# Statistical Learning

Lecture 01b

ANU - RSFAS - AHW

Last Updated: Tue Feb 22 14:58:43 2022

# Statistical Learning Problems

- Identify the risk factors for prostate cancer.
- Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements.
- Customize an email spam detection system (spam or ham).
- Identify the numbers in a handwritten zip code.
- Classify a tissue sample into one of several cancer classes, based on a gene expression profile.
- Establish the relationship between salary and demographic variables in population survey data.
- Classify the pixels in a LANDSAT image, by usage.

# The Supervised Learning Problem

Starting point:

- Outcome measurement Y (also called dependent variable, response, target).
- Vector of $p$ predictor measurements $X$ (also called inputs, regressors, covariates, features, independent variables).
- In the regression problem, $Y$ is quantitative (e.g price, blood pressure).
- In the classification problem, $Y$ takes values in a finite, unordered set (survived/died, digit 0-9, cancer class of tissue sample).
- We have training data $(x_1, y_1), \ldots, (x_N, y_N)$. These are observations (examples, instances) of these measurements.

# Objectives

On the basis of the training data we would like to:

- Accurately predict unseen test cases.
- Understand which inputs affect the outcome, and how.
- Assess the quality of our predictions and inferences.

# Philosophy

- It is important to understand the ideas behind the various techniques, in order to know how and when to use them.
- One has to understand the simpler methods first, in order to grasp the more sophisticated ones.
- It is important to accurately assess the performance of a method, to know how well or how badly it is working [simpler methods often perform as well as fancier ones!]
- This is an exciting research area, having important applications in science, industry and finance.

# Unsupervised Learning

- No outcome variable, just a set of predictors (features) measured on a set of samples.
- Objective is more fuzzy – find groups of samples that behave similarly, find features that behave similarly, find linear combinations of features with the most variation.
- Difficult to know how well your are doing.
- Different from supervised learning, but can be useful as a pre-processing step for supervised learning.

# The Netflix Prize

- The Competition started in October 2006. **Training data is ratings for 18,000 movies by 400,000 Netflix customers, each rating between 1 and 5**.
- The training data is very sparse! **About 98% missing.**
  - real data are messy
  - missing data is a large area of research in statistics
- The objective is to predict the rating for a set of 1 million customer-movie pairs that are missing in the training data.
- Netflix's original algorithm achieved a **root MSE of 0.953**.
- The first team to achieve a 10% improvement wins one million dollars.
- **Is this a supervised or unsupervised problem?**

# Statistical Learning versus Machine Learning

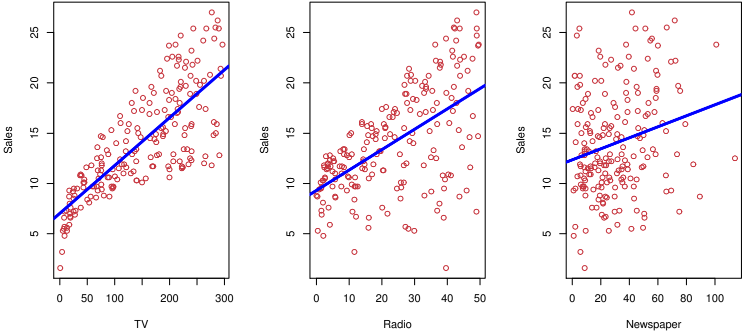
- Machine learning arose as a subfield of Artificial Intelligence.
- Statistical learning arose as a subfield of Statistics.
- There is much overlap – both fields focus on supervised and unsupervised problems:
    - Machine learning has a greater emphasis on large scale applications and prediction accuracy.
    - Statistical learning emphasizes models and their interpretability, and precision and uncertainty.
- But the distinction has become more and more blurred, and there is a great deal of "cross-fertilization".
- If you pick up a machine learning book . . . at least 85% of the ideas were developed in statistics!

