Class Notes

Statistical Computing & Machine Learning

Class 1

Statistical and Machine Learning

The two terms, "statistical learning" and "machine learning," reflect mainly the artificialities of academic disciplines.

- Statisticians focus on the statistical aspects of the problems
- Computer scientists are interested in "machines", including hardware and software.

"Data Science" is a new term that reflects the reality that both statistical and machine learning are about data. Techniques and concepts from both statistics and computer science are essential.

Statistics concepts

- Sampling variability
- Bias and variance
- Characterization of precision
- Function estimation frameworks, e.g. generalized linear models
- Assumed probability models
- Prior and posterior probabilities (Bayesian framework)

Computing concepts

- Algorithms
- Iteration
- Simulation
- Function estimation frameworks, e.g. classification trees, support vector machines, artificial intelligence techniques
- Kalman filters

Cross fertilization

- Assumed probability models supplanted by simulation
  - Randomization and iteration
  - Cross validation
  - Bootstrapping
- Model interpretibility Rather than an emphasis on the output of a function, interest in what the function has to say about how the world works.

Example 1: Machine translation of natural languages

Computer scientists took this on. * Identification of grammatical structures and tagging text. * Dictionaries of single-word equivalents, common phrases.

Story from early days of machine translation:

- Start with English: “The spirit is willing, but the flesh is weak.”
- Translate into Russian
- Translate back into English. Result: “The vodka is good, but the meat is rotten.”

Statistical approach:

- Take a large sample of short phrases in language A and their human translation into language B: the dictionary
- Find simple measures of similarity between phrases in language A (e.g. de-stemmed words in common)
- Take new phrase in language A, look up it’s closest match in the dictionary phrases in language A. Translation is the corresponding dictionary entry in language B

Where did the sample of phrases come from?

- European Union documents routinely translated into all the member languages. Humans mark correspondence.
- “Mechanical Turk” dispersal of small work tasks.

Result: Google translate.

Example 2: From library catalogs to latent semantic indexing

Early days: computer systems with key words and search systems (as in library catalogs)

Now: dimension reduction (e.g. singular value decomposition), angle between specific documents and what might be called “eigendocuments”

Result: Google search

Mathematics and Data

- Data tables: cases and variables.
- A quantitative variable is a vector.
- A categorical variable can be encoded as a set of “dummy” vectors.
- Response variable and explanatory variable

- The linear projection problem: find the point spanned by the explanatory variables that's closest to the response. That linear combination is the best-fitting model.
  - One explanatory and the response
  - Two explanatory on board and the response on the board (perfect, but meaningless fit)
  - Two explanatory in three-space and the response (residual likely)

Theoretical concepts ISL §2.1

Prediction versus inference

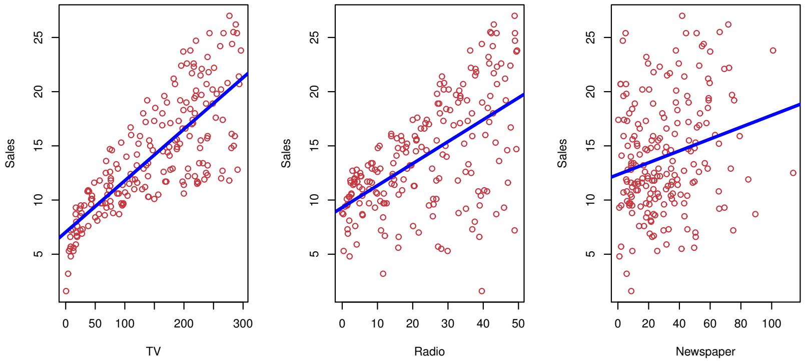
- Black-box predictions. If the box works, why worry about what's inside?
  - Example: Malignancy of cancer from appearance of cells. Works for guiding treatment. Does it matter why malignant cells have the appearance they do?
  - Story: Mid-1980s. Heart rate variability spectral analysis and holter monitors. (Holters were cassette tape recorders set to record ECG very, very slowly. Spectral analysis breaks down the overal signal into periodic components.) Very large spike at 0.03 Hz seen in people who will soon die.
- Causal influences. We want to use observations to inform our understanding of what influences what.
  - Story continued: The very large spike was the "wow and flutter" in the cassette tape mechanism. This had an exact periodicity: a spike in the spectrum. If the person was sick, their heart rate was steady: they had no capacity to vary it as other conditions in the body (arterial compliance, venus tone) called for. Understanding what happens in cardiac illness is, in part, about understanding how the various control systems interact.

