An Overview of Statistical Learning

Lu Haibo

Mondy, Feb 25, 2019

What is Statistical Learning?

Statistical learning refers to a vast set of tools for *understanding data*

- supervised
  - predicting an *output* based on *inputs*
- unsupervised
  - there are inputs but no supervising output, learn *relationships* and *structure*

Main Problems In Statistical Learning

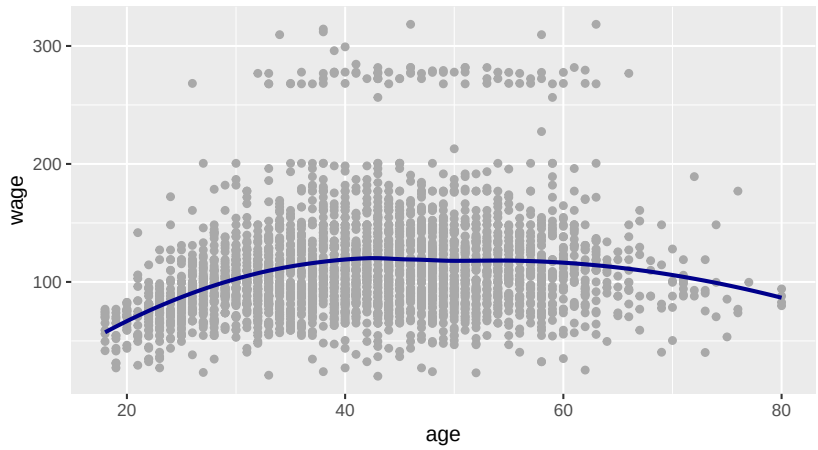
- *Regression* problem
  - predict a *continuous* or *quantitative* output
- *Classification* problem
  - predict a *categorical* or *qualitative* output
- *Clustering* problem
  - not trying to predict an output variable

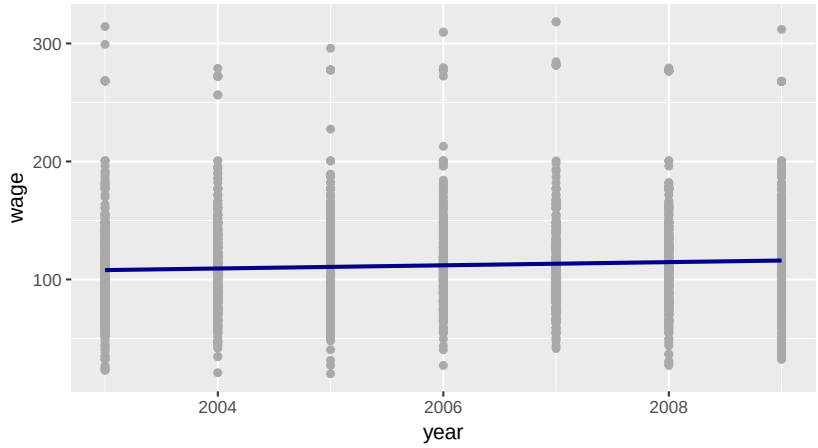
Example:

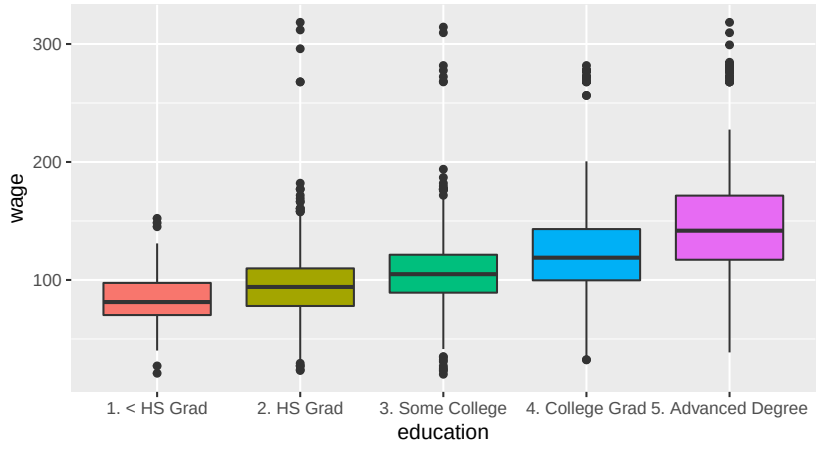
- Wage Data

Wish to understand the association between an employee's `age` and `education`, as well as the calender `year`, on his `wage`.

