Trevor Hastie
Robert Tibshirani

The Elements of Statistical Learning | YouTube Tutorial

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.

On the basis of train data ↓ Objectives:

- Accurately predict unseen test cases
- Understand which input affect the outcome, and how
- Assess the quality of our predictions and interfaces

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.

- ML is subfield of AI
- Statistical learning is a subfield of Statistics

ML has greater emphasis on large scale applications and prediction accuracy
SL emphasises models and their interpretability and precision and uncertainty.

Notation:

Input vector: $X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$ → feature, predictor.

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

→ captures errors & other discrepancies

- With a good f, we can make predictions of Y at new points $X = x$.
- Depending on the complexity of f, we understand how X_j of X affects Y.

An ideal function for f.

$$f(x) = \mathbb{E}(Y \mid X = x) \quad \rightarrow \text{regression function.}$$

The regression function:

$$f(x) = f(x_1, x_2, x_3) = \mathbb{E}[Y \mid X_1 = x_1, X_2 = x_2, X_3 = x_3]$$

This optimal predictor minimizes the MSE:

$\mathbb{E}[(Y - g(x))^2 \mid X = x]$ over all functions g at all points $X = x$.

$\epsilon = Y - f(x)$ → irreducible error

For any estimate $\hat{f}(x)$ of $f(x)$

$$\mathbb{E}[(Y - \hat{f}(x))^2 \mid X = x] = \underbrace{[f(x) - \hat{f}(x)]^2}_{\text{reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{irreducible}}$$

How to estimate f ?

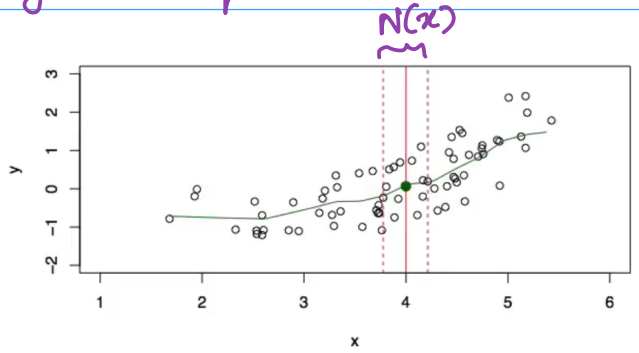
If at an $X = x$, there are only few/no points, the $\#[Y|X=x]$ cannot be computed.

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

↳ Neighborhood of x.

also called smoothing ↓ kernel & spline smoothing

Nearest neighborhood averaging can be good for $p \leq 4$, and large N



When p is large, Nearest neighbors tend to be far away.

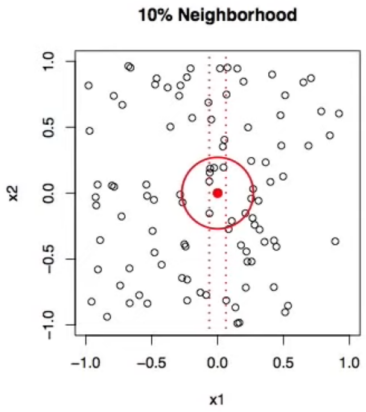
$\therefore$ The curse of dimensionality.

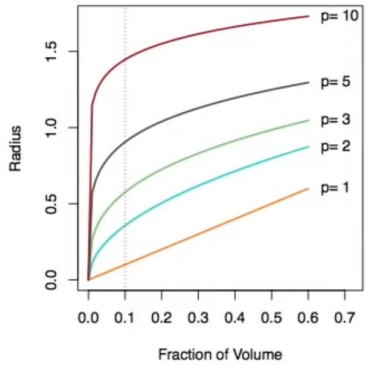
(say 10%)

→ We need to get a fraction of N values of y_i to average to bring the variance down.

→ In higher dimensions, this 10% neighborhood need not be local

The curse of dimensionality





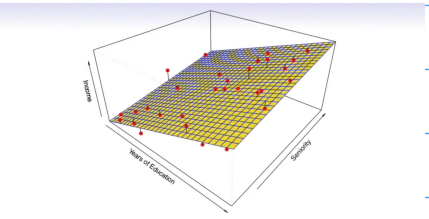
Parametric & structured models.

Linear model:

$$f_L(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Linear models are almost never correct but serves a good and interpretable approximation to the unknown true function $f(x)$

Trade offs.

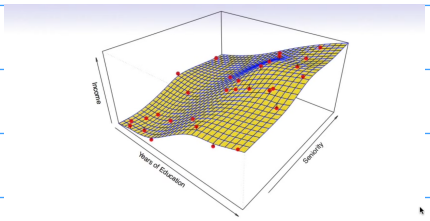


Linear regression model fit to the simulated data.

$\hat{f}_L(\texttt{education}, \texttt{seniority}) = \hat{\beta}_0 + \hat{\beta}_1 \times \texttt{education} + \hat{\beta}_2 \times \texttt{seniority}$

underfit

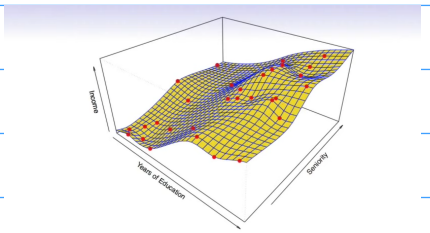
but easy to interpret



More flexible regression model $\hat{f}_S(\texttt{education}, \texttt{seniority})$ fit to the simulated data. Here we use a technique called a *thin-plate spline* to fit a flexible surface. We control the roughness of the fit (chapter 7).

good fit

difficult to interpret



Even more flexible spline regression model $\hat{f}_S(\texttt{education}, \texttt{seniority})$ fit to the simulated data. Here the fitted model makes no errors on the training data! Also known as *overfitting*.

overfit

Tradeoffs.

- Prediction accuracy versus interpretability.
  — Linear models are easy to interpret; thin-plate splines are not.
- 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.