Topics in Variable Selection I

Statistical decision theory and covariance penalties

Elvis Cui

PhD in Biostatistics, UCLA

November 28, 2019

Overview

Model set-up

High dimensional model:

- training set
$$d = \{(x_i, y_i), i = 1, \cdots, N\}$$
- Training set is fixed at their observed values.
- An unknown vector μ of expectations $\mu_i = \mathbb{E}(y_i)$ has yielded the observed vector of reponses y according to some given probability model.
- Assume $\mathbb{E}(y) = \mu$ and $Cov(y) = \sigma^2 I$, we denote this as
$$y \sim (\mu, \sigma^2 I)$$
- This is a model with dimension N (why?).

Statistical decision theory

Statistical desicison theory: suppose $y \sim (\mu, \sigma^2 I)$ (may not be normal), a regression rule $r(\cdot)$ has been used to produce an estimate of vector μ,

$$\hat{\mu} = r(y)$$

For instance, $\hat{\mu} = r(y) = X(X^T X)^{-1} X^T y$ is the LSE in the linear regression model $\mu = X\beta$. Suppose we want to know how accurate $\hat{\mu}$ will be in predicting a new vector of observations y_0 from the model

$$y_0 \sim (\mu, \sigma^2 I), \text{ indenpendent of } y$$

Note that $y_0 = (y_{01}, \cdots, y_{0p})^T \in \mathbb{R}^p$ and so is y.

Statistical decision theory

Define prediction error in terms of **squared discrepancy**

$$Err_i = \mathbb{E}_0\{(y_{0i} - \hat{\mu}_i)^2\}$$

for each component i, where $\mathbb{E}_0$ indicates expectation w.r.t. y_{0i} random but $\hat{\mu}_i$ fixed. Then the overall prediction error is the average

$$Err. = \frac{1}{N}\sum_{i=1}^{N} Err_i$$

An apparent error for each component i: $err_i = (y_i - \hat{\mu}_i)^2$.

Covariance penalty

Lemma 1

Let $\mathbb{E}$ indicate expectation over both y and y_0. Then

$$\mathbb{E}(Err_i) = \mathbb{E}(err_i) + 2Cov(\hat{\mu}_i, y_i)$$

where the last term is the covariance between the i^{th} components of $\hat{\mu}$ and y,

$$Cov(\hat{\mu}_i, y_i) = \mathbb{E}\{(\hat{\mu}_i - \mu_i)(y_i - \mu_i)\}$$

Note: *$\mathbb{E}(\hat{\mu}_i)$ may not equal μ_i.*

Covariance penalty

- The lemma says, the apparent error err_i underestimates the tru prediction error Err_i by the **covariance penalty** $2Cov(\hat{\mu}_i, y_i)$.
- Covariance penalty estimates of prediction error take the form

$$\widehat{Err}_i = err_i + 2\widehat{Cov}(\hat{\mu}_i, y_i)$$

  where $\widehat{Cov}(\hat{\mu}_i, y_i)$ approximates $Cov(\hat{\mu}_i, y_i)$.
- Overall prediction error is estimated by

$$\widehat{Err}_{\cdot} = \frac{\sum err_i}{N} + \frac{2}{N}\sum_{i=1}^{N} \widehat{Cov}(\hat{\mu}_i, y_i)$$

- The form of $\widehat{Cov}(\hat{\mu}_i, y_i)$ depends on the context assumed for the prediction problem.

Mallows' C_p estimate for linear estimators

Suppose we estimate μ via linear combinations of y:

$$\hat{\mu} = c + My$$

where $c \in \mathbb{R}^N$ and $M \in \mathbb{R}^{N \times N}$ are known. Then the covariance matri between $\hat{\mu}$ and y is

$$Cov(\hat{\mu}, y) = \sigma^2 M$$

Denote M_{ii} as the i^{th} diagonal element of M,

$$\widehat{Err}_i = err_i + 2\hat{\sigma}^2 M_{ii}$$

where $err = \sum_i (y_i - \hat{\mu}_i)^2 / N$. Thus,

$$\widehat{Err}_. = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{\mu}_i)^2 + \frac{2\hat{\sigma}^2}{N} Tr(M)$$

Mallows' C_p estimate

Definition 2 (Mallows' C_p estimate)

Given a linear estimator $\hat{\mu} = r(y) = c + My$, one estimation of prediction error called **Mallows' C_p estimate** is the following:

$$\widehat{Err}_{\cdot} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\mu}_i)^2 + \frac{2\hat{\sigma}^2}{N}Tr(M)$$

For OLS, $M = X(X^X)^{-1}X$ has $Tr(M) = p$, the number of predictors, so

$$\widehat{Err}_{\cdot} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\mu}_i)^2 + \frac{2\hat{\sigma}^2}{N}p$$

Degrees of freedom

For OLS regression, the degrees of freedom p (the rank of matrix X) determines the covariance penalty $(2/N)\sigma^2 p$. Comparing this with $\widehat{Err}_{\cdot}$ leads to **a general definition of degrees of freedom df** for a regression rule $\hat{\mu} = r(y)$:

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

Covariance penalty under Gaussian assumption

- If we are willing to assume normality:

$$y \sim \mathcal{N}_p(\mu, \sigma^2 I)$$

  we can **drop the assumption of linear estimators**.
- That is, for any differentiable estimator $\hat{\mu} = r(y)$, the covariance term is given by

$$Cov(\hat{\mu}_i, y_i) = \sigma^2 \mathbb{E}\{\partial \hat{\mu}_i / \partial y_i\}$$

- Use this formula to estimate Err_i and $Err_.$.