	year	age	education	wage
231655	2006	18	1. < HS Grad	75.04
86582	2004	24	4. College Grad	70.48
161300	2003	45	3. Some College	130.98
155159	2003	43	4. College Grad	154.69
11443	2005	50	2. HS Grad	75.04
376662	2008	54	4. College Grad	127.12



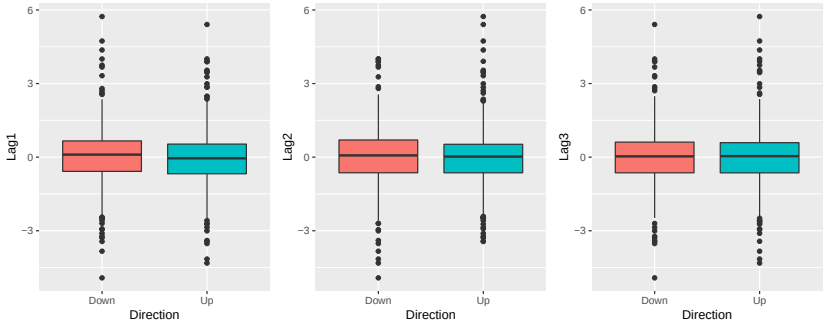




- Stock Market Data

We wish to predict whether the index will *increase* or *decrease* on a given day using the past 5 days' percentage changes in the index.

Year	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today	Direction
2001	0.38	-0.19	-2.62	-1.05	5.01	1.19	0.96	Up
2001	0.96	0.38	-0.19	-2.62	-1.05	1.30	1.03	Up
2001	1.03	0.96	0.38	-0.19	-2.62	1.41	-0.62	Down
2001	-0.62	1.03	0.96	0.38	-0.19	1.28	0.61	Up
2001	0.61	-0.62	1.03	0.96	0.38	1.21	0.21	Up
2001	0.21	0.61	-0.62	1.03	0.96	1.35	1.39	Up



- Gene Expression Data

We consider the `NCI60` data set, which consists of 6830 gene expression measurements for each of 64 cancer cell lines. Instead of predicting a particular output variable, we are interested in determining whether there are groups, or clusters, among the cell lines based on their gene expression measurements.

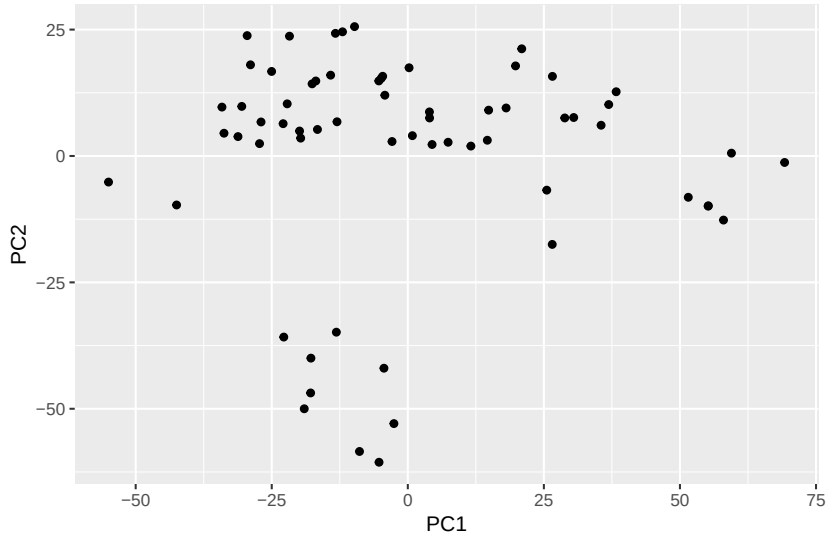


Figure 1: Representation of the NCI60 gene expression data set in a two-dimensional space, PC1 and PC2.

What Is Statistical Learning?

