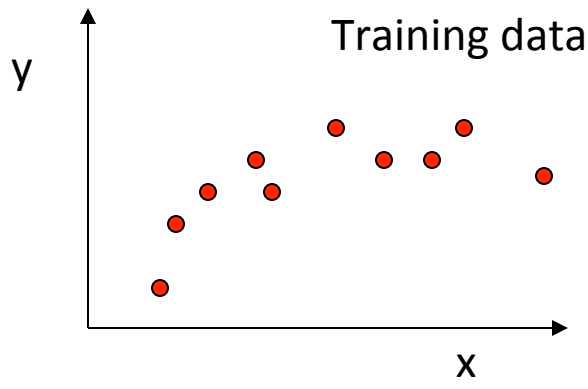
Backward Elimination

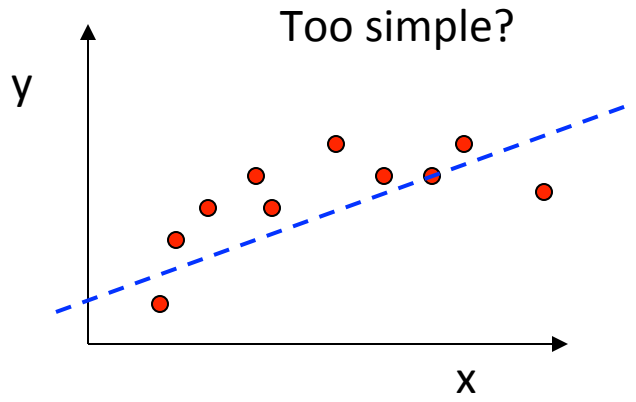
- start with the full model
- drop the predictor that produces the smallest F value (or highest p-value)
- Continue until F-value < threshold_f
  - (or p-value > threshold_p)
- Sometimes constraint N>p

