SML Week 1

C. Andy Tsao

Institute of Statistics
National Dong Hwa University

September 15, 2013

Outline

Motivating problems

- What is your income based on the items you bought? [Regression]
- Prostate Cancer [Regression] (`http://www.cancer.gov/cancertopics/factsheet/detection/PSA`)
- Animal Recognition (Is it a dog?) [Classification]
- Email Spam
- Hand-written Digit Recognition

cf. Figure 1.

Formulation of SML problem

Let $(y_i, x_i)_{i=1}^n \sim_{iid} P_{Y,X}$ with $y's \in \mathcal{Y}$, $x's \in \mathcal{X}$.
Objective: Find $F \in \mathcal{F}$ such that
$E_{Y,X} L(Y, F(X))$ is minimized.

- For classification problem, $\#\mathcal{Y} = K < \infty$. And $\mathcal{Y} = \mathcal{R} = (-\infty, \infty)$ for general prediction problem.
- Examples: Hand-digit recognition, spam-detection, diagnosis, precipitation prediction, etc.
- Learning (by examples [mainly training data]) vs. Rule-based classification
- supervised, semi-supervised, unsupervised.

Statistical Decision Theory: Versions of Expected Losses

[Point Estimation Problem]
Let $X_1, \cdots, X_n \sim f_\theta$, for example, pdf of $N(\theta, \sigma^2)$ or pmf of $Bernoulli(\theta)$
Objective: Find $\hat{\theta}_*$

Statistical Decision Theory: Versions of Expected Losses

[Point Estimation Problem]
Let $X_1, \cdots, X_n \sim f_\theta$, for example, pdf of $N(\theta, \sigma^2)$ or pmf of $Bernoulli(\theta)$
Objective: Find $\hat{\theta_*}$ which minimizes

- Risk: $R(\theta, \hat{\theta}) = E_{X|\theta} L(\theta, \hat{\theta}(X))$
- Bayes expected loss: $E_{\theta|x} L(\theta, \hat{\theta}(x))$ wrt π
- Bayes Risk: $r(\pi, \hat{\theta}) = E_{X,\theta} L(\theta, \hat{\theta}(X))$ wrt π

among all $\hat{\theta} \in \mathcal{D}$, collection of all estimators

Which decision δ is better?

With respect to risks

- $\delta_1 >_R \delta_2$ iff
  $R(\theta, \delta_1) \leq R(\theta, \delta_2)$ for all θ and inequality holds for some $\theta \in \Theta$
- δ is inadmissible in $\mathcal{D}$ iff
  there exists δ_* which is R-better than δ.
- δ is admissible in $\mathcal{D}$ iff
  it is not inadmissible in $\mathcal{D}$

Bayes Procedure

We say δ_π is a Bayes procedure wrt π iff

$$\delta_\pi = arg \min_{\delta \in \mathcal{D}} r(\pi, \delta).$$

Complete Class Theorem

- A class $\mathcal{C}$ is *complete* if for any decision δ not in $\mathcal{C}$, there exists a decision $\delta^{'}$ which dominates δ.
- Under some regularity conditions, the class of Generalized Bayes procedures form a complete class.
- Implication: Search no further. Work with Generalized Bayes procedures.

$E(Y|x)$

Consider $X \in R^p$, $Y \in R$ with joint prob distribution $P_{Y,X}$. Seek a ftn f for predicting Y given X. Under $L(Y, f(X)) = (Y - f(X))^2$, *squared error loss*, in the spirit of Bayes risk, find f minimize the *Expected Predicted Error* (EPE(f)) among all possible functions

$$EPE(f) = E(Y - f(X))^2 = \int (y - f(x))^2 dP_{Y,X}(y, x) \quad (1)$$
$$= E_X E_{Y|X}\left([Y - f(X)]^2 | X\right).$$

Conditioning on $X = x$, $f(x)$ is a *constant*. Pointwise minimization

$$f_\pi(x) = argmin_c E_{Y|x}\left([Y - c]^2 | x\right).$$

Minimizer $f_\pi(x) = E(Y|x)$, best prediction of Y given x. That is

$$EPE(f_\pi) \leq EPE(f), \text{for all } f \in \mathcal{F}$$

KNN as $E(Y|x)$

Let $T = (Y_i, X_i)_{i=1}^n$ be the training data.

- $\hat{f}(x) = Ave(y_i|x_i \in N_k(x))$, where "Ave" =average, $N_k(x)$ is the neighborhood containing the k points in T closest to x.
- expectation is approximated by averaging over sample space.
- conditioning at a point is relaxed to conditioning on some region "close" to the target point x.

How good is KNN? Rationale? Search is over?

KNN as $E(Y|x)$

Let $T = (Y_i, X_i)_{i=1}^n$ be the training data.

- $\hat{f}(x) = Ave(y_i|x_i \in N_k(x))$, where "Ave" =average, $N_k(x)$ is the neighborhood containing the k points in T closest to x.
- expectation is approximated by averaging over sample space.
- conditioning at a point is relaxed to conditioning on some region "close" to the target point x.

How good is KNN? Rationale? Search is over?

- Sample size usually small
- As p increases, $N_k(x)$ becomes huge
- Convergence
  - Converge holds. $\hat{f} \to f$ as $n \to \infty$
  - Slower rate of convergence.

LS as $E(Y|x)$

- $f(x) \approx x^T\beta$
- Plug this f into (1), β can be solved $\beta = [E(XX^T)]^{-1}E(XY)$.

KNN vs. LS

- LS assumes $f(x) \approx$ globally by a linear function
- KNN assumes $f(x) \approx$ locally by a const function