Principles of machine learning

Course: Analytics, Machine Learning, and the Digital Economy

- **Lecturer Radjabova Dilnora**

Prediction

- By producing a good estimate for f where the variance of ε is not too large, then we can make accurate predictions for the response variable, Y, based on a new value of $\boldsymbol{X}$.

Prediction

- We can predict Y using $\hat{Y} = \hat{f}(\boldsymbol{X})$

where $\hat{f}$ represents our estimate for f, and $\hat{Y}$ represents the resulting prediction for Y.

Prediction (cont.)

- The accuracy of $\hat{Y}$ as a prediction for Y depends on:
  - Reducible error
  - Irreducible error

Prediction (cont.)

- Note that $\hat{f}$ will not be a perfect estimate for f; this inaccuracy introduces error.

Prediction (cont.)

- Even if we could perfectly estimate f, there is still variability associated with ε that affects the accuracy of predictions = *irreducible* error.

Prediction (cont.)

- Even if we could perfectly estimate f, there is still variability associated with ε that affects the accuracy of predictions = *irreducible* error.

Prediction (cont.)

- Average of the squared difference between the predicted and actual value of Y.
- Var(ε) represents the *variance* associated with ε.

$$E[(Y-\hat{f}(X))^2|X=x] = \underbrace{[f(x)-\hat{f}(x)]^2}_{Reducible} + \underbrace{\mathrm{Var}(\epsilon)}_{Irreducible}$$

- Our aim is to minimize the reducible error!!

Example: Direct Mailing Prediction

- We are interested in predicting how much money an individual will donate based on observations from 90,000 people on which we have recorded over 400 different characteristics.

Example: Direct Mailing Prediction

- We don’t care too much about each individual characteristic.

Example: Direct Mailing Prediction

- Learning Problem:
  - For a given individual, should I send out a mailing?

Inference

- Instead of prediction, we may also be interested in the type of relationship between Y and the X's.

Inference

- Key questions:
  - Which predictors actually affect the response?
  - Is the relationship positive or negative?
  - Is the relationship a simple linear one or is it more complicated?

Example: Housing Inference

- We wish to predict median house price based on numerous variables.

Example: Housing Inference

- We want to *learn* which variables have the largest effect on the response and how big the effect is.

Example: Housing Inference

- For example, how much impact does the number of bedrooms have on the house value?

How do we estimate *f*?

- First, we assume that we have observed a set of **training data**.

$$\{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \ldots, (\mathbf{X}_n, Y_n)\}$$

- Second, we use the training data and a **machine learning method** to estimate *f*.
  – Parametric or non-parametric methods

Parametric Methods

- This reduces the *learning problem* of estimating the target function f down to a problem of estimating a set of **parameters**.

- This involves a two-step approach…

Parametric Methods (cont.)

- **Step 1:**
  – Make some assumptions about the functional form of *f*. The most common example is a linear model:

$$f(\mathbf{X}_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

  – In this course, we will examine far more complicated and flexible models for *f*.

Parametric Methods (cont.)

- **Step 2:**
  - We use the *training data* to fit the model (i.e. estimate f….the unknown parameters).