# Nate Silver

- He has popular science book *The Signal and the Noise*.
- "Silver successfully called the outcomes in 49 of the 50 states in the 2008 U.S. Presidential election, he was named one of The World's 100 Most Influential People by Time in 2009".
  https://en.wikipedia.org/wiki/Nate_Silver
- Data journalism – Five thirty Eight: http://fivethirtyeight.com
  - There 538 electors in the US Electoral College
    - Electoral votes for each of the 50 states are determined by the number of representatives
    - 435 house members; 100 senate members
    - DC gets 1 house member and two senate members as if it were a state for presidential voting
- 2016 US election forecast:
  http://projects.fivethirtyeight.com/2016-election-forecast/
- 2020 US Election forecast:
  https://projects.fivethirtyeight.com/2020-election-forecast/

What does a 70% prediction mean?

# What is Statistical Learning?



- Shown are **Sales** (in thousands of units) vs **TV, Radio and Newspaper** budgets in thousands of USD), with a blue linear-regression line fit **separately** to each.

- Can we predict **Sales** using these three?

$$\text{Sales} \approx f(\text{TV, Radio, Newspaper})$$

- Here **Sales** is a response or target that we wish to predict. We generically refer to the response as $Y$.
- **TV** is a feature, or input, or predictor, we name it $X_1$.
- Likewise name **Radio** as $X_2$, and so on.
- We can refer to the input vector collectively as:

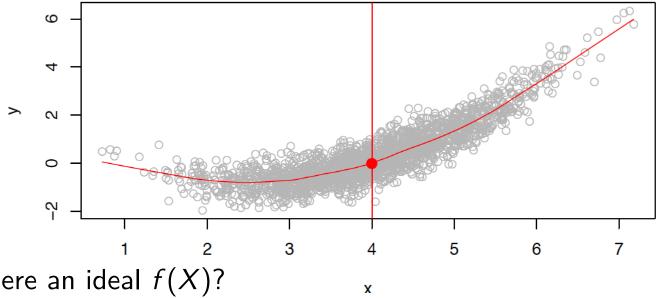
$$X = (X_1, X_2, X_3)$$

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

- where $\epsilon$ captures measurement errors and other discrepancies.

# What is $f(X)$ good for?

- With a good $f$ we can make predictions of $Y$ at new points $X = x$.
- We can understand which components of $X = (X_1, X_2, \ldots, X_p)$ are important in explaining $Y$, and which are irrelevant.
  - E.g. **Seniority** and **Years of Education** have a big impact on **Income**, but **Marital Status** typically does not.



- Is there an ideal $f(X)$?
- In particular, what is a good value for $f(X)$ at any selected value of $X$, say $X = 4$?
- There can be many $Y$ values at $X = 4$. A good value is

$$f(4) = E(Y|X = 4)$$

- This is the expected value (average) of $Y$ given $X = 4$.
- $f(x) = E(Y|X = 4)$ is called the regression function.

# The Regression Function

- We can extend the idea to a vector $X$:

$$f(x) = f(x_1, x_2, x_3) = E(Y|X_1 = x_1, X_2 = x_2, X_3 = x_3)$$

- Why is the expected value a good estimator? It is the optimal predictor of $Y$ with regard to mean-squared prediction error.

- $f(x) = E(Y|X = x)$ is the function that minimizes

$$E[(Y - g(x))^2|X = x]$$

for all functions $g$ at all points $X = x$.

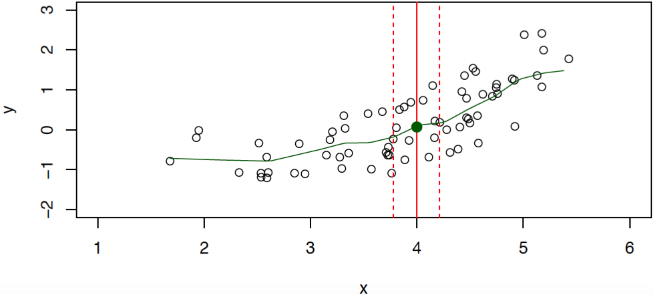
- $\epsilon = Y - f(x)$ is the irreducible error
  - i.e. even if we knew $f(x)$, we would still make errors in prediction, since at each $X = x$ there is typically a distribution of possible $Y$ values.

