Topic 12. The Principles of Data Mining

Reading Assignment

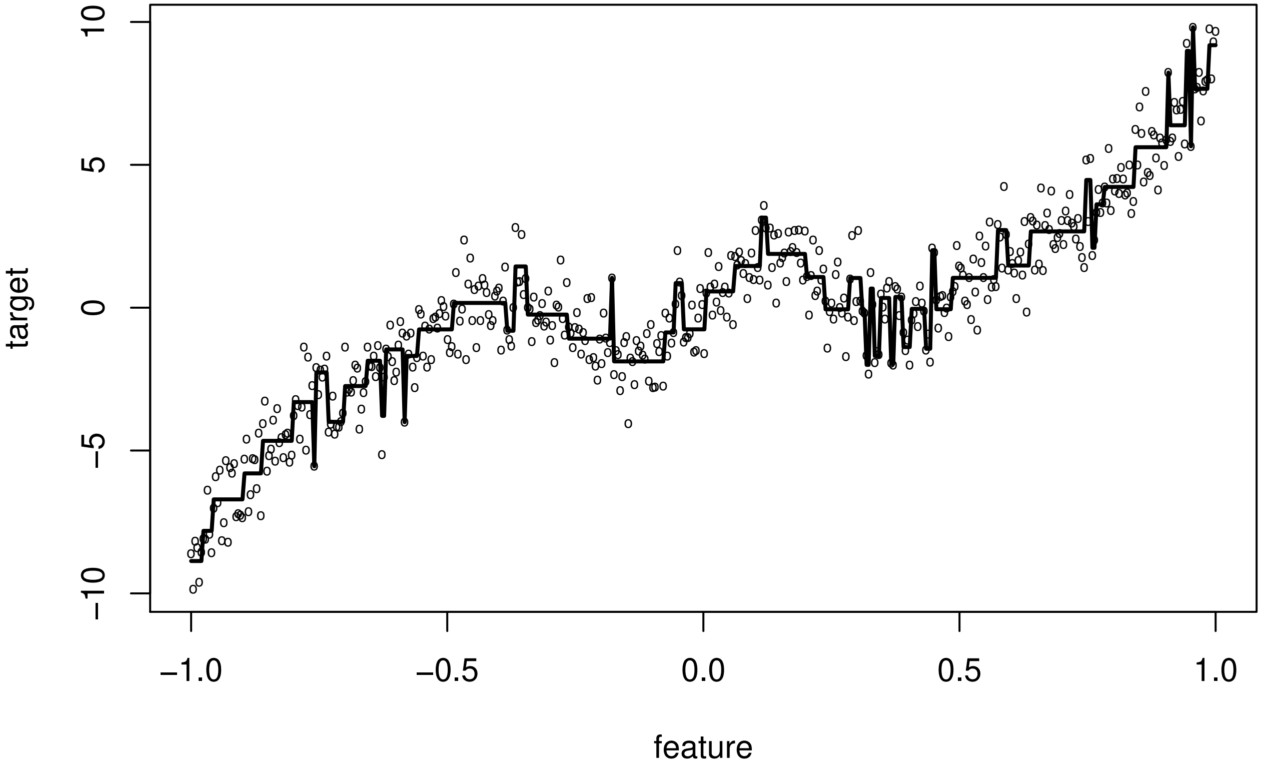
Berry and Linoff (2000)

- Pages 36, 64, 90, 129, 181–182, 224–226. Lessons learned.

