The Elements of Statistical Learning

Ch2: Overview of Supervised Learning

Philip Lin

Data Science in Hsinchu

2015.07.18

DSHC is a non-profit studying group.

This slide is created to help us to discuss the content of "Hastie, T., Tibshirani, R. & Friedman, J., 2009. The Elements of Statistical Learning" and is not used in any profit-oriented activity.

Framework

Today, we begin with an illustrative example in prediction

Least Square v.s. Nearest Neighbor

based on this example, we talk about

- Statistical Decision Theory
- Curse of Dimensionality
- Model complexity and degree of freedom
- Concept of Bias-Variance trade-off
- Restrictions on regression model

Suppose we have 100 simulated individuals (training set τ) and their features are recorded

$$x_1(\textit{height}),\ x_2(\textit{weight}) \text{ and } y(\textit{nation}{:}\{0,1\})$$

The question is

How to predict the "nation" of next incoming individual based on his height and weight ?

The figure is cited from "Hastie, T., Tibshirani, R. & Friedman, J., 2009. The Elements of Statistical Learning, Available at: http://www.springerlink.com/index/10.1007/b94608."

Least Square

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i$$

1. Based on the **model** given above, we estimate parameter by

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - x_i^T \beta)^2$$

2. Convert fitted values $\hat{y}$ to a fitted class vairable $\hat{G}$ according to

$$\hat{G} = \begin{cases} \text{ORANGE} & \text{if } \hat{y} > 0.5 \\ \text{BLUE} & \text{if } \hat{y} \leq 0.5 \end{cases}$$

3. Construct decision boundary

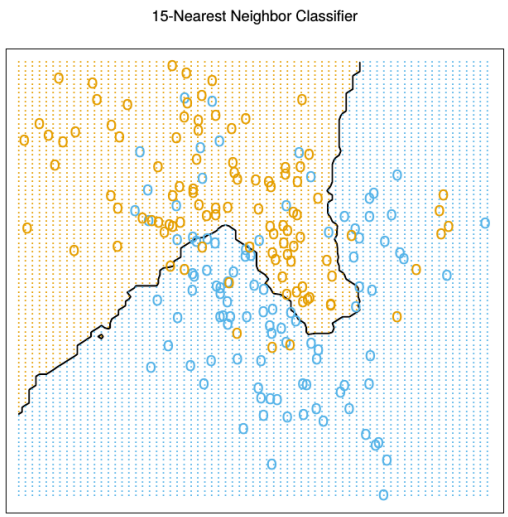
$$\{x : x^T \hat{\beta} = 0.5\}$$

Nearest Neighbor

1. Collect k-nearest neighborhood of x as set $N_k(x)$, we define the estimate

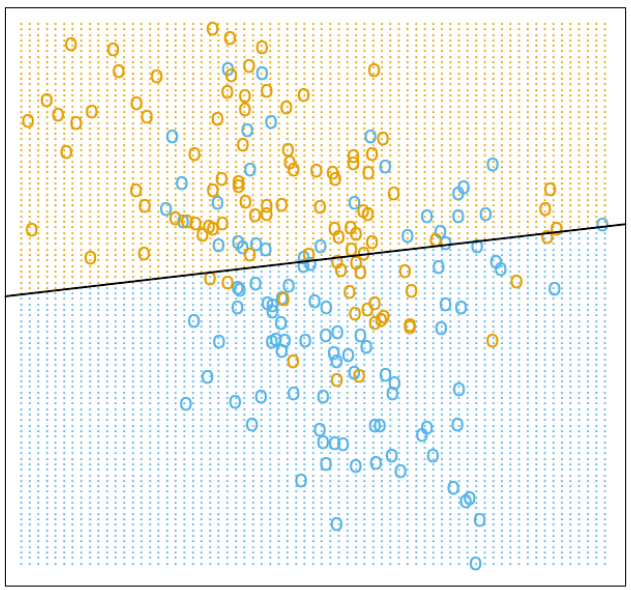
$$\hat{y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$