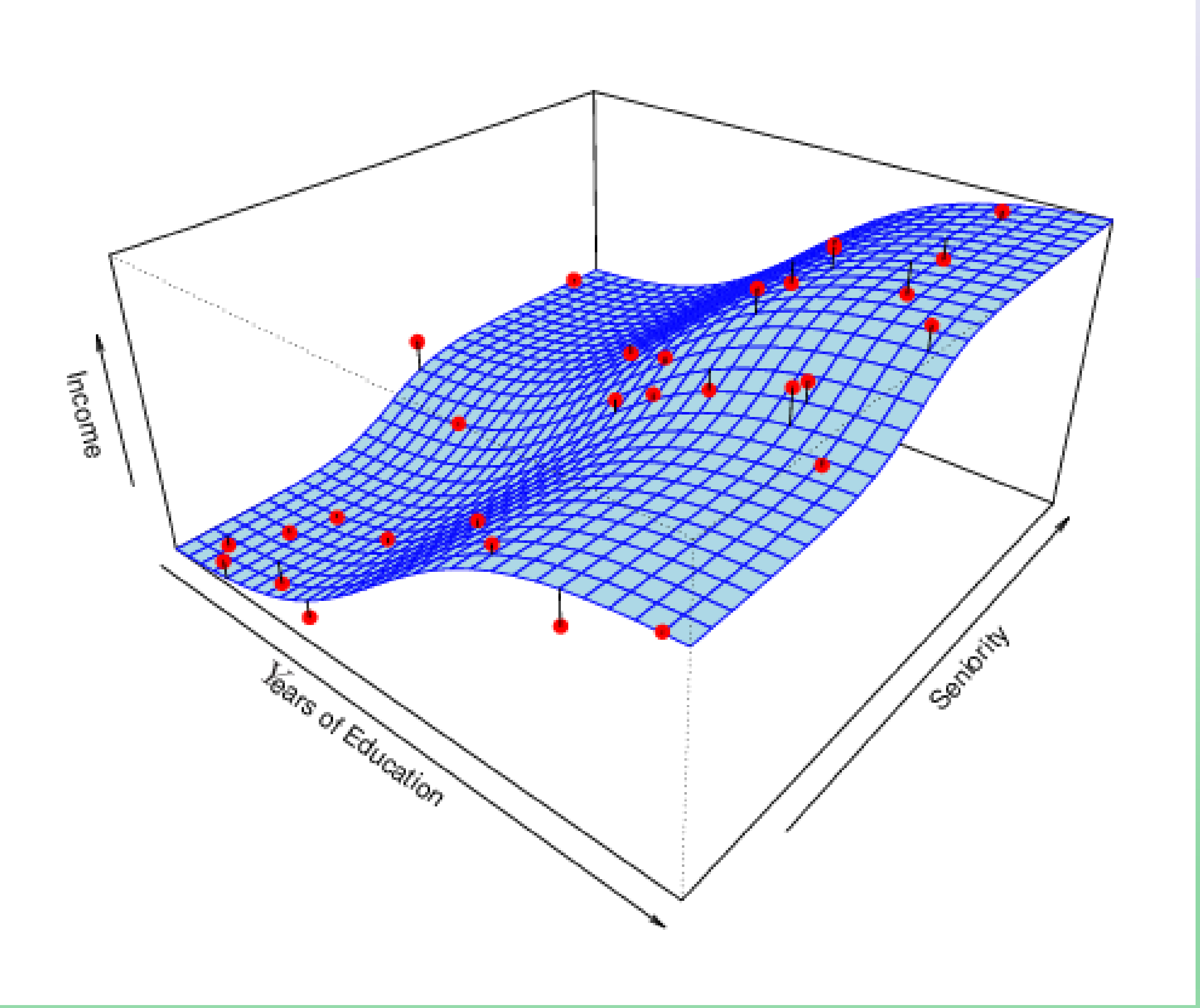
Parametric Methods (cont.)

- **Step 2:**
  - The most common approach for estimating the parameters in a linear model is via ordinary least squares (OLS) linear regression.

Parametric Methods (cont.)

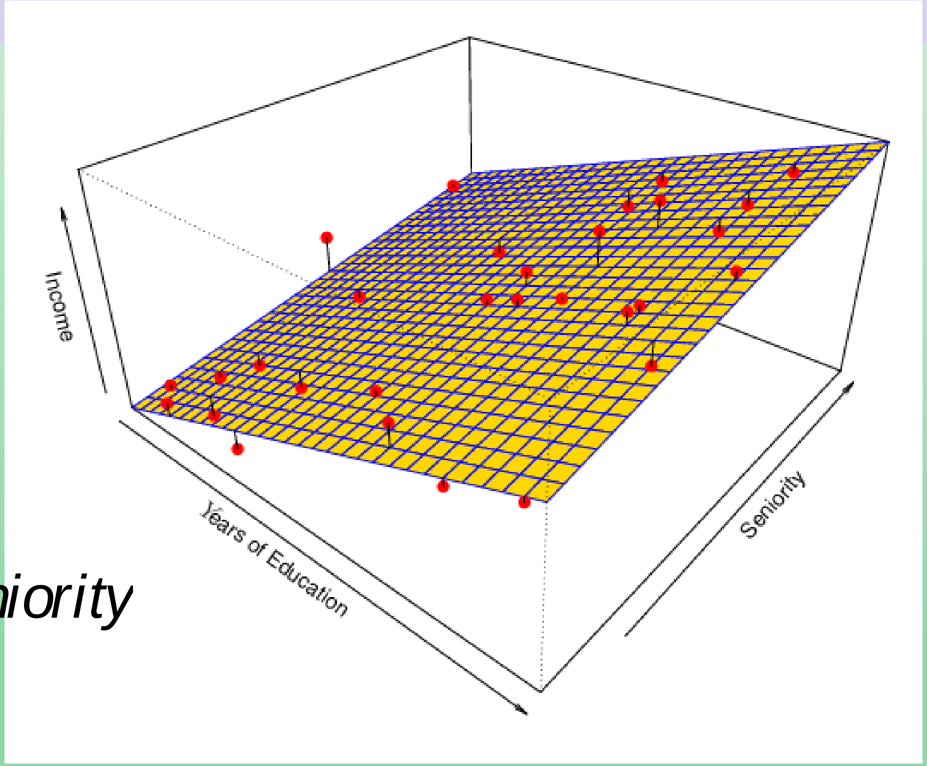
- **Step 2:**
  - However, there are superior approaches, as we will see in this course.