# How to estimate $f$?

- Typically we have few if any data points with $X = 4$ exactly.
- So we cannot compute $E(Y|X = x)$!
- Relax the definition and let

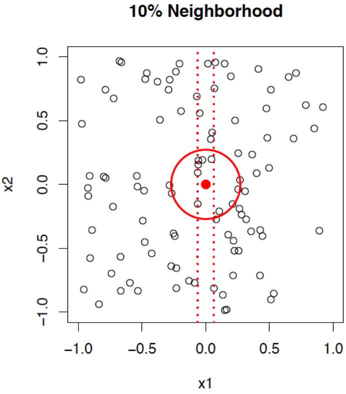
$$\hat{f}(x) = \text{Ave}(Y|X \in \mathcal{N}(x))$$

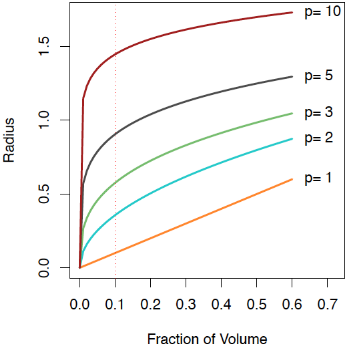
where $\mathcal{N}(x)$ is some neighborhood of $x$.



- Nearest neighbor averaging can be pretty good for small p – i.e. $p \leq 4$ and largish $N$.
- We will discuss smoother versions, such as kernel and spline smoothing later in the course.
- Nearest neighbor methods can be lousy when $p$ is large.
- Reason: the curse of dimensionality. Nearest neighbors tend to be far away in high dimensions.
  - We need to get a reasonable fraction of the $N$ values of $y_i$ to average to bring the variance down – maybe we want 10% of the data.
  - A 10% neighborhood in high dimensions need no longer be local, so we lose the spirit of estimating $E(Y|X = x)$ by local averaging.

# The Curse of Dimensionality





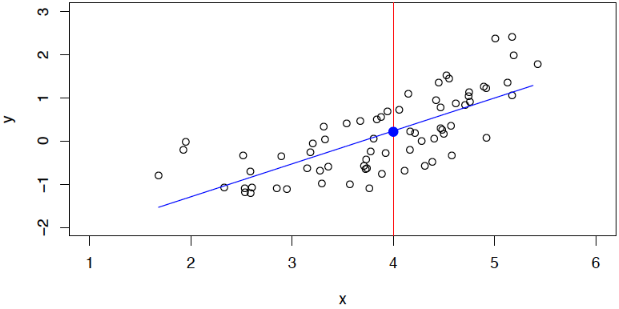
# Parametric and Structured Models

The linear model is an important example of a parametric model:

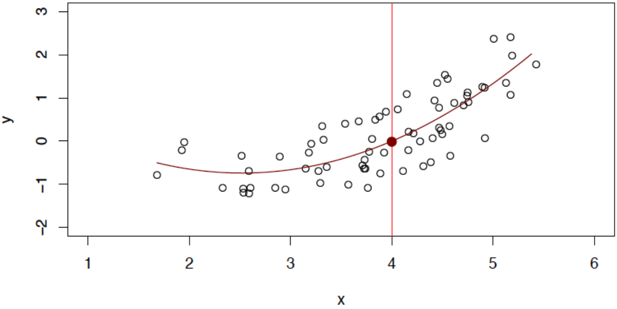
$$f_L(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_p X_p$$