Accuracy versus interpretability

Sometimes "flexibility" is used instead of "accuracy".

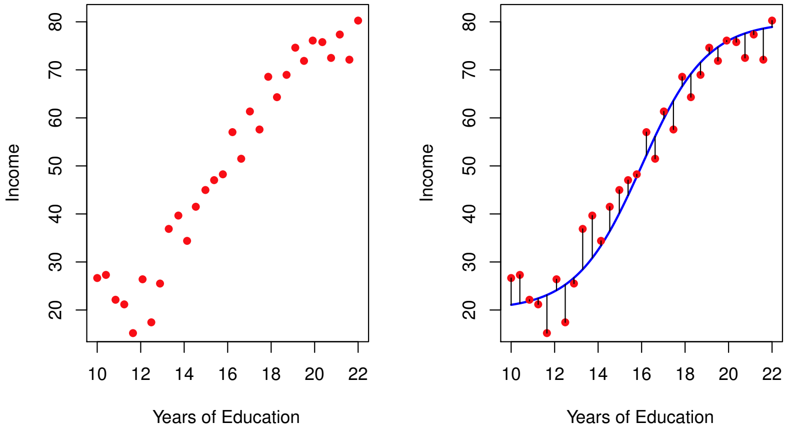
Not flexible:

Individual fits miss how the explanatory variables interact.



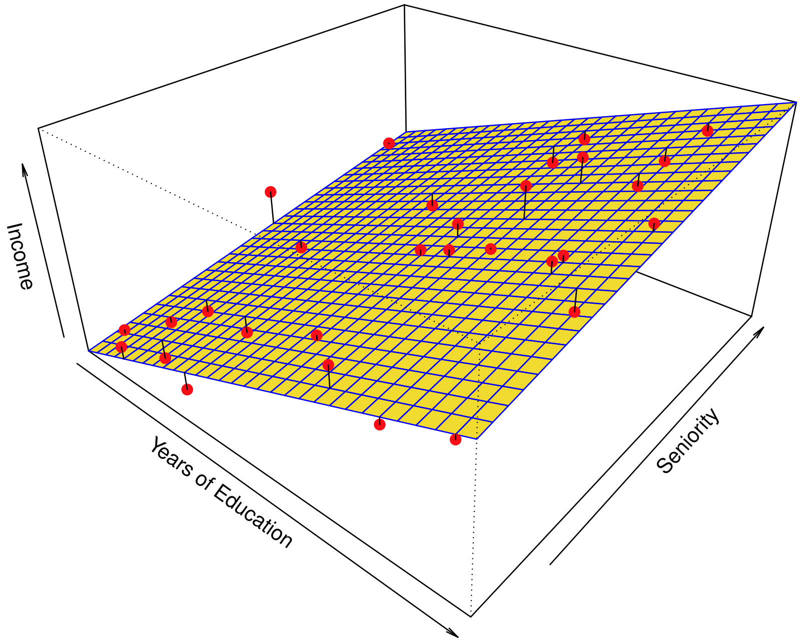
Flexible:

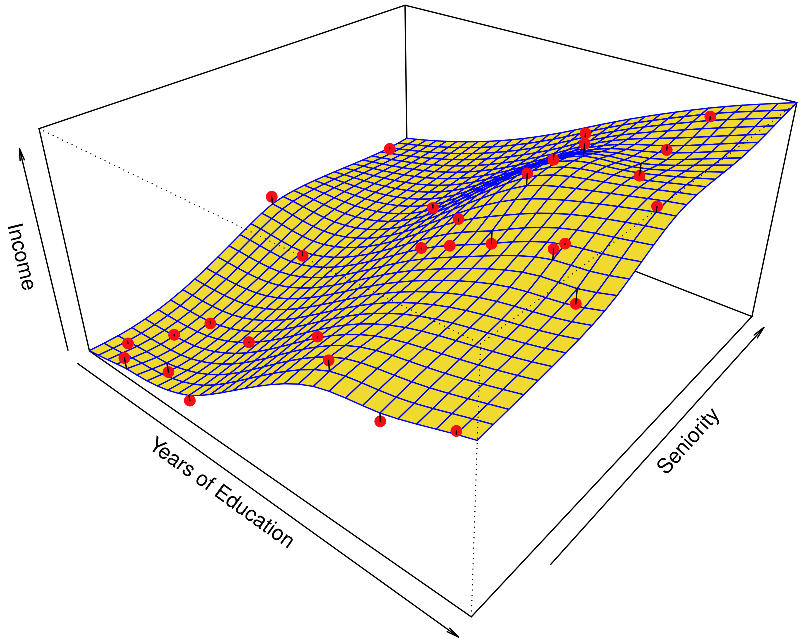
Such detailed patterns are more closely associated with physical science data than with social/economic data.



