University of Toronto- Time series club
Lecture 3
Introduction to function fitting

Lim, Kyuson

05/18/2022

Today's outline

- Understand function fitting, bias-variance tradeoff, identify parametric and nonparametric methods.

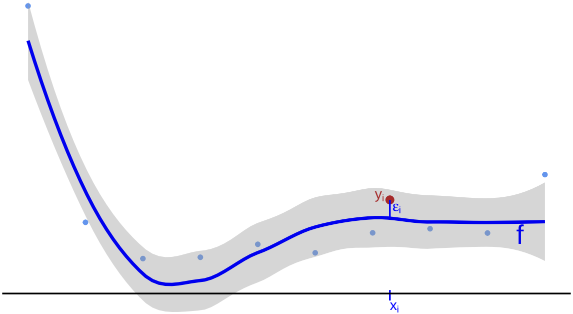
Recall: Function fitting

- Let $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$ be all the inputs.
- Let $f(\boldsymbol{x}_i)$ be input to output y_i relationship.

Recall: A diagram

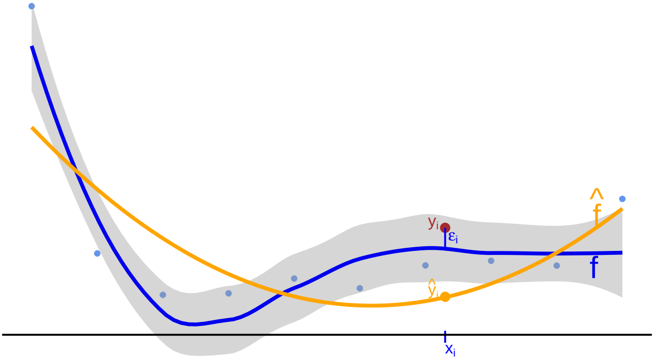
- We allow y_i to be different from $f(\boldsymbol{x}_i)$, perturbed by an observation-specific noise ϵ_i.

$$y_i = f(\boldsymbol{x}_i) + \epsilon_i.$$



Recall: Estimation and prediction

- In practice, we don't know f.
- We estimate it from the data and denote it $\hat{f}$.
- To predict y_i for some input $\boldsymbol{x}_i$, we'd use $\hat{y}_i = \hat{f}(x_i)$.



Recall: Source of error

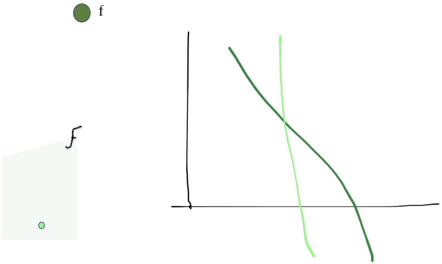
The process introduce two source of errors.

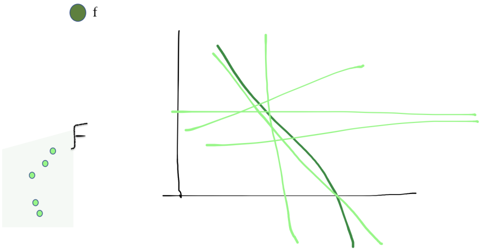
- Reducible error/Approximation error: $\hat{f}$ isn't close to f.
  - This error is reducible (using a better algorithm).
- Irreducible error: y_i isn't close to $f(\boldsymbol{x}_i)$.
  - Incur this error even f is known.

How do we estimate f?

Generally have two steps,

- Propose a model family $\mathcal{F}$.
  - E.g., set of all linear functions of x_i.
- Define a procedure to choose $\hat{f} \in \mathcal{F}$ based on the data.
  - E.g., the choice of $\hat{f}$ that minimizes $\sum_i \left(y_i - \hat{f}\,(x_i)\right)^2$.




f
F

Why we choose smaller size of $\mathcal{F}$?

- There is a bias-variance trade-off in fitting.
- Finding the best function in a large class $\mathcal{F}$ can be hard.

Why we choose smaller size of $\mathcal{F}$?

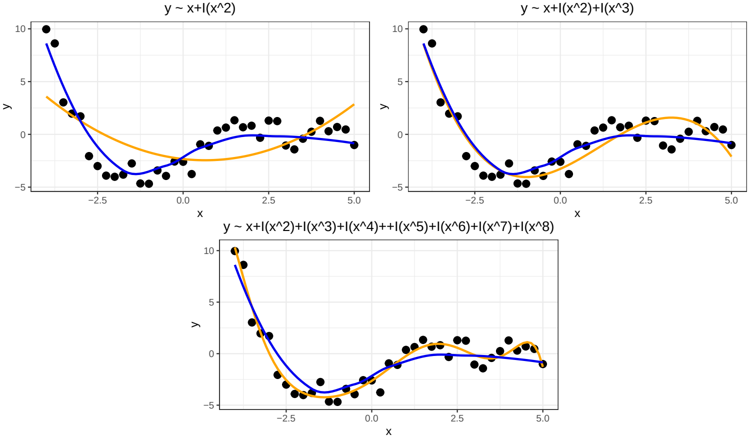


Figure 1: Blue line- true function, black points - training set, orange line - fitted function.

Why we choose smaller size of $\mathcal{F}$?

- We want our model to perform well on out-of-sample data.