Example: To develop an accurate model that can be used to predict sales on the basis of the three median budgets.

TV	Radio	Newspaper	Sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9
8.7	48.9	75.0	7.2

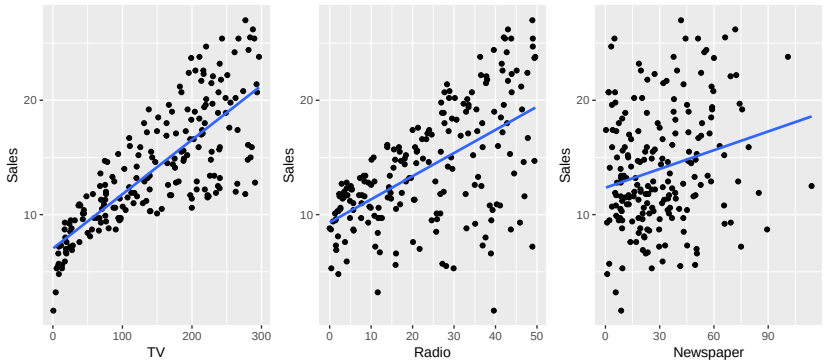


Figure 2: Simple least squares fit of sales to each predictor

- *input variables*: advertising budgets, `TV`, `Radio`, `Newspaper`
  - typically denoted as X
  - predictor, independent variables, features
- *output variables*: `sales`
  - typically denoted as Y
  - response, dependent variable

Suppose that we observe a quantitative response Y and p different predictors, $X_1, X_2, \ldots, X_p$. We assume that there is some relationship between Y and $X = (X_1, X_2, \ldots, X_p)$, 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 essence, statistical learning refers to a set of approaches for estimating f.

Why Estimate f?

Two main reasons that we may wish to estimate f

- *Prediction*
  - a set of inputs X are readily available, but the output Y cannot be easily obtained, since the error term has mean zero, we can predict Y using
    $$\hat{Y} = \hat{f}(X)$$
  - $\hat{f}$ treated as a *black box*
  - don't concern the exact form of $\hat{f}$, provided that it yields accurate predictions for Y
  - the accuracy of $\hat{Y}$ depends on two quantities
    - *reducible error*: $\hat{f}$ do not estimate f perfectly
    - *irreducible error*: even if we estimate f perfectly, so that $\hat{Y} = f(X)$, there is still some error from ϵ that cannot be predicted using X
  - Why irreducible error? ϵ may contain
    - unmeasured variables that are useful in predicting Y, and we don't measure them
    - unmeasurable variation
  - Consider a given estimate $\hat{f}$ and a set of predictors X, which yields the prediction $\hat{Y} = \hat{f}(X)$, then
    $$\begin{aligned} E(Y - \hat{Y})^2 &= E[f(X) + \epsilon - \hat{f}(X)]^2 \\ &= \underbrace{[f(X) - \hat{f}(X)]^2}_{Reducible} + \underbrace{Var(\epsilon)}_{Irreducible} \end{aligned}$$
- *Inference*
  - our goal is not necessarily to make predictions for Y, but to understand how Y chages as a function of $X_1, \ldots, X_p$.
  - $\hat{f}$ cannot be treated as a black box, we need to know its exact form
  - one may be interested in answering
    - *Which predictors are associated with the response?*: Identifying the important predictors among a large set of possible variables.
    - *What is the relationship between the response and each predictor?*: negative, positive, linear, nonlinear…
    - *Can the relationship between Y and eachpredictor be adequately summarized using a linear equation, or is the relationship more complicated?*

How Do We Estimate f?

- *training data*: observations, $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$, we will use to train, or teach, our method how to estimate f
- statistical learning methods for find a function $\hat{f}$ such that $Y \approx \hat{f}(X)$
  - Parametric Methods
    - make an assumption about the functional form of f. For example:
      $$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$
    - instead of estimating an entirely arbitrary function $f(X)$, one only needs to estimate the $p + 1$ coefficients $\beta_0, \beta_1, \ldots, \beta_p$
    - the potential disadvantage of a parametric approach is that the model we choose will usually not match the true unknown form of f
    - *flexibility*
      - a less flexible model: poor estimate
      - a more flexible model: *overfitting*, follow the errors, or *noise*, too closely