Example: Income vs. Education Seniority



Example: OLS Regression Estimate

- Even if the standard deviation is low, we will still get a bad answer if we use the incorrect model.

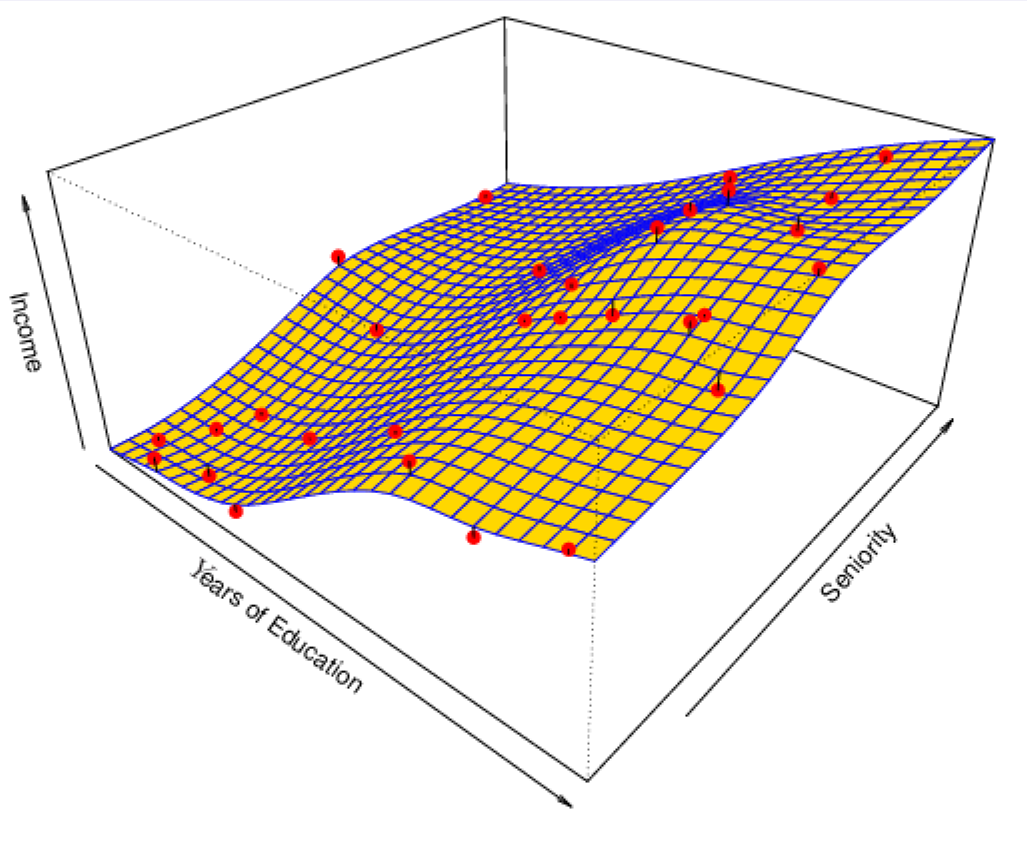


$$f = b_0 + b_1 \times Education + b_2 \times Seniority$$

Non-Parametric Methods

- As opposed to parametric methods, these do not make explicit assumptions about the functional form of f.
- Advantages:
  - Accurately fit a wider range of possible shapes of f.
- Disadvantages:
  - Requires a very large number of observations to acquire an accurate estimate of f.