1. Validation

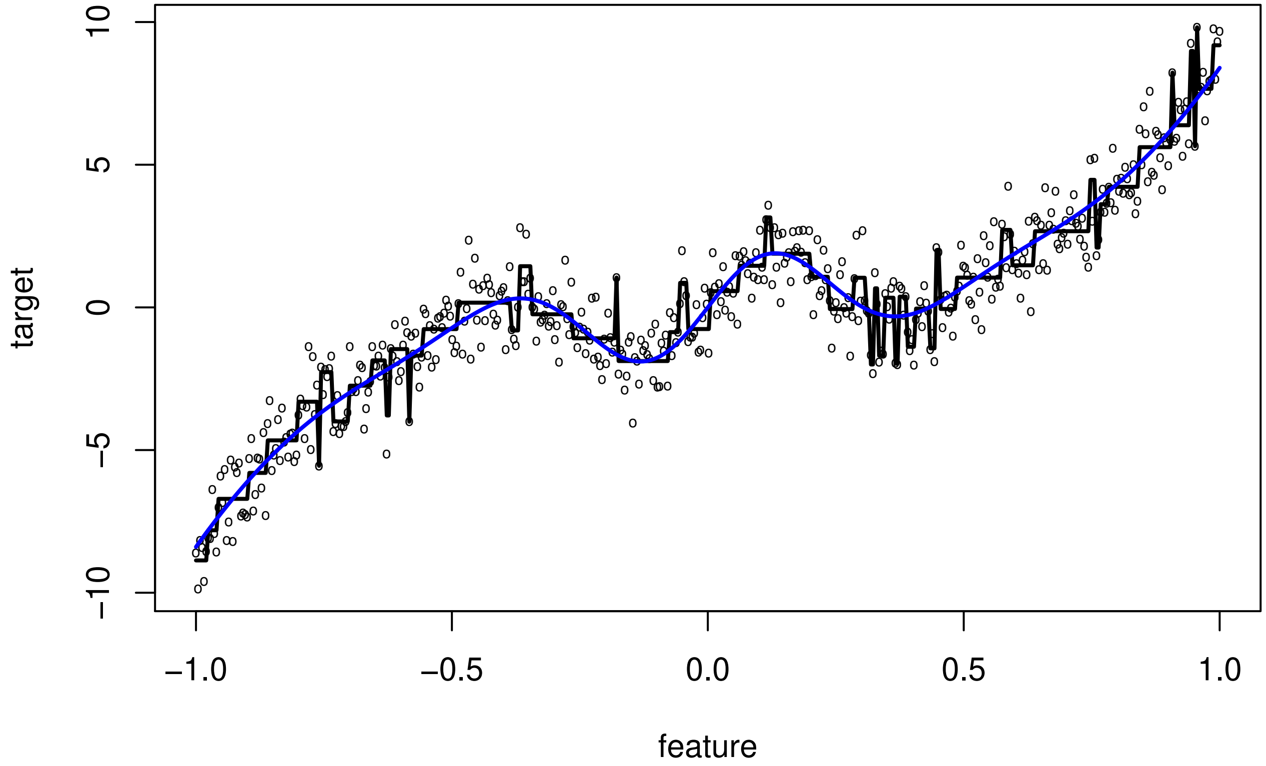
- When tools learn they will over fit if complexity is determined in the training sample.
- The consequence is that the trained tool will predict poorly when used to score new data; aka poor generalization.
- When a tool's behavior is studied over repeated samples, over fitting entails small bias (average of fits close to truth) but large variance (each fit differs considerably from the average of fits).
- Using a validation sample to choose the tool's complexity balances bias and variance and leads to models that generalize well.
- If data are abundant, use a validation sample. For medium sized data sets, use n-part cross validation. For small data sets use leave-one-out cross validation.

Fig 43. Tree Compared to Data, $cp = 0.0005$



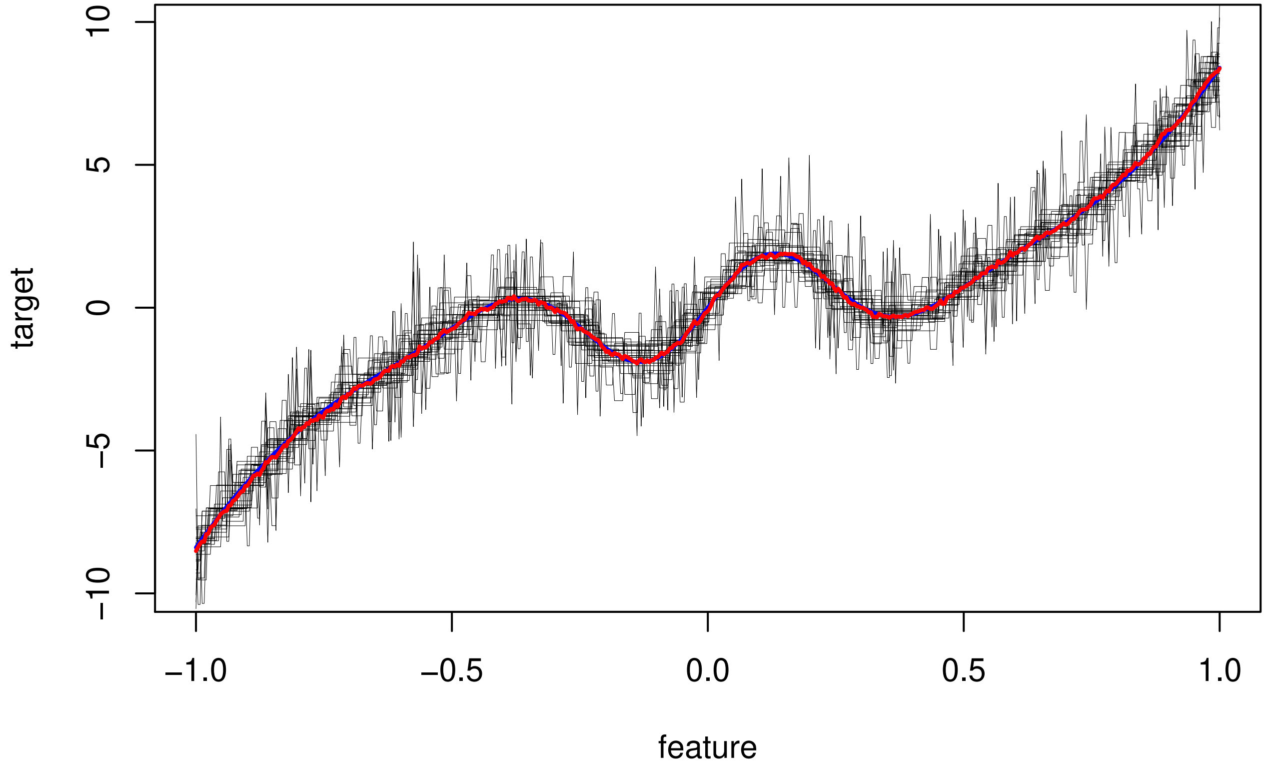
The prediction of the fitted tree is shown in as a black line; the learning sample is shown as black circles. (The best tree had cp=0.001.