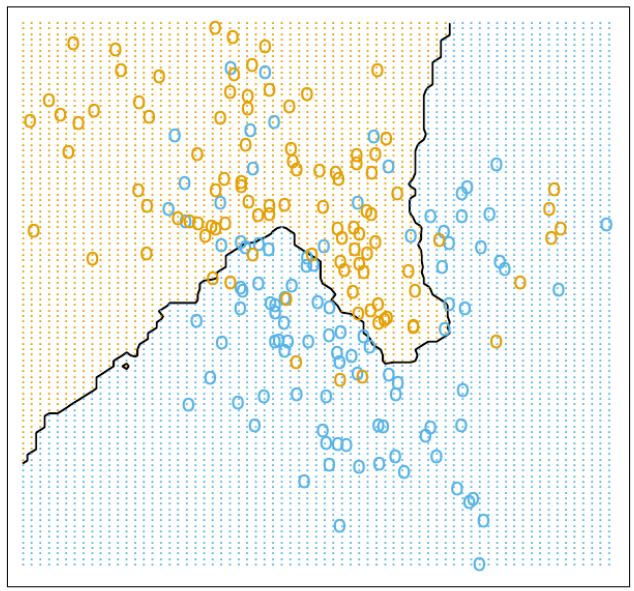
2. Since $\hat{y}$ is the proportion of ORANGE's in the neighborhood, we assign class ORANGE to $\hat{G}$ if $\hat{y} > 0.5$ amounts to a majority vote.
3. For a given k (e.g. $k = 15$), we can find the decision boundary in the same manner.



The figure is cited from "Hastie, T., Tibshirani, R. & Friedman, J., 2009. The Elements of Statistical Learning, Available at: http://www.springerlink.com/index/10.1007/b94608."

Which one do you prefer ? Let's vote !!!





The figure is cited from "Hastie, T., Tibshirani, R. & Friedman, J., 2009. The Elements of Statistical Learning, Available at: http://www.springerlink.com/index/10.1007/b94608."

Shortly compare these two procedures:

- We assume a **model** structure on LS, but no model assumption needed in $k-$NN.
- The whole decision boundary in LS uses the information of all data. In contrast, $k-$NN seems more flexiable since k controls the numbers of neighbors. Of course, we can set $k = N$. (**Global vs Local**)
- In this case $(N = 100, p = 2)$, we use 3 parameters to construct the decision boundary in LS. How many parameters should we estimate in k-NN ?

  Which one is more complicated ? How to measure the **complexity** of a method ?

Can we compared these two methods in a systematic way ? or
Is there a mathematical scheme that includes these two types of approaches ?
(**Statistical Decision Theory**)

Statistical Decision Theory

Set

- $X \in \mathbb{R}^p$ denote a real valued random input vector.

  $Y \in \mathbb{R}$ denote a real valued random output vector.

  $P(X, Y)$ denote the joint distribution of input and output vector.

- Define loss function

$$L(Y, f(X)) = (Y - f(X))^2$$

  and expected prediction error criterion

$$\text{EPE}(f) = E_{X,Y}(Y - f(X))^2$$

- This theory seeks for a good estimator of f that can minimize EPE. One common and popular way is to condition on X

$$\text{EPE}(f) = E_X E_{Y|X}\Big([Y - f(X)]^2 | X\Big)$$

  and minimize EPE pointwisely.

$$f(x) = \arg\min_c E_{Y|X}\Big([Y - c]^2 | X = x\Big)$$

- Obviously, the solution is the well-known *Regression*.