- A linear model is specified in terms of $p + 1$ parameters $(\beta_0, \beta_1, \dots, \beta_p)$
- We estimate the parameters by fitting the model to training data.
- **Although it is almost never correct, a linear model often serves as a good and interpretable approximation to the unknown true function $f(X)$.**

$$f_L(X) = \hat{\beta}_0 + \hat{\beta}_1 X$$



$$f_Q(X) = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2$$

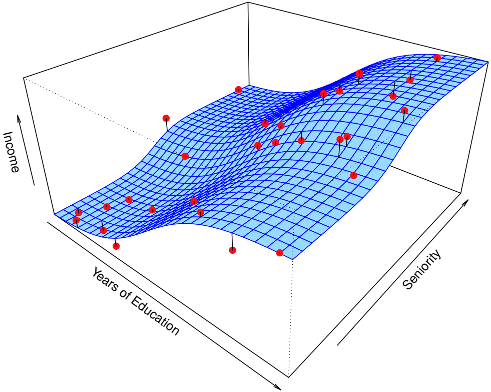


# We can extend these ideas.

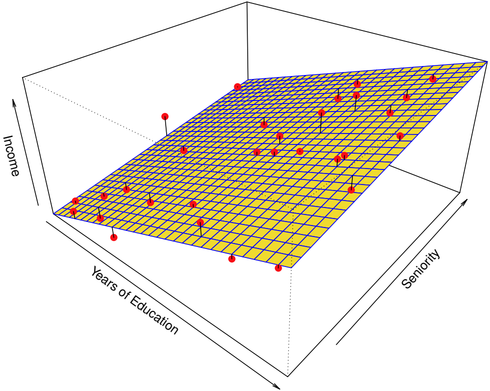
- Blue is the true relationship:

$$\text{Income} \approx f(\text{Years of education}, \text{Seniority})$$

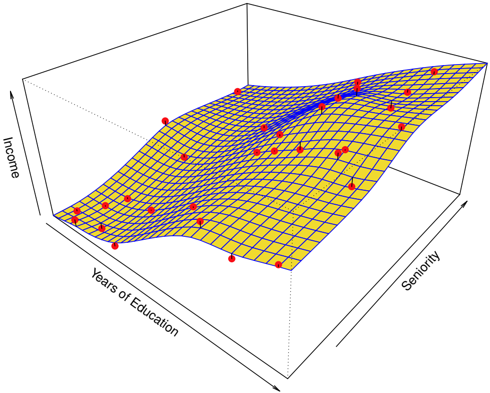
- Red dots are observed values for 30 individuals



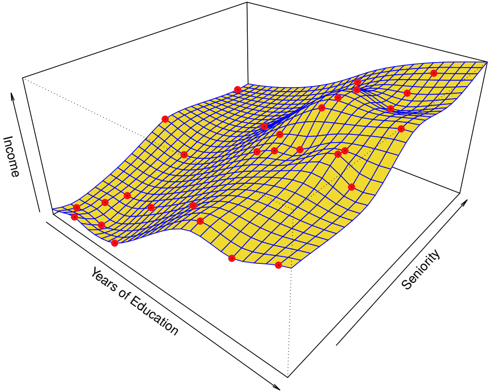
- Linear model fit



- Flexible surface



- Even more flexible.
- Here the fitted model makes no errors on the training data! Also known as overfitting.



# Some Trade-Offs

- Prediction accuracy versus interpretability – Linear models are easy to interpret
- Good fit versus over-fit or under-fit.
- How do we know when the fit is just right?
- Parsimony versus black-box.
  - We often prefer a simpler model involving fewer variables over a black-box predictor involving them all.

# Assessing Model Accuracy

- Suppose we fit a model $\hat{f}(x)$ to some training data:

$$\mathrm{Tr} = \{x_i, y_i\}_1^N ,$$

and we wish to see how well it performs.

