Quant III

Lab 10: High Dimensional Model

Junlong Aaron Zhou

November 17, 2020

Regression

- $E(Y|X) = X'\beta$
- Minimize mean squared error
- $Min_{\beta}||Y - X'\beta||_2^2$
- where $||X||_2 = \sqrt{\sum X^2}$
- Bias-variance trade-off

Regularization

- One approach: restrict the size of β
- Why?
- Put some constraint: $\sum \beta^2 \leq t$
- Change our target:

$$\min_{\beta} ||Y - X'\beta||_2^2$$
$$s.t. \sum \beta^2 \leq t$$

Ridge regression

- The previous target is equivalent to:

$$\underset{\beta}{min}||Y - X'\beta||_2^2 + \lambda||\beta||_2^2$$

- How to choose λ?
- Our final object: generalizability of our model
- Choose λ via cross validation

LASSO regression

- Ridge does not select variable, why?
- LASSO:

$$\min_{\beta} ||Y - X'\beta||_2^2 + \lambda||\beta||$$

- Elastic net:

$$\min_{\beta} ||Y - X'\beta||_2^2 + \alpha\lambda||\beta|| + (1-\alpha)\lambda||\beta||_2^2$$

Bayesian Model Average

- The idea is that inference is always conditioned on a model.
- Hierarchical DGP:
  - The nature picks a model.
  - It generates data from that model.
- Estimation is also hierarchical, accordingly:
  - Estimate probability that data are generated from a given model.
  - Estimate parameters of that model.
  - Estimate parameters for each (if feasible) possible model.

BMA: setup

- Let $\mathcal{M} = (M_1, ..., M_K)$ be the set of all possible models.
- Let $\boldsymbol{\Theta} = (\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(K)})$ be sets of parameters associated with each model.
- We then estimate:
  - Posterior probability of each model: $\Pr(M = M_k | \boldsymbol{Y})$, $k = 1, ..., K$.
  - Posterior distribution of parameters of each different model: $p(\boldsymbol{\theta}^{(k)} | M_k, \boldsymbol{Y})$, $k = 1, ..., K$.
  - We can select the highest posterior probability model as our 'preferred' model.
  - Alternatively, we can report model-averaged parameter estimates:

$$p(\boldsymbol{\theta} | \boldsymbol{Y}) = \sum_{k=1}^{K} p(\boldsymbol{\theta}^{(k)} | M_k, \boldsymbol{Y}) \Pr(M = M_k | \boldsymbol{Y}).$$

BMA: regression example (1)

- Suppose that $\boldsymbol{X}$ consists of two predictors, which implies that we can have the following possible models (in a linear additive world):

$$
\begin{aligned}
y_i | M_0 &\sim \mathcal{N}(\beta_0^{(0)}, \sigma_0^2),\\
y_i | M_1 &\sim \mathcal{N}(\beta_0^{(1)} + \beta_1^{(1)} x_{i1}, \sigma_1^2),\\
y_i | M_2 &\sim \mathcal{N}(\beta_0^{(2)} + \beta_2^{(2)} x_{i2}, \sigma_2^2),\\
y_i | M_3 &\sim \mathcal{N}(\beta_0^{(3)} + \beta_1^{(3)} x_{i1} + \beta_2^{(3)} x_{i2}, \sigma_3^2),
\end{aligned}
$$

BMA: regression example(2)

- Suppose we estimate each of the four sets of regression coefficients.
- Based on these estimates, we can calculate the posterior probability for each model $\Pr(M = M_k|\boldsymbol{Y})$, $k = 0, ..., 3$.
- Suppose $\Pr(M = M_1|\boldsymbol{Y}) = 0.9$: this means that there is only 90 percent chance that the model with the first predictor is the 'correct' one (closest to the true model).

BMA: regression example (3)

- The model-averaged coefficients are shrunken towards zero

$$
\begin{aligned}
E(\beta_1|\boldsymbol{Y}) = & \Pr(M = M_0|\boldsymbol{Y}) \times 0 + \Pr(M = M_1|\boldsymbol{Y}) \times \beta_1^{(1)} \\
& + \Pr(M = M_2|\boldsymbol{Y}) \times 0 + \Pr(M = M_3|\boldsymbol{Y}) \times \beta_1^{(3)}.
\end{aligned}
$$

- This is shrinkage/regularization (in addition to shrinkage due to priors).

BMA: alternative specification

- Let $Y \sim \mathcal{N}(\boldsymbol{X}'\boldsymbol{\beta}, \sigma^2)$, where $\boldsymbol{X}$ has p columns, so we have p regression coefficients $\boldsymbol{\beta} = (\beta_0, ..., \beta_{p-1})$
- Suppose now that we specify the following priors for each β_j, $j = 1, ..., p$:

$$\beta_j \sim \pi_j \mathbf{I}\{\beta_j = 0\} + (1 - \pi_j)\mathcal{N}(\beta_0, \tau_0^2).$$

- A priori with the probability π_j, β_j is equal to zero, and with the probability $1 - \pi_j$ it is drawn from the normal distribution.
- Prior of π_j will be the prior over the model where variable j is not included.
- A posteriori, we will estimate the probability that the coefficient β_j is identically equal to zero, that is, $\pi_j | \boldsymbol{Y}$.
- Strong predictors will have large posterior π_j, and vice versa.