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

Suppose we have a sample of N pairs x_i, y_i drawn i.i.d. from the distribution characterized as

$$y_i = f(x_i) + \epsilon_i, \quad f \text{ is the regression function; } \textbf{additive error model}$$
$$\epsilon_i \sim (0, \sigma^2) \text{ (mean zero, variance } \sigma^2)$$

we construct an estimator for f **linear** in the y_i

$$\hat{f}(x_0) = \sum_{i=1}^{N} l_i(x_0; \mathcal{X}) y_i$$

where the weights $l_i(x_0; \mathcal{X})$ depend on the entire training sequence of x_i, denoted by $\mathcal{X}$.

Thus, we can see that both LS and $k-$NN are under the same structure (Regression) but they approximate $f(x)$ in different way.

- for LS, $f(x)$ is approximated by a **globally linear function**

$$l_i(x_0; \mathcal{X}) = \left\{ x_0^T (X^T X)^{-1} X^T \right\}_i$$

- for $k-$NN, $f(x)$ is approximated by a **locally constant function**

$$l_i(x_0; \mathcal{X}) = \frac{1}{k} I\left\{ x_i \in N_k(x_0) \right\}$$

In fact, there are some inherent assumption made in their approximations

- For LS

$$f(x) = E(Y|X = x) \approx x^T \beta$$

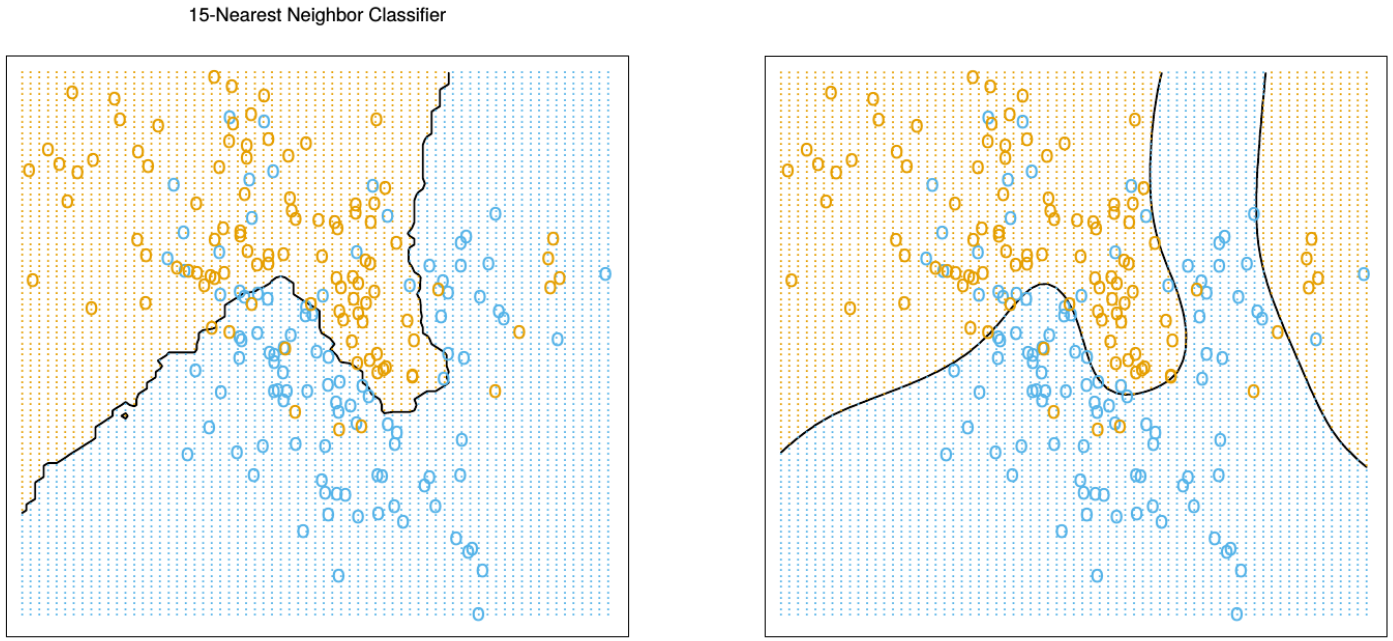
  - assume function $f(x)$ is approximately linear in β (Very STRONG !!!)