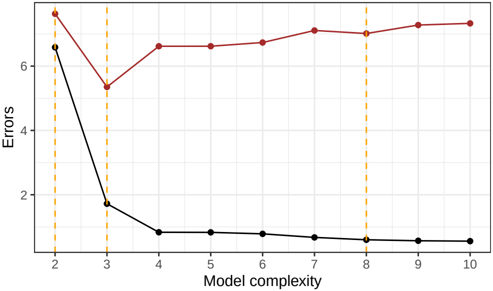
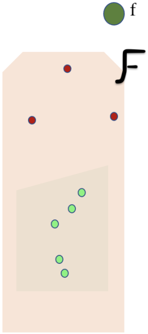
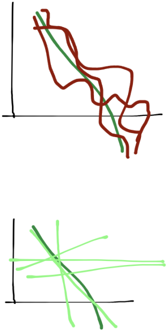


Figure 2: Black: in-sample, Brown: out-of-sample

Why we choose smaller size of $\mathcal{F}$?

- Finding the best function in a large class $\mathcal{F}$ can be hard.
  - High variance: different samples (x_i, y_i) might result in very different $\hat{f}$, even when f hasn't changed.



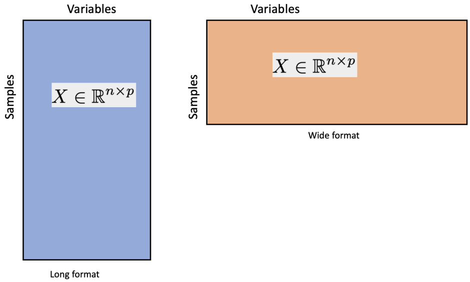


- This variance gets worse for the high-dimensional $\boldsymbol{x}_i$.
- Incurring bias, for the sake of better stability, can improve the predictions.

Discussion

- What are the advantages of more or less flexible regression models? When would you prefer one versus the other?
  - How does your answer depend on the input dimension of $\boldsymbol{x}_i$?
  - How does your answer depend on the sample size?

Discussion - note



- Long format (sample size $> p$ number of inputs).
  - More flexible models such as random forests, gradient boosting, deep learning.
- Wide format (sample size $< p$ number of inputs).
  - Less flexible models such as LASSO, Elastic net.

Discussion - note

- Summary
  - If we don't have too many samples, we should prefer a simpler model.
  - If we have many samples, we can afford a more complex model.

Post-training Analysis

- There's a certain set of checks we should always do after we fit a model, no matter what family it is.
  - Yes, even deep learning models.
- We can do better than looking at the validation loss.
  - Residual analysis, error modeling, outliers, high-leverage.

Residual Analysis

- Make a histogram of residuals $e_i = y_i - \hat{y}_i$.
- Plot residuals against a few input variables/fitted values.
    - If we notice systematic variation in them, this is information we can squeeze into f.

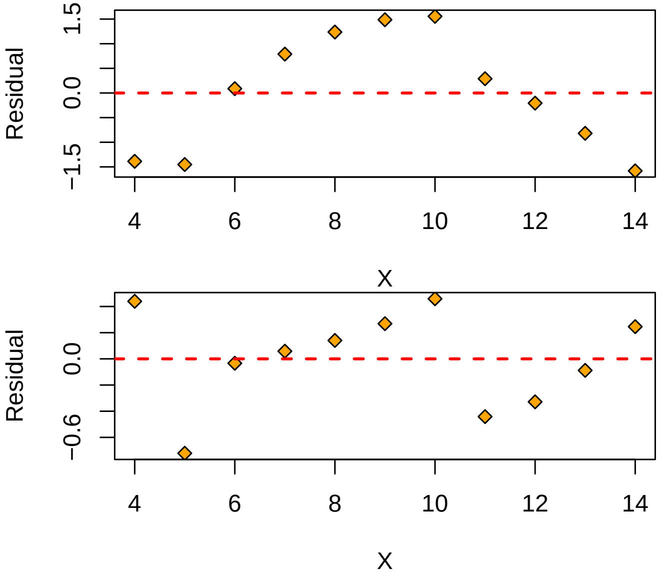


Figure 3: Above: residuals of linear fit (shows a quadratic pattern); Bottom: residuals of a quadratic fit.

Error Modeling

- We can use models to seek out systematic variation in e_i.
  - Cluster the x_i associated with large errors e_i.
  - Train a model with e_i as response and clusters as predictors.

Outliers and Leverage

- Look for outliers either in x_i or y_i directions.
- High leverage points are those that, if they were perturbed slightly, would dramatically alter the fit.

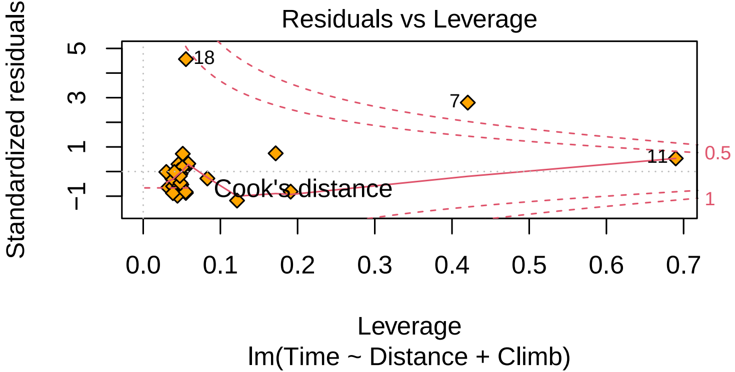


Figure 4: Data point 7 has leverage greater than 1.

Note

- Statistics and data science involved more than simply running a machine learning algorithm on data.
- Examples:
  - What is a question of interest? How can I collect data or design an experiment to address this question?
  - What inferences can be drawn from the data?
  - What actions should I take as a result of what I've learned?
  - Do I need to worry about bias, confounding, generalizability, concept drift (statistical properties of the response variable change over time), and etc.

Examples of Model Families

- Lab
  - Let's get a feel for how different model families look like.
  - We will use `Advertising` dataset (how does advertising affect sales?)

References

- Function Fitting Intro by Kris Sankaran.
- **ISLR** Chapter 2.