ORF525 - Class Notes

Bachir EL KHADIR

February 24, 2016

Class 1

Definition 1 (Oridnary Lease Squares Regression). $f_i = \{f(x) = \beta^T X\}$ $\hat{\beta}^{OLS} = \arg\min_\beta ||Y - X\beta||_2^2$ $F(\beta) = Y^TY + \beta^T X^T X\beta - 2\beta^T X^T Y$ $\frac{\partial F(\beta)}{\partial \beta} = 2X^T X\beta - 2X^T Y = 0 \implies \hat{\beta} = (X^T X)^{-1} X^T Y$

Definition 2 (Model-based Interpretation of OLS). *Statistical Model* $Y = \beta^T X + \varepsilon, \varepsilon \sim \mathcal{N}(0, 1)$ *Joint-Loglikelihood*

$$l_n(\beta, \sigma^2) = f \sum_{i=1}^n \log p_{\beta,\sigma^2}(Y_i, X_i) = \sum_{i=1}^n \log p_{\beta,\sigma^2}(Y_i|X_i) + \underbrace{\sum_{i=1}^n \log p(X_i)}_{\textit{does not depend on } \beta \textit{ or } \sigma^2}$$

$\implies$

$$\begin{aligned}
\arg\max_{\beta,\sigma^2} l_n(\beta, \sigma^2) &= \arg\max_{\beta,\sigma^2} \underbrace{\sum_{i=1}^n \log p_{\beta,\sigma^2}(Y_i|X_i)}_{\textit{Conditional log-likelihood}} \\
&= \arg\max_{\beta,\sigma^2} \frac{1}{2\sigma^2} \sum (Y_i - \beta^T X_i)^2 + n \log(\frac{1}{\sqrt{2\pi\sigma^2}})
\end{aligned}$$

$\implies \hat{\beta}^{MLE} = \arg\min \sum (Y_i - \beta^T X_i)^2 = \hat{\beta}^{OLS}$

1 Linear Regression with Basis Expansion

From linear to non linear

- Input vairables can be transofrmation of original feautres: Handraft features, Box-Cox tranformation (find the best transmformation)
- Input can have interactions, eg $X_1X_2\ldots$
- Inputs can have basis expansions. Instead of $f(x) = \beta^T x$ we can have $f(x) = \sum_j \beta_j \underbrace{h_j}_{\text{Adaptative learning}} (x)$.

Definition 3 (Categorical Variable)**.** *A variable that can take on only one of a limited values.* ***Dummy coding***

2 High Dimensional Regression

Definition 4 (High Dimensional Regression)**.** *Data when dimension d is bigger than the sample size n.*

$$Y = \begin{pmatrix} Y_1 \\ \cdots \\ Y_n \end{pmatrix}$$

$$X = \begin{pmatrix} X_{11} & \dots & X_{1n} \\ & \cdots & \\ X_{n1} & \dots & X_{nn} \end{pmatrix}$$

Question: $\hat{\beta}^{OLS} = (\underbrace{X^T X}_{\text{not invertible}})^{-1} X^T Y$, what should we do?

- Ridge Estimation $\hat{\beta}^{\lambda} = (\underbrace{X^T X + \lambda I}_{\text{Tuning Parameters}})^{-1} X^T Y$

$$\iff \hat{\beta}^{\lambda} = \arg\min_{\beta \in \mathbb{R}^d} ||Y - X\beta||_2^2 + \lambda ||\beta||_2^2$$

$$\iff \hat{\beta}^{t} = \arg\min_{||\beta||_2^2 < t} ||Y - X\beta||_2^2$$

- Computation of Ridge:
  - Convex Optimzation (QP)
  - Never naively use a *general-purpose* solver. (CVX, AMPL)
- Question: How to choose the tuning parameter λ? Model selection: $\Lambda = \{\lambda_1, \dots, \lambda_n\}$ Basic Method: $D = D_1 \cup D_2$, let $\hat{\beta}^{\lambda_1}, \dots, \hat{\beta}^{\lambda_k}$ be ridge estimators on D_1. We define the data split score $DS(k) = \frac{1}{n^2} \sum_{D_2} (Y_i - X_i^T \hat{\beta}^{\lambda_k})^2$ We then pick the model with the smallest DS score. Intuition: Conditioning on D_1, $DS(k)$ is an unbiased estimator of $R(\hat{\beta}^{\lambda_1})$. Pro:Theoritically and conceptually simple. Con: Waste of the training sample. $\implies$ Cross validation.