Example: Thin-Plate Spline Estimate



- Non-linear regression methods are more flexible and can potentially provide more accurate estimates.
- However, these methods can run the risk of over-fitting the data (i.e. follow the errors, or noise, too closely), so too much flexibility can produce poor estimates for *f*.

Predictive Accuracy vs. Interpretability

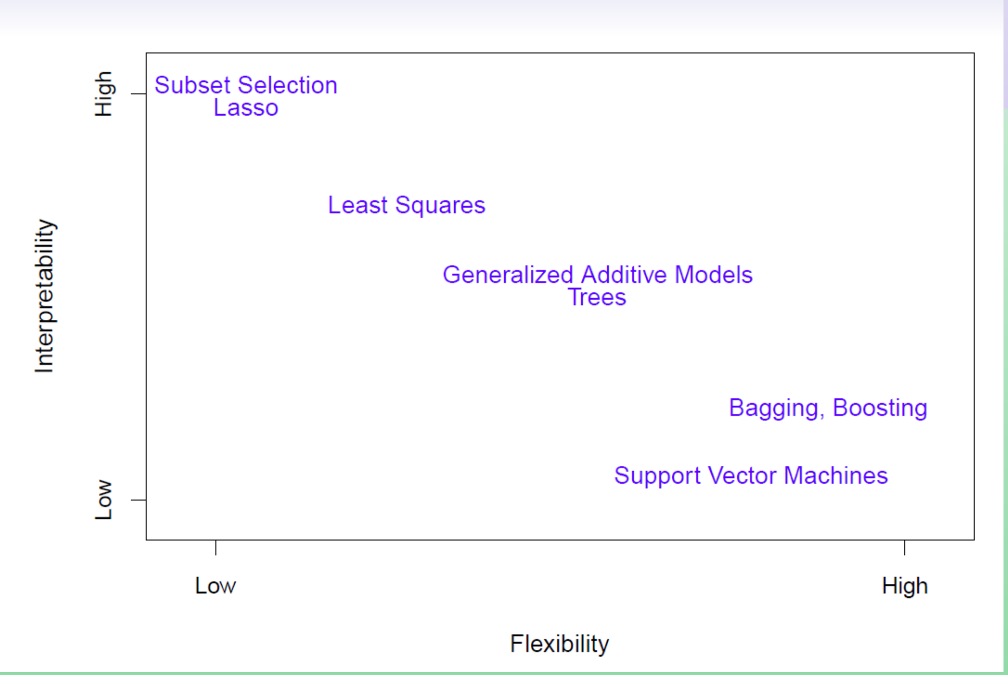
- Conceptual Question:
  - Why not just use a more flexible method if it is more realistic?

Predictive Accuracy vs. Interpretability

- Reason 1:
  - A simple method (such as OLS regression) produces a model that is easier to interpret (especially for inference purposes).

Predictive Accuracy vs. Interpretability (cont.)

- Reason 2:
  - Even if the primary purpose of learning from the data is for prediction, it is often possible to get more accurate predictions with a simple rather than a complicated model.



Reference

- 1. Shalev-Shwartz and Ben-David. Understanding Machine Learning: From Theory to Algorithms (Cambridge University Press, 2014)
- 2. Daum´e. A Course in Machine Learning.
- 3. The Art of Statistics: How to Learn from Data by David Shpigelter
- 4. Learning From Data – January 1, 2012 by Yaser S. Abu-Mostafa (Author), Malik Magdon-Ismail (Author), Hsuan-Tien Lin (Author)
- 5. Statistics: The Art and Science of Learning from Data by Alan Agresti
- 6. Learning From Data: An Introduction To Statistical Reasoning by M.Glenber.
- 7. Statistics: Learning from Data (with JMP Printed Access Card) by Rocky Pek
- 8. The Elements of Statistical Learning by Gerim Garold
- 9. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems 2nd Edition
- by Aurélien Géron (Author)