Supervised Learning

Feng Li

feng.li@cufe.edu.cn

School of Statistics and Mathematics
Central University of Finance and Economics

Revision: October 29, 2013

Today we are going to talk about...

1. Prediction with least squared and nearest neighbor methods

Least squared

- $\hat{Y} = X'\hat{\beta}$
- $\mathrm{RSS} = (y - X\beta)'(y - X\beta)$
- $\hat{\beta} = (X'X)^{-1}X'y$
- the hat matrix $H = X(X'X)^{-1}X'$, symmetric and idempotent $(H^2 = H)$
- $\mathrm{tr}(H) = p$ is the number *effective parameters*.

Nearest Neighbor Methods

- $\hat{Y}(x) = \frac{1}{k}\sum_{x_i \in N_k(x)} y_i$
- $N_k(x)$ is the neighborhood of x defined by the k closest points x_i in the training sample.
- The Euclidean distance is usually used for measuring the distance between two points.

$$d(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

- Nearest neighbor methods only depend on one parameter k.
- However the effective parameters is N/k which is greater than the effective parameter p in linear model.
- Don't use sum-of-squared errors to pick up k. This will end up with $k = 1$ in the end.

Statistical Decision Theory

- Loss function $L(Y, f(X))$ to penalizing errors in predictions.
- $L(Y, f(X)) = (Y - f(X))^2$ is commonly used (squared error loss).
- The expected prediction error (using squared error loss function)

$$\mathrm{EPE} = E(Y - f(X))^2$$

- **Find** $f(x)$ **so that** arg $\min_c E_{Y|X}((Y - f(x))^2 | X = x)$
- An example nearest with neighbor: k-nearest neighbor assume $f(x)$ is well approximated by a locally constant function

$$\hat{f}(x) = Ave(y_i | x_i \in N_k(x))$$

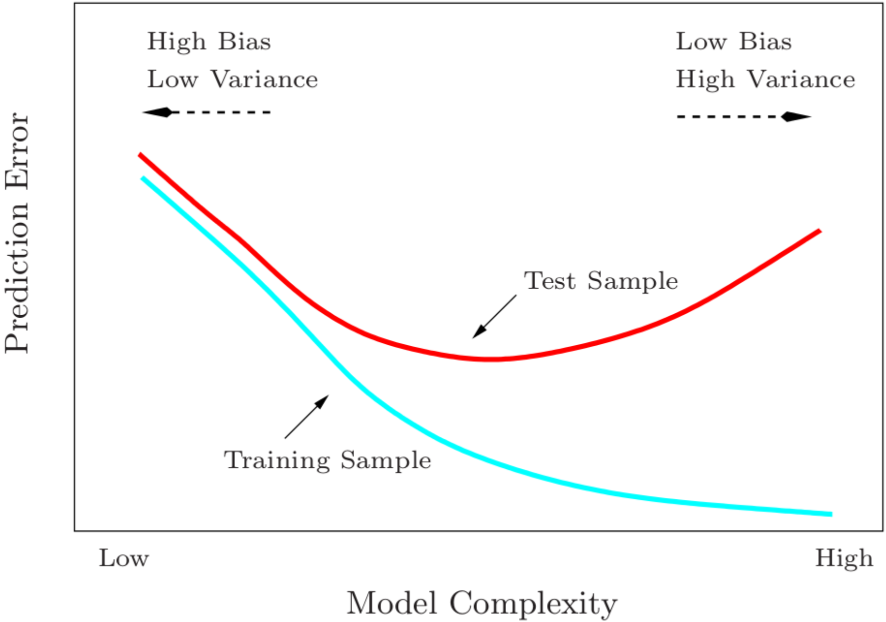
- An example with least squared: assuming $f(x)$ is well approximated by a global linear function

$$\hat{f}(x) = X\hat{\beta}$$

The bias variance trade off

- **Training sample:** The dataset used for building and estimating the model.
- **Testing sample**: The dataset used for prediction
- If the testing sample is nested in the training sample, we called it **in-sample fitting**. If the testing sample is not part of the training sample, it is called **out-of-sample fitting**.
- **The prediction error** at x_0 can be decomposed as three parts

$$\begin{aligned} EPE(x_0) &= E((Y - \hat{f}(x_0))^2 | X = x_0) \\ & \sigma^2 + Bias^2(\hat{f}(x_0)) + Var(\hat{f}(x_0)) \end{aligned}$$


High Bias
Low Variance
Low Bias
High Variance
Test Sample
Training Sample
Prediction Error
Low
High
Model Complexity

Supervised learning

- Supervised learning is also known as learning by example (with a teacher)
- A learning algorithm (OLS, nearest neighborhood,...) studies the training set and produce $\hat{f}(x)$.
- The algorithm can also modify $\hat{f}(x)$ in response to $y - \hat{f}(x)$
- A clear measure of success/lack-of-success and compare the effectiveness of different methods are available in supervised learning, e.g. cross-validation, loss functions
- **NOTE**: Finding a good $\hat{f}(x)$ is very important.

Upsupervised learning

- Upsupervised learning: learning without a teacher.
- Examples: K-means clustering and other clustering methods, density estimation problems.
- Upsupervised learning can be transformed into supervised learning with the tool **generalized association rules**