SML Week 2-3

C. Andy Tsao

Dept. of Applied Math
National Dong Hwa University

September 26, 2013

Outline

Reference: §2.4–2.9, Chapter 3 of HTF's ESL

$E(Y|X)$ linear in $X_1, \cdots, X_p$

- Both KNN and LS can be viewed as $E(Y|x)$ (in some sense) which minimizes the *expected (squared) prediction error* $EPE(f) = E_{Y,X}(Y - f(X))^2 = E_X E_{Y|X}[(Y - f(X))^2|X]$
- What if loss is chosen as L_1, $L(y, f(x)) = |y - f(x)|$, instead of the L_2 loss $(y - f(x))^2$?
  1. $\hat{f}(x) = median(Y|x)$ more robust but lesser convenient
- Discrete Y? Or $\#\mathcal{Y}$ is finite?

Bayes classifier for Discrete Scenario

- WLOG, assume $\mathcal{Y} = \{1, \cdot, K\}$

$$EPE(G) = E_{Y,X} L(Y, G(X)), \quad X, Y \sim P_{Y,X}$$
$$= E_X E_{Y|X} L(Y, G(X)) = E_X \sum_{k=1}^{K} L(k, G(X)) P(Y = k|X)$$

  - Minimize pointwise $\rightsquigarrow \hat{G}(x) = arg \min_{y \in \mathcal{Y}} E_{Y|x} L(Y, G(x))$. (*Bayes classifier*)
  - When $L(y, G(x)) = 1_{[y \neq G(x)]}$, 0-1 loss, $\hat{G}(x) = arg \min_{y \in \mathcal{Y}} [1 - P(y|X = x)] = arg \max_{y \in \mathcal{Y}} P(y|X = x)$.
- Bayes classifier
  - Good: Achieve the optimal error rate (Bayes error rate).
  - Bad: the conditional $P_{Y|x}$ usually unknown and can lead to unreasonable estimator in cases.

Ways to improve KNN, LSE

Estimation of $E(Y|x)$ through KNN or regression can fail

- Curse of dimensionality: KNN includes points afar leads to large error
- If special structure is known, further reduction in bias and variance is possible.

Prediction Problem: Emphasis on "Y" rather than "X"

- Statistical Model: Assumption on $P_{Y,X}$ (or ϵ), say $Y = f(X) + \epsilon$
- Supervised learning

Functional approximation

- Functional approximation
  - regression: $f(x) = x'\beta, \beta \in R^p$
  - *linear basis expansions*: $f_\theta(x) = \sum_k h_k(x)\theta_k$
  - $\{h_k(x)\}_k$ forms a basis for the feasible/approximate space F where the target f is located/approximated.
  - Examples: $x_1^2, x_1x_2, \cos(x_3)$. Polynomials, trig functions. Also $h_k(x) = \frac{1}{1+exp(-x'\beta)}$
- Residual Sum of Squares (RSS) $RSS(\theta) = \sum_{i=1}^n (y_i - f_\theta(x_i))^2$. Projection.

$E(Y|X)$ linear in $X_1, \cdots, X_p$

- Simple: easier computation, interpretation and communication
- Readily generalizable: transformation on Y and X, combination of X's'
- Conceptual Framework for more general methods, for example, nonlinear problems.

Definition

$(Y_i, x_i)_{i=1}^n$ with $x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})'$

- $Y_i = \beta_0 + \sum_{j=1}^p x_{ij}\beta_j + \epsilon_i, i = 1, \cdots, n.$
- ϵ_i id with $E(\epsilon_i) = 0$ and $Cov(\epsilon_i, \epsilon_j) = \sigma^2$ if $i = j$; 0 otherwise.
- (Typically) $\epsilon_i \sim_{iid} N(0, \sigma^2)$.

Alternatively,

- Systematic component: $E(Y|X) = \beta_0 + \sum_{j=1}^p X_j\beta_j$
- Random component: ϵ_i id with $E(\epsilon_i) = 0$ and $Cov(\epsilon_i, \epsilon_j) = \sigma^2$ if $i = j$; 0 otherwise.

How flexible is LR?

Assume $\epsilon_i \sim_{iid} N(0, \sigma^2)$

- $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
- $Y_i = \beta_0 + \beta_1 X_{1i} X_{2i} + \epsilon_i$
- $sin(Y_i) = exp(\beta_0 + \beta_1 X_i) + \epsilon_i$

How flexible is LR?

Assume $\epsilon_i \sim_{iid} N(0, \sigma^2)$

- $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
- $Y_i = \beta_0 + \beta_1 X_{1i} X_{2i} + \epsilon_i$
- $sin(Y_i) = exp(\beta_0 + \beta_1 X_i) + \epsilon_i$

Your turn

How flexible is LR?

Assume $\epsilon_i \sim_{iid} N(0, \sigma^2)$

- $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
- $Y_i = \beta_0 + \beta_1 X_{1i} X_{2i} + \epsilon_i$
- $sin(Y_i) = exp(\beta_0 + \beta_1 X_i) + \epsilon_i$

Your turn

- quantitative inputs, X
- transformation of quantitative inputs, $sin(X), \log(X), \sqrt{X}$
- powers, $X_2 = X^2, X_3 = X^3$
- interactions: $X_3 = X_1^2 X_2$.
- For GLM (general linear model), qualitative inputs, say $1_{[X>20]}$.