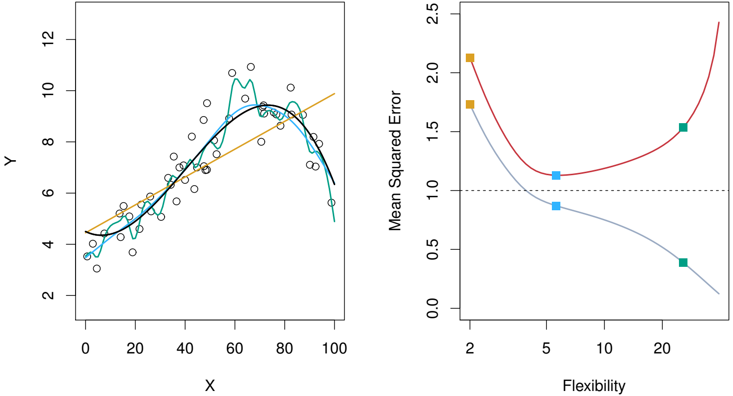
- We could compute the average squared prediction error over Tr:

$$MSE_{\mathrm{Tr}} = \mathrm{Ave}_{i \in \mathrm{Tr}} \left[ y_i - \hat{f}(x_i) \right]^2$$

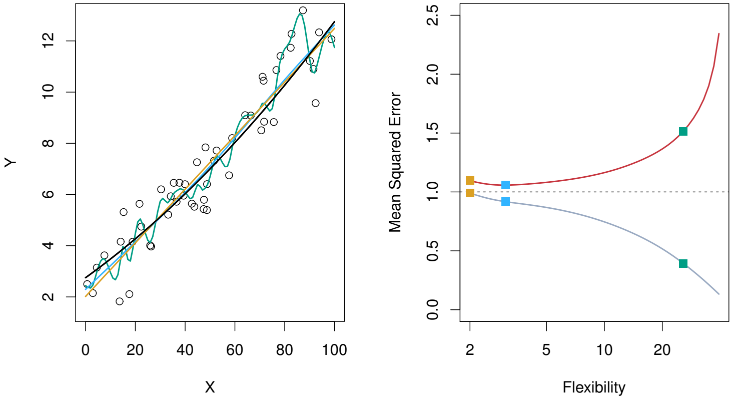
- This will lead to overfitting.

- Instead we should, if possible, compute it using fresh test data: $\mathrm{Te} = \{x_i, y_i\}_1^M$

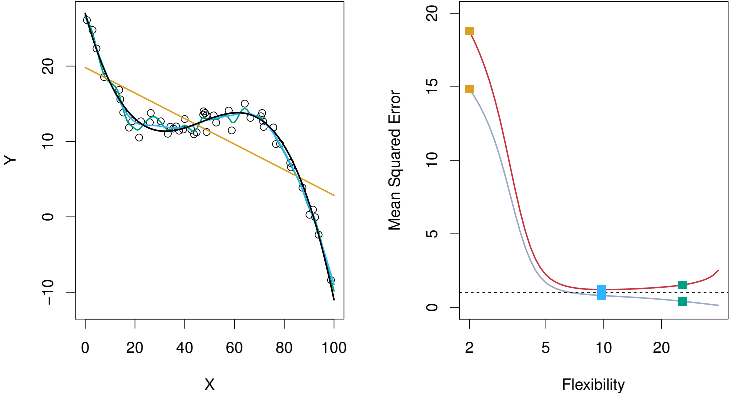
$$MSE_{\mathrm{Te}} = \mathrm{Ave}_{i \in \mathrm{Te}} \left[ y_i - \hat{f}(x_i) \right]^2$$



- Black curve is truth.
- Red curve on right is MSE of **Te** data, grey curve is MSE of **Tr** data.
- Orange, blue and green curves/squares correspond to fits of different flexibility.



- Here the truth is smoother, so the smoother fit and linear model do really well.



- Here the truth is wiggly and the noise is low, so the more flexible fits do the best.

# Bias-Variance Trade-Off

- Suppose we have fit a model $\hat{f}(x)$ to some training data **Tr**, and let $(x_0, y_0)$ be a **test observation** drawn from the population.

- If the true model is $Y = f(X) + \epsilon$ (with $f(x) = E(Y|X = x)$), then

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

- The expectation averages over the variability of $y0$ as well as the variability in **Tr**.

- Typically as the flexibility of $\hat{f}$ increases, its variance increases, and its bias decreases.

- So choosing the flexibility based on average test error amounts to a bias-variance trade-off.