- For $k-$NN

$$f(x) = E(Y|X = x) \approx \text{Avg}(y_i|x_i \in N_k(x))$$

  - Expectation is approximated by averaging over sample data
  - conditioning at a point is relaxed to conditioning on some region 'close' to the target point. (mild assumption and very intuitive)

Recall that our nation data is generated by simulation, i.e., we know the truth.



- $k-$NN seems to be closed to the truth.
- One can show that under mild regularity conditions on $P(X, Y)$,as $N, k \to \infty$ such that $k/N \to 0$, we have

$$\hat{f}(x) \to E(Y|X = x)$$

- For large training sample size N, the points in the neighborhood are likely to be close to x, and as k gets large the average will get more stable.

The figure is cited from "Hastie, T., Tibshirani, R. & Friedman, J., 2009. The Elements of Statistical Learning, Available at: http://www.springerlink.com/index/10.1007/b94608."

"It seems we have a universal approximator, delete you LS progrom right now, it's useless."

"....."

Note:

- Nearest-neighbor works only if we are able to find a fairly large neighborhood of observations close to any x and average them.
- Regression focuses on *conditional mean*, and is usually called *frequentist approach*. It is not the only way to analyze the statistical problem, other approaches such as *Bayesian*, . . . , *etc*.

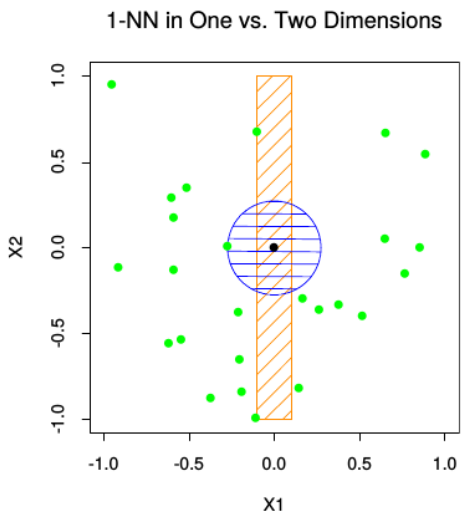
Curse of Dimensionality

Example 1. in high dimensions, all sample points are close to an edge of the sample

Consider N data points uniformly distributed in a $p-$dimensional unit ball centered at the origin. Suppose we consider a nearest-neighbor estimate at the origin. The median distance from the origin to the closest data point is given by

$$d(p, N) = \left(1 - \frac{1}{2}^{1/N}\right)^{1/p}$$

for $N = 500, p = 10, d(p, N) \approx 0.52$

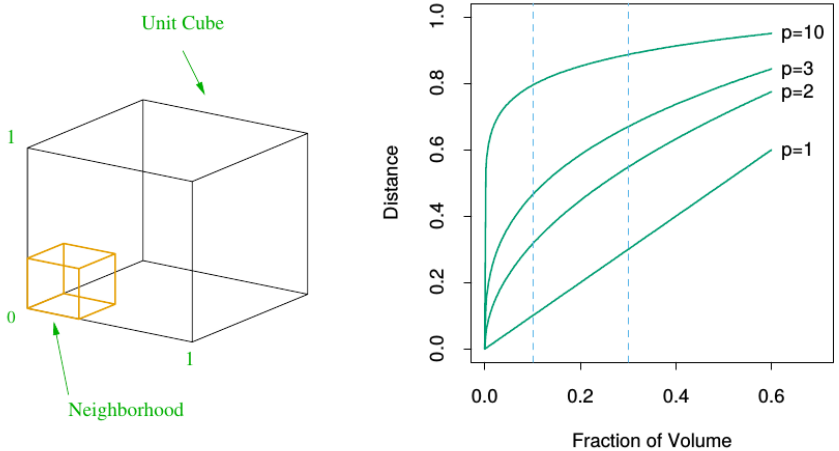


The figure is cited from "Hastie, T., Tibshirani, R. & Friedman, J., 2009. The Elements of Statistical Learning, Available at: http://www.springerlink.com/index/10.1007/b94608."