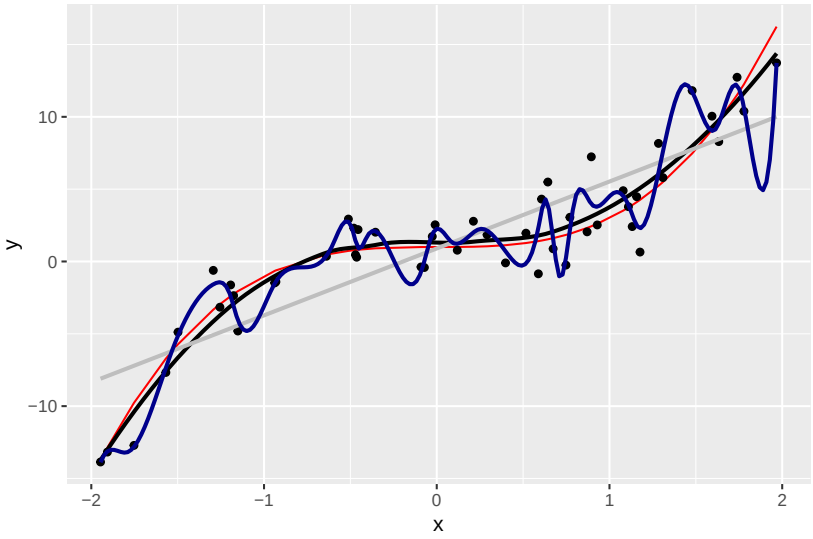


Figure 3: The green curve represent the true underlying relationship $f(X) = 1 + 2x^3$. The red line represent the simple linear regression line. The black curve represent the loess curve. The darkblue curve is a bspline regression with df=30.

- Non-parametric Methods - do not make explicit assumptions about the functional form of f - seek an estimate of f that gets as close to the data points as possible without being too rough or wiggly - avoid the assumption of a particular functional form for f - since they do not reduce the problem of estimating f to a small number of

parameters, a very large number of observations is required in order to obtain an accurate estimate for f - overfitting: should select a level of smoothness

The Trade-Off Between Prediction Accuracy and Model Interpretability

- *Why would we ever choose to use a more restrictive method instead of a very flexible approach?*
  - when inference is the goal, restrictive models are much more interpretable.
  - even if we were only interested in prediction, we will often obtain more accurate predictions using a less flexible method. (the potential for overfitting in highly flexible methods)

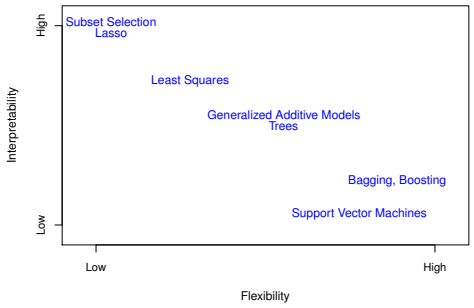


Figure 4: A representation of the trade off between flexibility and interpretability, using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases.

Assessing Model Accuracy

- *There is no one method dominates all other over all possible data sets.*
- It's important to decide for any given set of data which method produces the best result.

Measuring The Quality of Fit

- *mean squared error* (MSE): the most commonly-used measure in the regression setting

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

- *training MSE*: the MSE computed using the training data that was used to fit the model

- *test MSE*: the MSE computed using the test data that was not used to train the model
- select the model for which has the smallest test MSE
- there is no guarantee that the method with the lowest training MSE will also have the lowest test MSE. For the methods with quite small training MSE can have much larger test MSE