Class 2

[Data spliting]

2.0.1 Pros and cons o f data splitting

Pro: Theoritically and computationaly simple. **Con:** Waste if training data $\implies$ cross validation.

- training / test split: conditional (on the training) prediction error.

$$\mathbb{E}_{X,Y}[|Y - \hat{f}_{D_{train}}(X)|^2 | D_{train}]$$

- cross validation: converges to expected training data.

$$\mathbb{E}_D[\mathbb{E}[|Y - \hat{f}_{D_{train}}(X)|^2 | D_{train}]]$$

Definition 5 (J-Fold Cross validation)**.** *We split the data $\mathcal{D}$ into J-equally sized parts $\mathcal{D}_1, \ldots, \mathcal{D}_J$. This forms:*

$$\begin{aligned}
(DS1) :&\mathcal{D}_1 \ \ vs \ \mathcal{D} \setminus \mathcal{D}_1 \\
&\cdots \\
(DS1) :&\mathcal{D}_1 \ \ vs \ \mathcal{D} \setminus \mathcal{D}_n
\end{aligned}$$

For $\lambda_k \in \Lambda$ we calculate the data splitting scores Using $DS1, \ldots DS2$. Denote the result as $DS_1(k), \ldots, DS_J(k)$. The cross validation is

$$CV(k) := \frac{1}{J} \sum_{j}^{J} DS_j(k)$$

We then pick $\arg\min CV(k)$*. In pratice, picke the most parsimonious model whose error is no more than one standard deviation above the smallest CV score.*

Question: After CVm we pick $\hat{\lambda}_k$. Then what shall we do?

- Use λ_k to fit the entire data, then deliver
- Take the average of the estimators.

2.0.2 Model assessment vs selection

Definition 6 (Lasso)**.** *Bridge estimator with $\beta = 1$ Least absolute shrinkage and selection operator Sparsity: Intersection of ellipsoid ($||Y - X\beta||_2^2 = cte$) and a polytope $||\beta||_1 = cte$)*

Sparsity: many ekements of β are 0 $\implies$ model selection. (select variable with coefficient $\neq 0$)

———————-

Class 3

[Persistency]

Ridge	Lasso
Not Sparse	Sparse
Handles collinearity	Doesn't handle collinearity

Definition 7 (Collinearity). *A phenomenon in which two or more predictor variables are highly correlated.*

Question : Combine Ridge and Lasso? Answer: **Elastic-Net**

$$\hat{\beta}^{\text{Elastic}} = \arg\min ||Y - X\beta||_2^2 + \lambda(\alpha||\beta|||_1 + (1-\alpha)||\beta||_2^2)$$

- $\alpha = 1 \implies$ Lasso.
- $\alpha = 0 \implies$ Ridge.

Question: two tuning parameters, how to choose then? Answer: Use a two stage approach:

- Use $\alpha = 1$, fit a full Lasso path, visualize the regularization path.
- Use $\alpha = 0.6$, fit the regularization path pagain. Then we examine whether there si significant change of the final path:
  - If not $\implies \alpha = 1$ (Lasso)
  - o/w $\implies \alpha = 0.6$ (Elastic)

Insight of the Lasso Estimator

Definition 8 (SQRT-Lasso). *An equivalent representation of the lasso is called SQRT-Lasso:*

$$\hat{\beta}^{Elastic} = \arg\min ||Y - X\beta||_2^2 + \lambda||\beta||_1 \tag{1}$$

Symptoticcaly $\lambda^{optimaly} \sim 2.1\sqrt{\frac{t}{\log d}n}$, $n > 10k$ + *The model has to be linear*

Theorem 1 (Robust Optimization Representation of Lasso). *The SQRT-Lasso problem in (1) is equivalent to the following robust linear regression problem:*

$$\min_{\beta} \max_{U \in \Omega_\lambda} ||Y - (X+U)\beta||_2$$

Where $\Omega_\lambda := \{U = (U_1, \ldots, U_d) \in \mathbb{R}^{n \times d}, \max_j ||U_j||_2 \leq \lambda\}$

Proof. We only need to prove $\max_{U \in \Omega_\Lambda} ||Y - (X+U)\beta||_2 = ||Y - X\beta||_2 + \lambda||\beta||_1$

- $\max_{U \in \Omega_\Lambda} ||Y - (X+U)\beta||_2 \leq ||Y - X\beta||_2 + \lambda||\beta||_1$ $||Y - (X+U)\beta||_2 \leq ||Y - X\beta||_2 + \sum_j |\beta_j| \, ||U_j||_2 \leq ||Y - X\beta||_2 + \lambda||\beta||_1$