Example 2. Neighborhoods are no longer local.

Suppose send out a hypercubical neighborhood about a target point to capture a fraction r of the observations. Since this corresponds to a fraction r of the unit volume, the expected edge length will be

$$e_p(r) = r^{1/p}$$



$$e_2(0.5) = 0.707, e_{10}(0.01) = 0.63, e_{10}(0.1) = 0.80$$

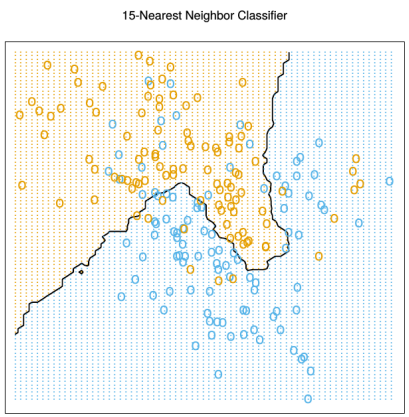
Example 3. The sampling density is proportional to $N^{1/p}$

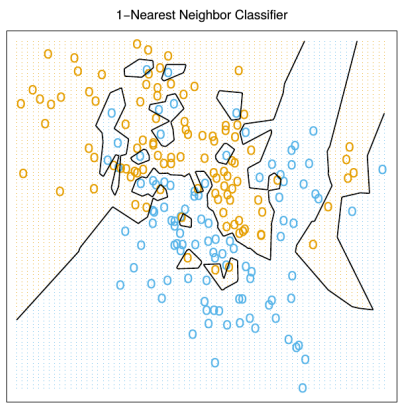
$$N_1 = 10, N_2 = 10^2, N_{10} = 10^{10}$$

The figure is cited from "Hastie, T., Tibshirani, R. & Friedman, J., 2009. The Elements of Statistical Learning, Available at: http://www.springerlink.com/index/10.1007/b94608."

Model complexity

which boundary seems more complicated ?





- In LS, we used 3 parameters (intercept, and slope of height and weight) to describe the boundary and only 1 (number of neighborhood) in $k-$NN. Is it reasonable to use number of parameters we used to describe the model complexity ?

The figure is cited from "Hastie, T., Tibshirani, R. & Friedman, J., 2009. The Elements of Statistical Learning, Available at: http://www.springerlink.com/index/10.1007/b94608."

One way to measure the complexity of a model is to measure its *prediction error.* [9]

- Assume we have training data $\boldsymbol{y}$ and testing data $\boldsymbol{y}^0$, they are independent and identically followed from the same distribution $(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$, so that we have

$$E_0\Big[\sum(y_i^0 - \mu_i)^2\Big] = E\Big[\sum(y_i - \mu_i)^2\Big]$$

- We can build the following relationship

$$\begin{aligned}
\sum(y_i - \hat{\mu}_i)^2 &= \sum(y_i - \mu_i + \mu_i - \hat{\mu}_i)^2 \\
&= \sum(y_i - \mu_i)^2 + \sum(\mu_i - \hat{\mu}_i)^2 + 2\sum(y_i - \mu_i)(\mu_i - \hat{\mu}_i) \\
\sum(y_i - \mu_i)^2 + \sum(\mu_i - \hat{\mu}_i)^2 &= \sum(y_i - \hat{\mu}_i)^2 + 2\sum(y_i - \mu_i)(\hat{\mu}_i - \mu_i) \\
E\Big\{E_0\Big[\sum(y_i^0 - \hat{\mu}_i)^2\Big]\Big\} &= E\Big[\sum(y_i - \hat{\mu}_i)^2\Big] + 2\sum \mathrm{Cov}(\hat{\mu}_i, y_i)
\end{aligned}$$

here, we define **(effective) degree of freedom** as

$$\mathrm{df} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathrm{Cov}(\hat{\mu}_i, y_i)$$