Complexity versus Goodness of Fit

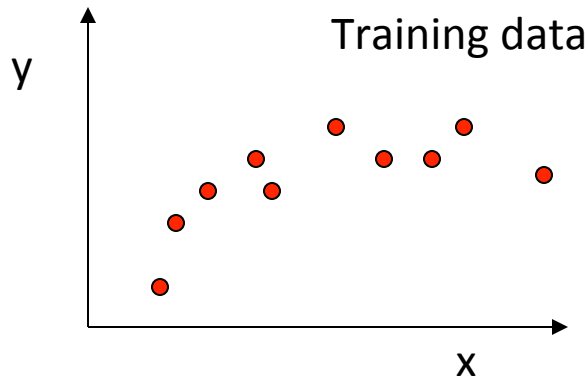


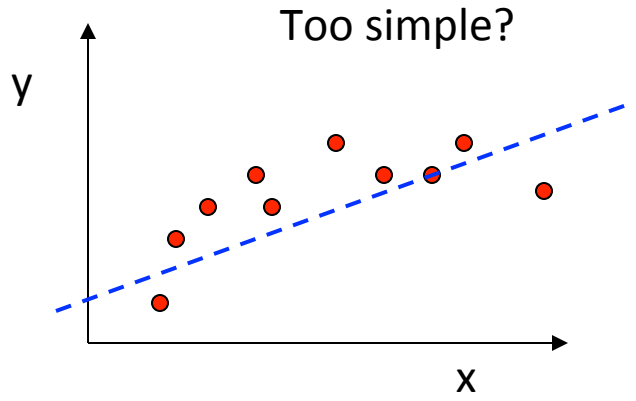
Complexity versus Goodness of Fit

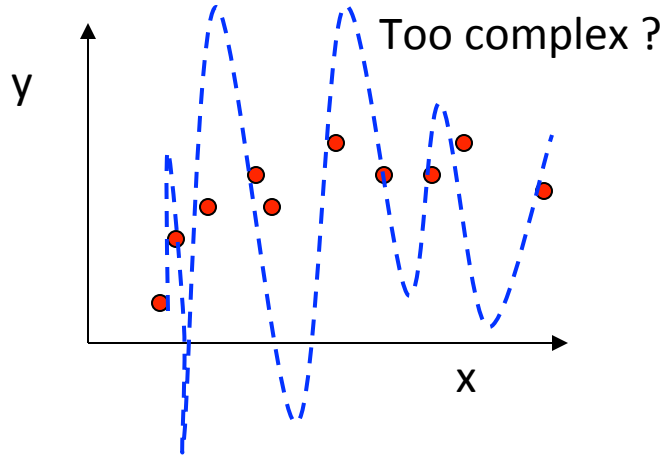




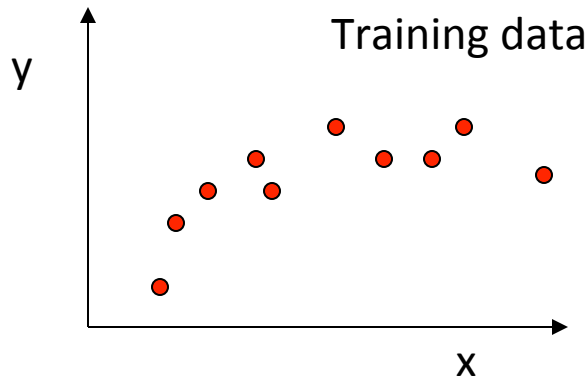
Complexity versus Goodness of Fit

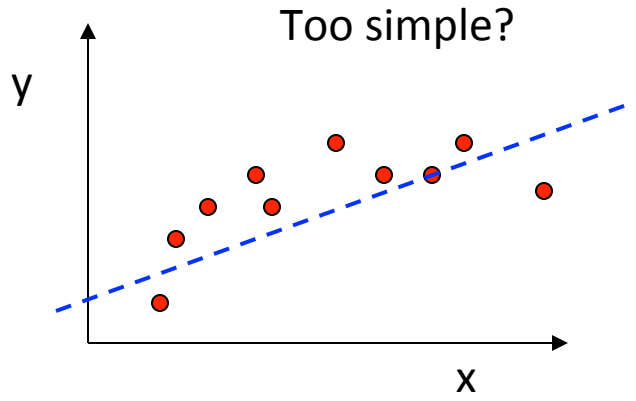


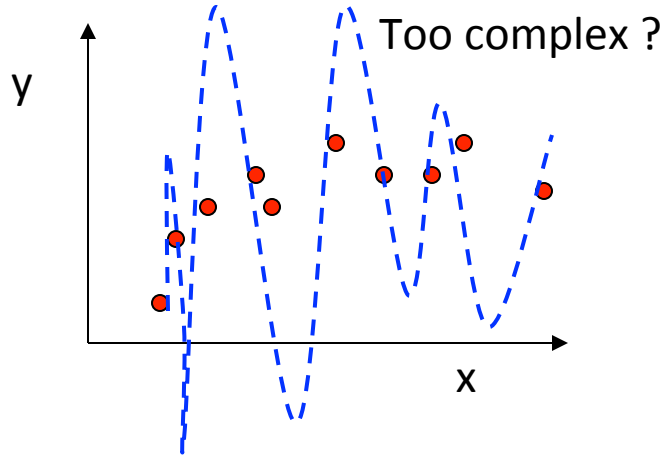