- $||Y - X\beta||_2 + \lambda||\beta||_1 \leq \max_{U \in \Omega_\Lambda} ||Y - (X+U)\beta||_2$

$$u = \begin{cases} \frac{Y - X\beta}{||Y - X\beta||_2} & \text{if } Y \neq X\beta \\ \text{arbitrary unit vector} & \text{o.w} \end{cases}$$

And define:

$$U_j^* = -\lambda sign(\beta_j) u$$

$(sign(0) = 1)$

We can verify that $|U_j|_2 \leq \lambda$

$$\begin{aligned} |(Y - (X+U^*)\beta|_2 &\geq |(Y - X\beta - \sum_j \beta_j U_j^*|_2 \\ &\geq |(Y - X\beta - \sum_j |\beta_j| \frac{Y - X\beta}{||Y - X\beta||_2} \\ &= |(|Y - X\beta)|_2 + \lambda|\beta|_1) \frac{Y - X\beta}{||Y - X\beta||_2}|_2 \\ &= |Y - X\beta|_2 + \lambda|\beta|_1 \end{aligned}$$

□

Definition 9 (Theory of Lasso (Greenshtein and Ritov '2006))**.** *We define*

$$R(\beta) = E_{Y,X}(Y - \beta^T X)^2, \hat{R}(\beta) = \frac{1}{n}\sum_j (Y_j - \beta^T X_i)^2$$

$\hat{\beta} = \arg\min_{|\beta|_1 \leq L} \hat{R}(\beta)$*: Lasso estimator* $\beta^* = \arg\min_{|\beta|_1 \leq L} R(\beta)$*: Lasso estimator*

Definition 10 (Persistence)**.** *An estimator* $\hat{\beta}$ *is persistent within a class* $\mathcal{B}_n$ *if* $R(\hat{\beta}) - \inf_{\beta \in \mathcal{B}_n} R(\beta) \to_{\mathbb{P}} 0$ *as* $n \to \infty$

Theorem 2 (Lasso)**.** *Assume* $|Y_i| \leq B$ *and* $|X|_\infty \leq B$*. Then*

$$P\left(R(\hat{\beta}) - R(\beta^*) \leq 2(1 + L^2)\sqrt{\frac{2B^4 \log(\frac{2d^2}{\delta})}{n}} \right) \geq 1 - \delta$$

Proof.

$$\begin{aligned} R(\hat{\beta}) - R(\beta^*) &= R(\hat{\beta}) - \hat{R}(\hat{\beta}) + \hat{R}(\hat{\beta}) - R(\beta^*) \\ &\leq R(\hat{\beta}) - \hat{R}(\hat{\beta}) + \hat{R}(\beta^*) - R(\beta^*) \\ &\leq 2 \sup_{\|\beta\|_{L_1} \leq L} |R(\beta) - R(\hat{\beta})| \end{aligned}$$

Let $Z = (Y, X^T)^T$, $r = (-1, \beta^T)^T$

$$R(\beta) = \mathbb{E}(Y - \beta^T X)^2 = \mathbb{E}(r^T Z Z^T r) = r^T \mathbb{E}(Z Z^T) r = r^T \Sigma r$$
$$\hat{R}(\beta) = \frac{1}{n} \sum (Y_i - \beta^T X_i)^2 = r^T \frac{1}{n} \sum Z_i Z_i^T r = r^T \hat{\Sigma} r$$

Therefore

$$\begin{aligned}\sup_{\|\beta\|_{L_1} \leq L} |R(\beta) - R(\hat{\beta})| &= \sup |r^T (\hat{\Sigma} - \Sigma) r| \\ &\leq \|r(\beta)\|_1^2 \|\hat{\Sigma} - \Sigma\|_\infty\end{aligned}$$

By Hoeding

$$\mathbb{P}(\|\hat{\Sigma} - \Sigma\|_\infty > t) \leq \sum \mathbb{P}(\hat{\Sigma}_{ij} - \Sigma_{ij} > t) \leq 2d^2 \exp(-\frac{nt^2}{2B^4})$$

$\square$

Theorem 3 (Persistency of the Lasso)**.** $\forall k > 0, d = O(n^k)$, $\mathcal{B}_n = \{\beta, |\beta|_1 \leq L_n, L_= o(\frac{n}{\log n})^{\frac{1}{4}}\}$ *Then:* $R(\hat{\beta}) - \inf_{\beta \in \mathcal{B}_n} R(\beta) \to_{\mathbb{P}} 0$ *as* $n \to \infty$