What's the meaning of df in this way ? (Note that df is unrelated to testing set)

[9] Bradley Efron (2004) The Estimation of Prediction Error, JASA.

$$y_i = \mu(x_i) + \epsilon_i \quad , \epsilon_i(0, \sigma^2)$$

and

$$\text{df}(\hat{y}) = \frac{1}{\sigma^2} \sum_{i=1}^{N} \text{Cov}(\hat{\mu}_i, y_i)$$

- If $\text{Cov}(\hat{\mu}_i, y_i) = \text{Cov}(y_i, y_i) = \sigma^2$, then $\text{df} = N$.

  In this case, $\hat{\mu}$ uses N effective degree of freedom

  If $\text{Cov}(\hat{\mu}_i, y_i) = 0$, then $\text{df} = 0$.
- However, $\text{Cov}(\hat{\mu}_i, y_i)$ is not an observable statistic unless we have distribution of $\hat{\mu}_i$.
- Stein (1981), **assume** $y \sim N(\mu, \sigma I)$, then

$$\text{Cov}_i = \sigma^2 E\Big\{ \frac{\partial \hat{\mu}_i}{\partial y_i} \Big\}$$

  where $E\Big\{ \frac{\partial \hat{\mu}_i}{\partial y_i} \Big\}$ can be unbiased estimated by $\Big\{ \frac{\partial \hat{\mu}_i}{\partial y_i} \Big\}$

Least Square

$$\begin{aligned}\hat{\mathrm{Cov}}_i =& \sigma^2 \Big\{ \frac{\partial \hat{\mu}_i}{\partial y_i} \Big\} \\ =& \Big\{ \frac{\partial}{\partial y_i} \left[X(X^T X)^{-1} X^T y \right]_i \Big\} \\ =& \sigma^2 H_{ii}, \quad \text{since } \hat{y} = Hy \text{ in linear predictor} \\ \mathrm{df} =& \frac{1}{\sigma^2} \sum \mathrm{Cov}(\hat{\mu}_i, y_i) = \mathrm{tr}(H) = \mathrm{tr}(X(X^T X)^{-1} X^T) = p\end{aligned}$$

k-NN

$$\begin{aligned}\hat{f}(x_i) =& \sum_{j=1}^{N} l(x_j; x_i) y_j = \sum_{j=1}^{N} \frac{1}{k} I\Big\{ x_j \in N_k(x_i) \Big\} \cdot y_j \\ \mathrm{Cov}_i =& \sigma^2 \Big\{ \frac{\partial \hat{\mu}_i}{\partial y_i} \Big\} = \sigma^2 \frac{1}{k} \\ \mathrm{df} =& \frac{1}{\sigma^2} \sum \mathrm{Cov}(\hat{\mu}_i, y_i) = \sum_{i=1}^{N} \frac{1}{k} = \frac{N}{k}\end{aligned}$$

What did they mean in both case ?

- for LS, number of parameter we used.
- for k-NN, note that if the neighborhoods were nonoverlapping, there would be N/k and we sould fit one parameter (a mean) in each neighborhood.

Remarks:

- effective degree of freedom measures the model complexity in the view of prediction error, however, it is not the only measure in describing model complexity.
- Other measure such as minimun descriptive length, VC dimensionality, . . . , etc.
- as we can see, one most important assumption in degree of freedom is that

  We assume our data from a Normal distribution.

  This is a heavy and questionable assumption in data analysis especially when our data is not from Normal. However, it is still an open question. Researchers tries to extend the Steins formula to release this assumption.