Fig 44. Tree Compared to Truth, $cp = 0.0005$



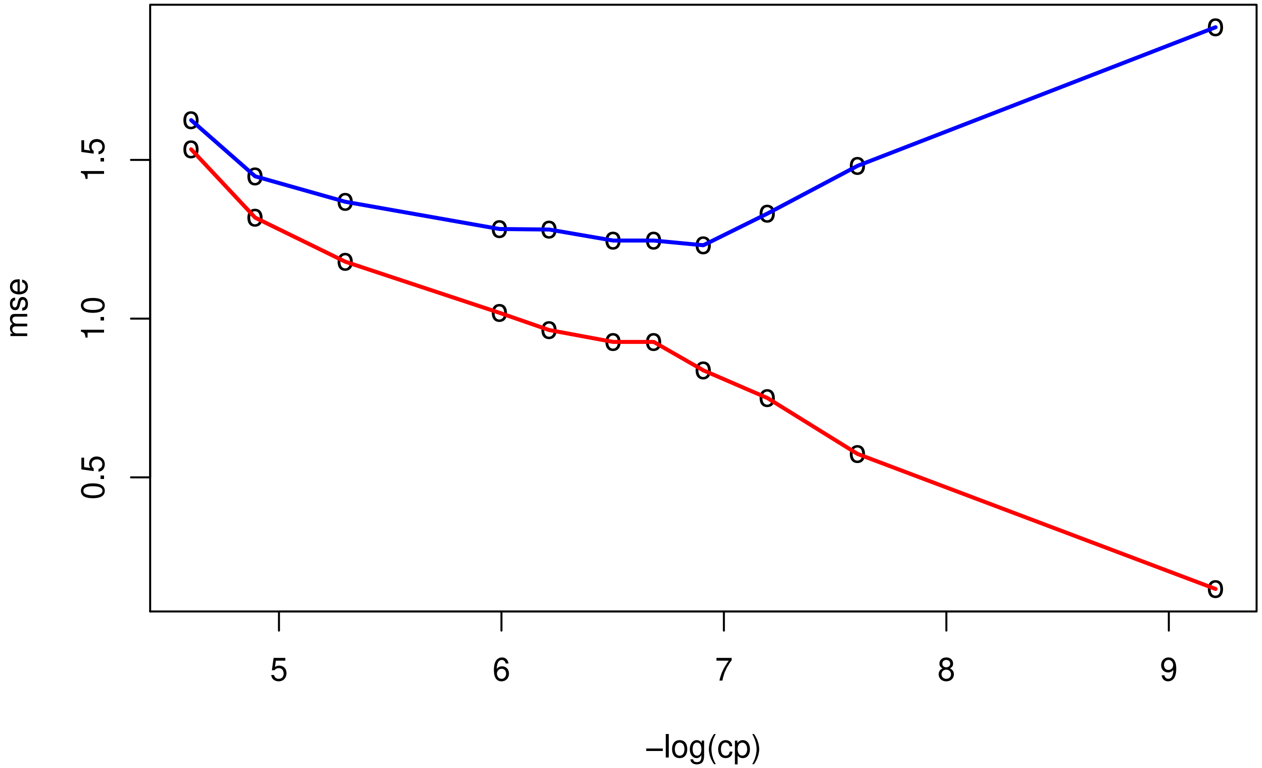
The model is $y = f + e$ where f is shown in blue. The prediction of the fitted tree is shown as a black line; the learning sample is shown as black circles.