And in multiple variables:

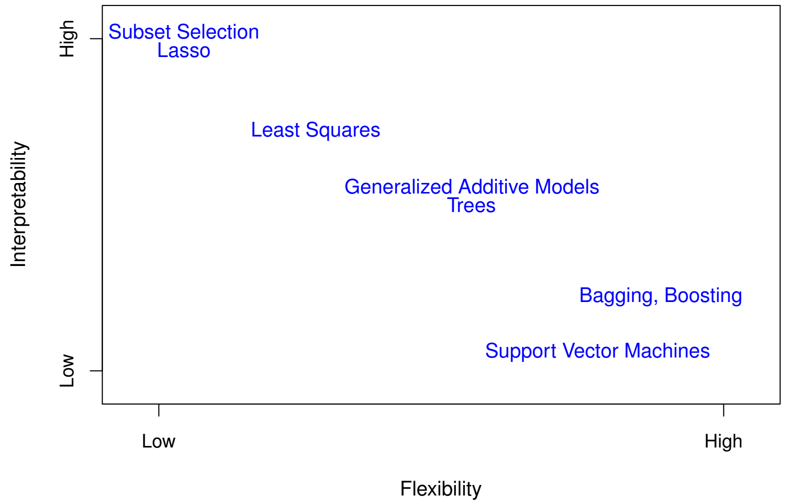
Not flexible:





Flexible:

A quick characterization of several model architectures (which they call "statistical learning methods")



Reducible versus irreducible error

What does this mean? (from p. 19)

$$\begin{aligned} E(Y-\hat{Y})^2 &= E[f(X)+\epsilon-\hat{f}(X)]^2 \\ &= \underbrace{[f(X)-\hat{f}(X)]^2}_{Reducible} + \underbrace{Var(\epsilon)}_{Irreducible} \end{aligned}$$

Notation:

- X — the inputs that determine the output Y.
- Y — the output, that is, the quantity we want to predict
- $\hat{Y}$ — our prediction
  - hat means estimated, no hat means "real" (whatever that might mean)
- $E(Y-\hat{Y})^2$ — the mean of the square difference between our prediction and the "real" value. E means "expectation value."
- $f(X)$ — what Y would be, ideally, for a given X
- $\hat{f}(X)$ — our estimate of $f(X)$
- ϵ — but Y is subject to other, "random" influences. ϵ represents these. ϵ is a misleading notation because it may not be at all small in practice. But ϵ is alway centered on zero (by definition).
- $|f(X)-\hat{f}(X)|$ — the magnitude of the difference between the "real" $f()$ and our estimate. This can be made small by

1. collecting more data
2. using a more flexible model

3. expanding the set of inputs considered

- $Var(\epsilon)$ — the "variance" of ϵ. This is the mean square of ϵ, that is, $E(\epsilon^2)$.

Regression versus classification

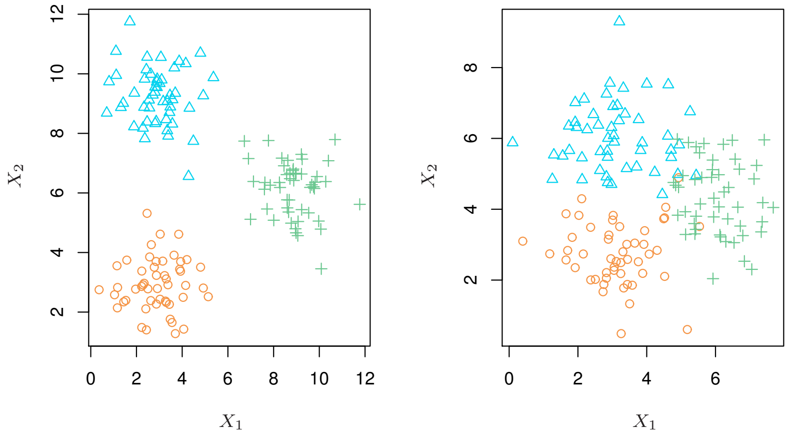
Regression: quantitative response (value, probability, count, ...)

Classification: categorical response with more than two categories. (When there are just two categories, regression (e.g. logistic regression) does the job.)

Supervised versus unsupervised

- Demographics groups in marketing.
- Poverty vs middle-class
- Political beliefs ... left vs right?

ISL Figure 2.8



In-class programming activity

Day 1 activity