Bias-Variance trade-off

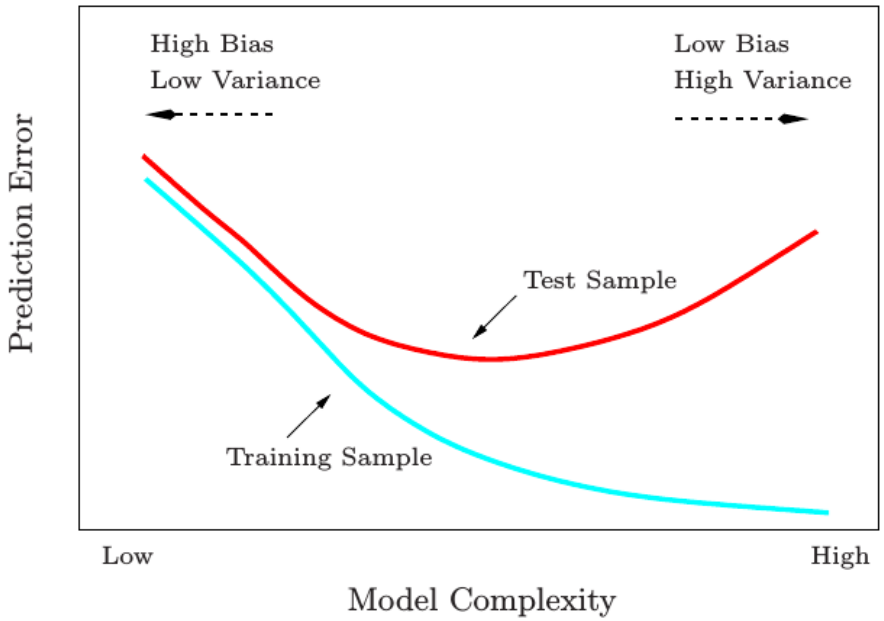
Suppose the data arise from a model

$$Y = f(X) + \epsilon \quad , \epsilon \sim (0, \sigma^2)$$

$k-$NN

$$\begin{aligned}
\text{EPE}_k(x_0) =& E_\tau E_{Y|X}\left[(Y - \hat{f}_k(x_0))^2 | X = x_0\right] \\
=& E_\tau E_{Y|X}\left\{\left[Y - f(x_0) - E_\tau \hat{f}_k(x_0) + f(x_0) - \hat{f}_k(x_0) + E_\tau \hat{f}_k(x_0)\right]^2 | X = x_0\right\} \\
=& E_{Y|X}\left[\left(Y - f(x_0)\right)^2 | X = x_0\right] + \left[E_\tau \hat{f}_k(x_0) - f(x_0)\right]^2 \\
& + E_\tau\left[\hat{f}_k(x_0) - E_\tau \hat{f}_k(x_0)\right]^2 \\
=& \sigma^2 + \text{Bias}^2(\hat{f}_k(x_0)) + \text{Var}_\tau(\hat{f}_k(x_0)) \\
=& \sigma^2 + \left[f(x_0) - \frac{1}{k}\sum_{l=1}^{k} f(x_{(l)})\right]^2 + \frac{\sigma^2}{k}
\end{aligned}$$

- The first term is *irreducible error*
- The second term will most likely increase with k, e.g. $k = 1$ v.s. $k = N$.
- The thrid term is simply the sampling error, decerase with k.



The figure is cited from "Hastie, T., Tibshirani, R. & Friedman, J., 2009. The Elements of Statistical Learning, Available at: http://www.springerlink.com/index/10.1007/b94608."