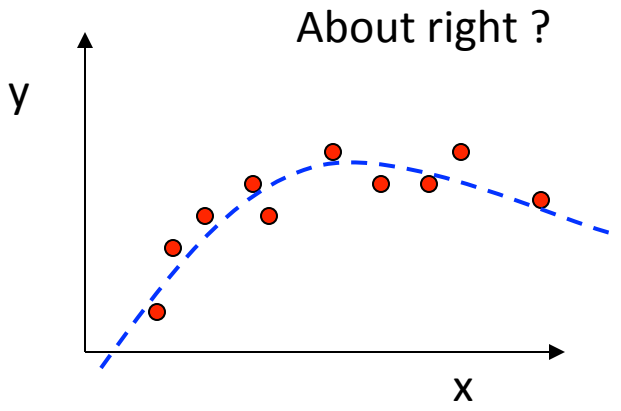


Complexity versus Goodness of Fit

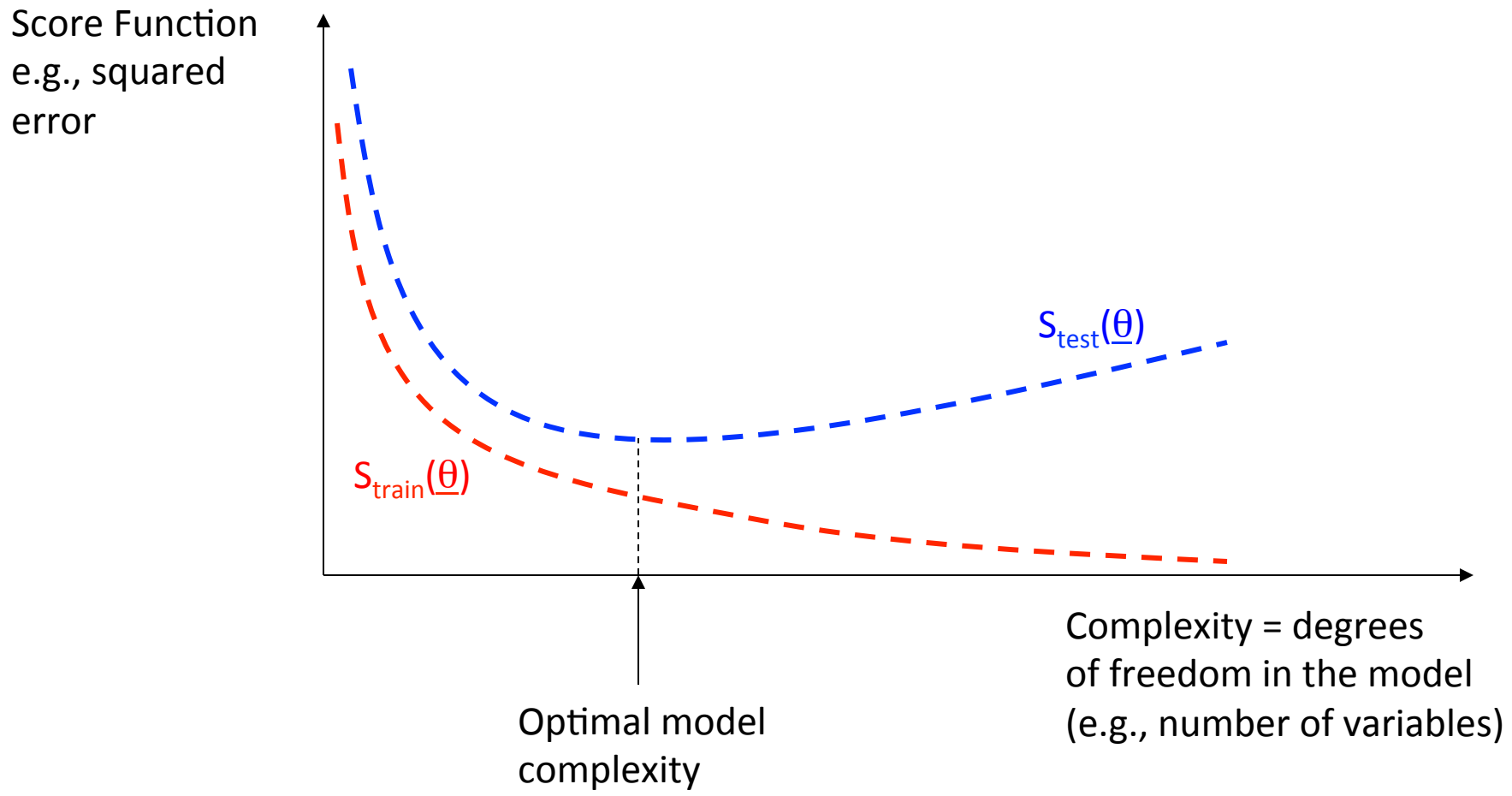








Complexity and Generalization



REFERENCES

- An introduction to variable and feature selection, I Guyon, A Elisseeff - The Journal of Machine Learning Research, 2003
- Wrappers for feature subset selection, R Kohavi, GH John - Artificial intelligence, 1997 – Elsevier