Technical proof

The formula $Cov(\hat{\mu}_i, y_i) = \sigma^2 \mathbb{E}\{\partial \hat{\mu}_i / \partial y_i\}$ can be obtained from integration by parts (let's do 1-D case):

$$\begin{aligned} Cov(\hat{\mu}, y) &= \int_{\mathbb{R}} \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu)^2/(2\sigma^2)} (y-\mu) \right] \hat{\mu}(y) dy \\ &= \sigma^2 \int_{\mathbb{R}} \phi(\frac{y-\mu}{\sigma}) \frac{\partial \hat{\mu}(y)}{\partial y} dy \\ &= \sigma^2 \mathbb{E}(\frac{\partial \hat{\mu}(y)}{\partial y}) \end{aligned}$$

Where $\phi(\cdot)$ denotes the pdf of standard normal.

Stein's Unbiased Risk Estimation (SURE)

Definition 3 (Stein's Unbiased Risk Estimation (SURE))

The SURE formula for multivariate model is

$$\widehat{Err}_i = err_i + 2\sigma^2 \frac{\partial \hat{\mu}_i}{\partial y_i}$$

with corresponding estimate for overall prediction error

$$\widehat{Err}_{\cdot} = err + \frac{2\sigma^2}{N} \sum_{i=1}^{N} \frac{\partial \hat{\mu}_i}{\partial y_i}$$

Deviance of likelihood as discrepency

- Covariance penalties can apply to measures of prediction error other than squared error $D(y_i, \hat{\mu}_i) = (y_i - \hat{\mu}_i)^2$.
- Suppose y_i are obtained from different members of a one-parameter exponential family $f_\mu(y) = \exp\{\lambda y - \gamma(\lambda)\} f_0(y)$,

$$y_i \sim f_{\mu_i}(y_i) \; \forall \; i = 1, \cdots, N$$

- Error is measured by the deviance

$$D(y, \hat{\mu}) = 2 \int_{\mathcal{Y}} f_y(Y) \log \left(\frac{f_y(Y)}{f_{\hat{\mu}}(Y)} \right) dY$$

Akaike Information Criterion: covariance penalty

- The apparent error $\sum_i D(y_i, \hat{\mu}_i)$ is

$$err = \frac{2}{N} \sum_{i=1}^{N} \log \left(\frac{f_{y_i}(y_i)}{f_{\hat{\mu}_i}(y_i)} \right) = \frac{2}{N} \left\{ \log(f_y(y)) - \log(f_{\hat{\mu}}(y)) \right\}$$

- **The general theory** gives overall covariance penalty:

$$\text{penalty} = \frac{2}{N} \sum_{i=1}^{N} Cov(\hat{\lambda}_i, \hat{\mu}_i)$$

where $\hat{\lambda}_i$ is the natural parameter corresponding to $\hat{\mu}_\lambda$ (e.g., for binomial case, $\hat{\lambda}_i = \log \frac{\hat{\mu}_i}{1-\hat{\mu}_i}$).

Akaike Information Criterion

- If $\hat{\mu}$ is obtained as the MLE of μ in a generalized linear model with p degrees of freedom, then

$$\text{penalty} \doteq \frac{2p}{N}$$

  is a good approximation.
- The corresponding version of $\widehat{Err}$ can then be written as

$$\widehat{Err} \doteq -\frac{2}{N}\{\log(f_{\hat{\mu}}(y)) - p\} + \text{constant}$$

  where the constant $\frac{2}{N}\log(f_y(y))$ does not depend on $\hat{\mu}$.
- The term $\{\log(f_{\hat{\mu}}(y)) - p\}$ is called the **Akaike information criterion** (AIC).

AIC: Sidenotes

- The AIC says to select the rule maximizing the **penalized** maximum likelihood:

$$\log(f_{\hat{\mu}^{(j)}}(y)) - p^{(j)}$$

  where $\hat{\mu}^{(j)}$ is rule j's MLE and $p^{(j)}$ its degrees of freedom.
- For GLMs, the AIC amounts to selecting the rule with the smallest $\widehat{Err}_{\cdot}$.
- An analogue to AIC is called **Bayesian information criterion** (BIC): for model $\mathcal{M}$, BIC is defined to be

$$BIC(\mathcal{M}) = \log\{f_{\hat{\mu}}(x)\} - \frac{p}{2}\log(n)$$

END !

Next time: Penalized Likelihood