Suppose we know the relationship between Y and X is linear,

$$Y = X^T\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

$$\begin{aligned}\text{EPE}(x_0) =& \sigma^2 + \text{Bias}^2(\hat{f}(x_0)) + \text{Var}_\tau(\hat{f}(x_0)) \\ =& \sigma^2 + 0^2 + E_\tau x_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}x_0\sigma^2\end{aligned}$$

suppose further N is large and τ were select at random, and assuming $E(X) = 0$, then $\boldsymbol{X}^T\boldsymbol{X} \to N\text{Cov}(X)$

$$\begin{aligned}E_{x_0}\text{EPE}(x_0) \sim& E_{x_0}x_0^T\text{Cov}(X)^{-1}x_0\sigma^2/N + \sigma^2 \\ =& \text{trace}[\text{Cov}(X)^{-1}\text{Cov}(x_0)]\sigma^2/N + \sigma^2 \\ =& \sigma^2(p/N) + \sigma^2\end{aligned}$$

- expected EPE increases linearly as a function of p. If N is large and/or σ^2 is small, this growth in variance is negligible.
- By imposing some heavy restrictions on the class of models being fitted, we have avoid the curse of dimensionality.
- *"All models are wrong, but some are useful"*, and linear model is such a case (in my experience).

Restrictions on regression model

Consider the RSS criterion for an arbitrary function f

$$\text{RSS}(f) = \sum_{i=1}^{N} \big(y_i - f(x_i)\big)^2$$

Remarks

- Minimizing RSS leads to infinitely many solutions: any function $\hat{f}$ passing through the training points (x_i, y_i) is a solution.
- In order to obtain useful results for finite N, we must restrict the eligible solutions to a smaller set of functions.
- How to choose the constraint ? (It's an art.)

Example 1.: penalized RSS

$$\sum_{i=1}^{N} \left(y_i - f(x_i)\right)^2 + \lambda \int [f''(x)]^2 \mathrm{d}x$$

- This constraint restrict rigid change of slope of $f(x)$.

Example 2.: for k-NN

$$f(x) = E(Y|X = x) \approx \mathrm{Avg}(y_i|x_i \in N_k(x))$$

- The strength of the constraint is dicated by the neighborhood size.
- NN method are based on the assumption that locally the function is constant.
- for example, local constant fit in infinitesimally small neighborhoods is no constrant at all.

Example 3.: likelihood inference on linear model

$$\sum_{i=1}^{N} \left(y_i - f(x_i)\right)^2 = \sum_{i=1}^{N} \left(y_i - x_i^T \beta\right)^2$$

- This constraint restrict structure of $f(x)$ to be linear and normal error assumption

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i, \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

- Likelihood function $L(\beta|x)$: given the data we observed, we find the most possible β in parameter space.

$$\begin{aligned} L(\beta|x) &= \prod_{i=1}^{N} L_i(\beta|x_i) \\ &= \prod_{i=1}^{N} \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ - \frac{(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2}{2\sigma^2} \right\} \right) \end{aligned}$$

and we persue $\hat{\beta}_{\text{MLE}} = \arg\max_{\beta} L(\beta|x)$ and name it *Maximum Likelihood Estimator*

Remarks

- Based on the model structure

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i, \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

  there are many approach to estimate β (e.g., moment estimator, UMVUE, bayesian estimator, . . . , etc). Why we talk about MLE only ?

- The speaker (Philip Lin) learned MLE only.
- its implementation is relatively easy and attains very elegant asymptotic performance.
  1. when $N \to \infty$, $\hat{\beta}_{\text{MLE}} \to \beta_0$
  2. when $N \to \infty$, $\hat{\beta}_{\text{MLE}}$ has minimum variance among the class of all unbiased estimator. (Cramar-Rao lower bound)

  In the view of efficiency (a measure on vairance of an unbiased estimator), we call MLE **most efficient**.

Feedback

- In k-NN, if we take $k = \frac{N}{p}$, then df $= p$. Could this degree of freedom comparable to the degree of freedom in LS case ? Obviously not, Why ?
- Curse of dimensionality breaks down k-NN becuase we catch neighborhood by means of Euclidean distance, k-NN is still applicable if we use other distance.

The End