Remark: Linear in parameters (β) not in X.

Estimation of LR

$Y = X\beta + \epsilon$

- Solve β st $Q(\beta) = ||Y - X\beta||^2$ is minimized
- Normal equation: β solves $X^t(Y - X\beta) = 0$.
- When X^tX is nonsingular, $\hat{\beta} = (X^tX)^{-1}X^tY$.
- Geometric Interpretation: $\hat{Y} = X(X^tX)^{-1}X^tY$ is the projection of Y onto the column space of the design matrix X.

Inference: HT and CI

- $\hat{\beta} \sim N(\beta, (X^t X)^{-1}\sigma^2)$
- $\hat{\sigma^2} = ||Y - \hat{Y}||^2/(n - p - 1)$.
- $(n - p - 1)\hat{\sigma^2} \sim \sigma^2\chi^2_{n-p-1}$.
- Gauss-Markov Theorem: For any estimable parameter $\theta = a^t\beta$, $a^t\hat{\beta}$ is BLUE provided GM condition holds.

Simple Linear Regression

- $Y_i = x_i\beta + \epsilon_i$ (No intercept)
- $Y = X\beta + \epsilon$ where $X = (x_1, \cdots, x_n)^t$
- $\hat{\beta} = (X^tX)^{-1}X^tY = \frac{\sum_1^n x_i y_i}{\sum_i x_i^2}$,
  $r_i = y_i - x_i\hat{\beta}$.
- In inner product with $< x, y >= \sum_i x_i y_i$
  $\hat{\beta} = \frac{<x,y>}{<x,x>}, \quad r = Y - X\hat{\beta}$.

Multiple Linear Regression w/ orthogonal x's

- $Y = X\beta + \epsilon$ where $X = (X_1, \cdots, X_n)^t$
- $\hat{\beta} = (X^t X)^{-1} X^t Y = \frac{\sum_1^n x_i y_i}{\sum_i x_i^2}$,
  $r_i = y_i - x_i \hat{\beta}$.
- $\hat{\beta}_j = \frac{<X_j, y>}{<X_j, X_j>}, \quad r = y - X\hat{\beta}$.
- When inputs are orthogonal, they have no effect on each other parameter estimates in the model.

Succession Orthogonalization w/ general x's

Orthogonality occurs in balanced, designed experiment but not in general

- Initialize $z_0 = x_0 = 1$
- For $j = 1, 2, \cdots, p$
  Regress x_j on $z_0, z_1, \cdots, z_{j-1}$ to get
  $\hat{\gamma}_{lj} = \frac{<z_l, x_j>}{<z_l, z_l>}, l = 0, 1, \cdots, j-1$
  $z_j = x_j - \sum_{k=0}^{j-1} \hat{\gamma}_{lj} z_k$.
- Regress y on the residual z_p to get $\hat{\beta}_p$.

Gram-Schmidt procedure for multiple regression

- z's are orthogonal to each other.
- Iterative projection of Y onto z's
- $\hat{\beta} = (\hat{\beta}_0, \cdots, \hat{\beta}_p)'$ is **a** LSE.

Succession Orthogonalization: Recap

- $\hat{\beta}_j$ represents the additional contribution of X_j on Y, after X_j has been adjusted by $X_0, X_1, \cdots, X_{j-1}$.
- $\hat{Y} = X\hat{\beta}$ is **the** projection of Y onto column space of X
- Non-unique $\hat{\beta}$. Unique $\hat{Y}$
- Alternative iteration for β: Iterative residual fitting.

Exercise 1: Write down the algorithm for iterative residual fitting and show that the obtained $\hat{\beta}$ also solves the normal equation.

Unsatisfying LSE

$Y|X_1, \cdots, X_q$, q large/huge

- Accuracy
  Even if $\hat{\beta} = (X^t X)^{-1} X^t Y$ exists, it may have large variance.
- Interpretation
  Non-uniqueness
- Scientific Important X might be missing
- Variable selection

Subset Selection

$Y|X_1, \cdots, X_q$, q large/huge. Want to pick $p(<< q)$ X's out of them.

- What have we learned before?

Subset Selection

$Y|X_1, \cdots, X_q$, q large/huge. Want to pick $p(<< q)$ X's out of them.

- What have we learned before?
- Accuracy versus parsimoniousness

Subset Selection

$Y|X_1, \cdots, X_q$, q large/huge. Want to pick $p(<< q)$ X's out of them.

- What have we learned before?
- Accuracy versus parsimoniousness
- Mission impossible: High accuracy, few indep variables
- Criterion-based approach: R^2_{adj}, AIC, etc
- Important First
- Simple versus Complex terms
- Use auto procedure only when necessary. Screening rather than determing.

Shrinking Methods

$$\hat{\beta^{ridge}} = argmin_{\beta} \left\{ (Y - X\beta)^t (Y - X\beta) + \lambda \beta^t \beta \right\}$$

- What does this mean? Alternatives?
- (Ex 2) It can be shown

$$\hat{\beta^{ridge}} = (X^t X + \lambda I)^{-1} X^t Y.$$