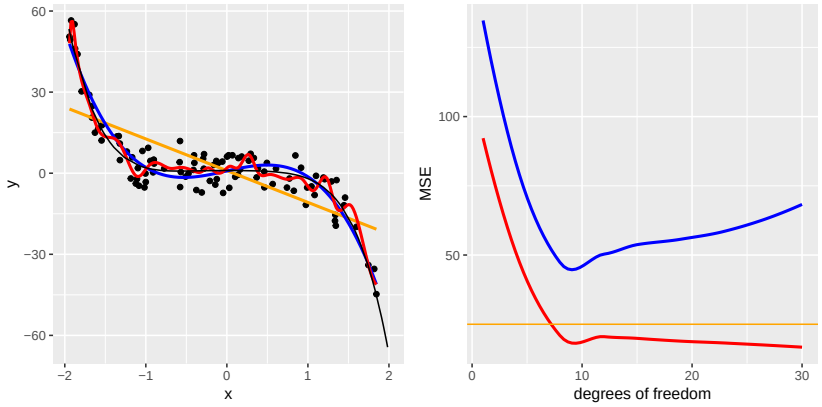


Figure 5: Left: Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange), and two smoothing spline fits (blue and red). Right: Traning MSE (red), test MSE (blue), and the irreducible error (orange line).

The Bias-Variance Trade-Off

The expected test MSE, for a given value x_0, can be decomposed into

$$E(y_0 - \hat{f}(x_0))^2 = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\epsilon)$$

- $E(y_0 - \hat{f}(x_0))^2$ defines the *expected test MSE*, refers to the average test MSE that we would obtain if we repeatedly estimated f using a large number of training sets, and tested each at x_0.
- *Variance*
  - refers to the amount by which $\hat{f}$ would change if we estimated it using a different training data set.
  - a method has high variance then small changes in the training data can result in large changes in $\hat{f}$.
  - in general, more flexible statistical methods have higher variance.
- *bias*
  - refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model.
  - generally, more flexible methods result in less bias.
- As a general rule, as we use more flexible methods, the variance will increase and the bias will decrease. The relative rate of change of these two quantities determines whether the test MSE increases or decreases.

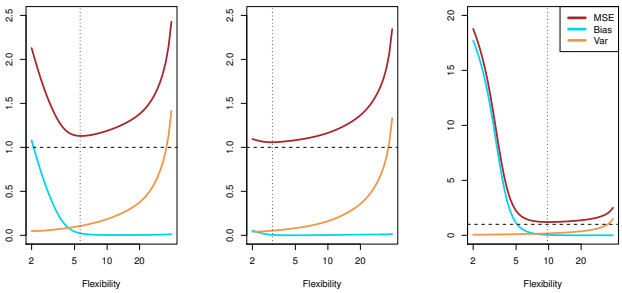


Figure 6: Squared bias (blue curve), variance (orange curve), $Var(\epsilon)$ (dashed line), and test MSE (red curve) for the three types of data sets. The vertical dotted line indicates the flexibility level corresponding to the smallest test MSE.

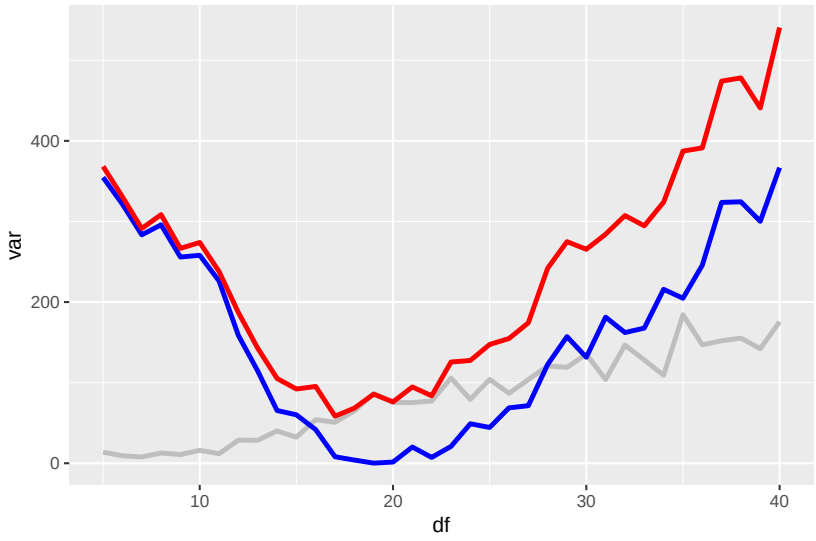


Figure 7: Bias Variance trade-off: variance (gray), bias2 (blue), test-mse (red)

The Classification Setting

The most common approach for quantifying the accuracy of our estimate $\hat{f}$ is the training *error rate*

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$