STAT 151A Lecture 30

Henry Liev

5 November 2025

Remark 0.1

Method	Initial Model	Add/Drop?	Criterion
Forward	Intercept only	Add	ANY
Backward	All Variables	Drop	ANY
Stepwise	Any	Either	ANY

Reminder: You shouldn't compute p-values/CI in the same dataset you used for model selection
Interactions: Don't paly well with variable selection
Categorical Variables: Handling depends on R package (leaps, step, stepAIC)
Intuitively: as $\lambda \uparrow$ we "shrink" coefficients $\hat{\beta}_j$ in full model towards zero
Shrink approaches: Ridge vs LASSO

Remark 0.2 (How to Implement Shrinkage)

Put all variables on some scale $\to$ standardized design matrix
$\mathbf{X}$ drop intercept, replace each value x_{ij} by $\frac{x_{ij}-\bar{x}_j}{s_{x_j}}$
Call to new design matrix $\mathbf{Z}$
Also: Center $\vec{Y} \to \vec{Y} - \bar{Y}\vec{1}$
Think about OLS
$\min_\beta \|\vec{y} - \mathbf{Z}\beta\|^2$
Goal: Get good predictions while forcing β in to be a bit smaller.
Attempt:
$\min_{\beta, S(\beta) \le c} \|\vec{y} - \mathbf{Z}\beta\|^2$
$S(\beta) = \sum \beta_j^2$ (Ridge)
$S(\beta) = \sum |\beta_j|$ (LASSO)
Equivalent to the following
$\min_\beta \left[\|\vec{y} - \mathbf{Z}\beta\|^2 + \lambda S(\beta)\right]$
For any c, $\exists \lambda$ st the two problems solve the same answer
Connection to variable selection
What if we choose $S(\beta) + \sum I_{\hat{\beta}_j \neq 0}$, which gives the count of variables in the model
All subsets regression:
Problem: This is a $0-1$ penalty (L^0 norm) not smooth, makes optimization non-convex
By switching to a smoother penalty, make problems much easier computationally. Guarantees global optimum.

Remark 0.3 (Ridge Regression)

$S(\beta) = \sum \hat{\beta}_j^2$, Let us rename $\mathbf{Z}$ as $\mathbf{X}$
Solve $\min \|y - \mathbf{Z}\beta\|^2 + \lambda \|\beta\|^2$
$\min_\beta (y - \mathbf{X}\beta)^\intercal (y - \mathbf{X}\beta) + \lambda \beta^\intercal \beta$
$\min_\beta y^\intercal y - 2y^\intercal \mathbf{X}\beta + \beta^\intercal \mathbf{X}^\intercal \mathbf{X}\beta + \lambda\beta^\intercal\beta = y^\intercal y - 2y^\intercal \mathbf{X}\beta + \beta^\intercal[\mathbf{X}^\intercal\mathbf{X} + \mathbb{I}\lambda]\beta$
$\nabla_\beta[-2\mathbf{X}^\intercal y + 2(\mathbf{X}^\intercal\mathbf{X} + \lambda\mathbb{I})\beta] = 0$
$\hat{\beta}_\lambda = (\mathbf{X}^\intercal\mathbf{X} + \lambda\mathbb{I})^{-1}\mathbf{X}^\intercal\vec{y}$
Linear function of $\vec{y}$
It is not a projection
Why? Projection $\mathbb{H}\vec{y}$ satisfies $\mathbb{H}^\intercal = \mathbb{H}, \mathbb{H}\mathbb{H} = \mathbb{H}$, but regularization does not satisfy idempotency
$\mathbb{H}\mathbb{H} \neq \mathbb{H}$

Remark 0.4 (Bias-Variance Tradeoff)

Assume NLM
Bias of $\hat{\beta}_\lambda$
$\mathbb{E}\left[(\mathbf{X}^\intercal\mathbf{X} + \lambda\mathbb{I})^{-1}\mathbf{X}^\intercal y\right]$
$= \mathbb{E}\left[(\mathbf{X}^\intercal\mathbf{X} + \lambda\mathbb{I})^{-1}(\mathbf{X}^\intercal\mathbf{X})^(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\vec{y}\right] = (\mathbf{X}^\intercal\mathbf{X} + \lambda\mathbb{I})^{-1}(\mathbf{X}^\intercal\mathbf{X})\beta = U_\lambda\beta$
$\text{Bias}(\hat{\beta}_\lambda) = \mathbb{E}(\hat{\beta}_\lambda) - \beta = [U_\lambda - \mathbb{I}]\beta$
As $\lambda \uparrow, \text{Bias} \uparrow$
Variance $\hat{\beta}^\intercal\hat{\beta}$ vs $\hat{\beta}_\lambda^\intercal\hat{\beta}_\lambda$
$\vec{y}^\intercal\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal y \geq y^\intercal\mathbf{X}(\mathbf{X}^\intercal\mathbf{X} + \lambda\mathbb{I})^{-1}\mathbf{X}^\intercal y$, strict inequality when $\lambda > 0$, $\text{Var}(\hat{\beta}) > \text{Var}(\hat{\beta}_\lambda)$
Inference: Not allowed on same data, generally not useful since $\hat{\beta}_{\lambda,j}$ are biased