Introduction to Statistical Learning

– (Part Two) –

Topic 3. Statistical learning - a high-level overview and illustrative examples

3.2. Estimation: how and why; tradeoff between accuracy and interpretability

Sonja Petrović

Created for ITMD/ITMS/STAT 514

Spring 2021.

Goals of this lecture

- Setting the context: estimating f
- Accuracy-interpretability trade off
  - [looking forward to the bias-variance trade off]
- Supervised vs. unsupervised learning
- Regression vs. classification

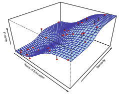
Section 1

Setting the context: estimating f, accuracy & interpretability

Review: estimating f [regression setting]

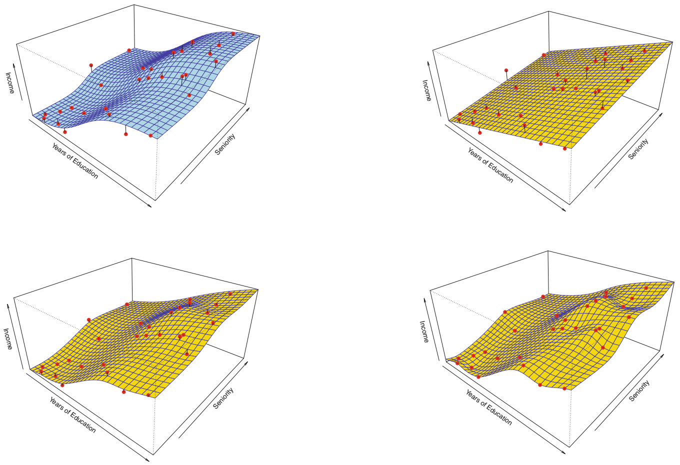
- Observe: a quantitative response Y, p different predictors, $X_1, X_2, \ldots, X_p$.
- Assume: some relationship between Y and $X = (X_1, X_2, ..., X_p)$, which can be written in the very general form

$$Y = f(X) + \epsilon.$$



- f is some fixed but unknown function of $X_1, X_2, \ldots, X_p$
- ϵ is a random error term, which is independent of X and has mean zero.
- In this formulation, f represents the *systematic* information that X provides about Y.

f, $Y = f(X) + \epsilon$, $\hat{f}$, $\hat{Y} = \hat{f}(X)$ [regression setting]



Accuracy vs. interpretability

- **Less flexible** methods = more restrictive, relatively small range of shapes for $\hat{f}$.
  - E.g.: linear regression
- **More flexible** methods = can generate a wider range of possible shapes to estimate f.

Accuracy vs. interpretability

- **Less flexible** methods = more restrictive, relatively small range of shapes for $\hat{f}$.
  - E.g.: linear regression
- **More flexible** methods = can generate a wider range of possible shapes to estimate f.
- **Why** ever choose more restrictive?!
  - Inference: restrictive $\leftrightarrow$ interpretable
    - E.g. Linear model: easy to understand the relationship between Y anad $X_1, \ldots, X_p$.
    - Flexible approach can lead to such complicated estimates of f that it is difficult to understand how any individual predictor is associated with the response.
  - Prediction: the interpretability of the predictive model is simply not of interest
    - Expect? - best to use most flexible model
    - Surprise: often more accurate prediction using a less flexible method *(looking ahead: the overfitting phenomenon)*.

Generalizability: a central theme

Construct predictors that generalize well to unseen data

- Capture useful trends in the data *(don't underfit)*
- Ignore meaningless random fluctuations in the data *(don't overfit)*

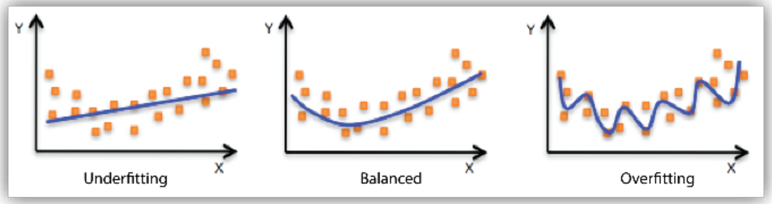
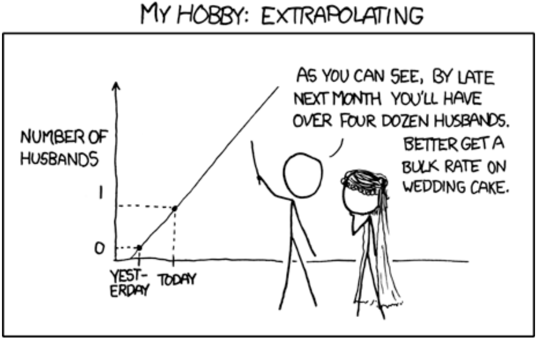


Figure 1: meaning of overfitting and underfitting

Avoid unjustifiably extrapolating beyond the scope of the data



Randall Munroe, xkcd

Figure 2: meaningless extrapolation

Reminder: supervised vs unsupervised learning

- Predictive Analytics (Supervised learning):
    - Q: To whom should I extend credit?
        - Task: Predict how likely an applicant is to repay loan.
    - Q: What characterizes customers who are likely to churn?
        - Task: Identify variables that are predictive of churn.
    - Q: How profitable will this subscription customer be?
        - Task: Predict how long customer will remain subscribed.
- Descriptive Analytics (Unsupervised learning):
    - Clustering customers into groups with similar spending habits
    - Learning association rules: E.g., 50% of clients who {recently got promoted, had a baby} want to {get a mortgage}

Supervised vs. unsupervised – from f's point of view:

Supervised learning

For each observation of the predictor measurement(s) $x_1, \ldots, x_n$, there is an associated response measurement y_i.

Unsupervised learning

for every observation $i = 1, \ldots, n$, we observe a vector of measurements x_i but no associated response y_i.

Regression vs. classification

Types of random variables:

quantitative (continuous) or qualitative (categorical, discrete).

We select learning methods based on type of response (predictor type less important)!

- Quantitative response $\mapsto$ regression problems
- Qualitative response $\mapsto$ classification problems
    - ... *but the lines do blur, so beware:*
        - Least squares linear regression is used with a quantitative response,
        - Logistic regression is typically used with a qualitative (two-class, or binary) response. As such it is often used as a classification method.

→ Up next: ←

- Assessing model accuracy
  - (from the point of view of both classification and regression)
    - [NEXT LECTURE]
- Training & testing data sets
  - Partitioning
  - Balancing
  - Cross-validation, etc.
    - [NEXT LECTURE; but in preparation for that: HANDS-ON LAB NOW]

Aha!

It is time for AhaSlides review! https://www.ahaslides.com/STATITMW11

Lab time!

$\rightarrow$ Hands-on: group breakout work $\leftarrow$

See worksheets handouts posted on Campuswire:

- Partitioning the data
- Validating the partition
- Balancing

License

Content of this lecture is based on the first two chapters of the textbook Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, *' An Introduction to Statistical Learning: with Applications in R'*. The book is available online.