Fig 45. Variability in Tree, $cp = 0.0005$



Trees were fitted to 100 samples from $y = f + e$ where f is shown in blue. The red line is the average of the 100 trees and the black lines are the predictions of the first 25 trees.

Fig 50. Learning and Validation mse



The mse in the learning sample is shown in red. The mse in the validation sample is shown in blue. The mse in the leaning sample underestimates the bias plus variance.

2. Features

- A critical determinant of success in data mining is finding the best (derived) features.

- If the (derived) features are well chosen, then even simple tools like regression will work well.

- Domain knowledge is the best source of suggestions for (derived) features.

- Taylor's theorem suggests that it is also reasonable to try squares and cross-products (aka interactions) of features.

3. Scaling Laws

- Information is acquired at the rate of $\sqrt{n}$, i.e., standard errors decline at the rate $\frac{1}{\sqrt{n}}$, where n is the number of cases.

- This is a very slow rate of acquisition and if the relationship one seeks is obscured by a low signal to noise ratio, then the number of required cases required to find it can be huge – one hundred thousand to several million.

- One is often in this situation in data mining which makes machine assisted model selection and feature selection essential.

4. Software

- Data acquisition and cleaning is messy and tedious. Flexible software with looping and control structures is usually required. Examples are perl, C++, R, Splus.

- Automated feature selection can also require flexible software with looping and control structures but some menu driven software is effective. An example is SAS Enterprise Miner.

- In the final stages of analysis, menu driven software is usually more convenient.

5. Regression

- The oldest method (Gauss, 1816).

- Efficient use of space and time.

- Works very well when features are well chosen.

- Statistics computed from the learning sample – F-tests, Mallows Cp, etc. – should only be used to screen features.

- Validation should be used for feature selection.

6. Neural Networks

- Neural nets as typically implemented cannot handle categorical features.

- Boosting can be used to resolve this difficulty and, indeed, make a dictionary method out of any combination of tools.

- Nets are notoriously difficult to fit!

- The main plus is that neural nets are universal approximators which amounts to saying that nets are automatic derived feature finders.

- When nets work, their performance can be spectacular.

7. Decision Trees

- Decision trees are popular because they are interpretable.
- Decision trees are easy to use and can handle both categorical and numerical features and targets.
- Due to their interpretability, they can make good feature detectors.
- Like nets, decision trees are universal approximators.

8. Unsupervised Learning

- Evaluation of unsupervised learning results is subjective.
- The main tools of unsupervised learning are principal components, hierarchical clustering, and K-means.
- These tools complement each other. Hierarchical clustering helps to get an initial feel for the data and to find likely clusters. This information feeds into K-means for refinement. Principal components helps visualize K-mean results.
- Classification trees and linear discriminant analysis can help interpret clustering results. Cannonical correlations can help visualize them.
- Success depends critically on good feature selection.

9. Classification

- The main tools are classification trees, thresholded regression, thresholded nets, logistic regression and linear discriminant analysis. The latter two are probabilistic tools, i.e. compute $P(C|X = x)$.

- Linear discriminant analysis is the best of the probabilistic classification tools.

- With derived features, linear discriminant analysis is flexible.

10. Association Rules

- Association rules mining finds interesting relationships in data.

- The apriori algorithm can quickly analyze very large commercial data sets.

- Association rules are among data mining's biggest successes (Hastie, Tibshiranai, and Freedman, 2001).

11. Boosting

- In general it is a good idea to try several tools.

- Boosting is a way to combine similar or different tools.

- The method works best when a large number of tools with a minimal number of features per tool are combined.

12. Oversampling

- One effect of oversampling is to change the loss function.
- One can choose a loss function to suit the problem in ways that do not discard data.
- Although theoretical results can be misleading because the assumptions are not nearly enough satisfied, this is one instance when theory gives the correct answer: Over sampling is counterproductive.

Blank page

Blank page