Linear Methods for Regression (Chap 3 ESL) Date: 10/23

I. Introduction.

A linear regression model assumes that the regression function $\mathbb{E}[Y|X]$ is linear in the inputs $X_1, \dots, X_p$. Reminder: $\mathbb{E}[Y|X] = \varphi(X)$.

II. Linear Regression Models and Least Squares.

. We have an input vector $X \in \mathbb{R}^p$ and we want to predict an output $Y \in \mathbb{R}$.
. Linear regression model: $Y = f(X) + \varepsilon$ where $f(X) = \beta_0^* + \sum_{j=1}^{p} X_j \beta_j^*$ and $\mathbb{E}[\varepsilon] = 0$
. The coefficients $\beta_1, \dots, \beta_p$ are unknown and the variables (X_j) can come from different sources: quantitative inputs, transformation of quantitative inputs ($\log$, $\sqrt{\cdot}$, $\cdot^2$), ...
. The model is linear in the parameters.
. Data: Collect $(y_1, x_1), \dots, (y_n, x_n)$ where $\forall i \in [n]$ $y_i \in \mathbb{R}$ and $x_i \in \mathbb{R}^p$.
. From those data we estimate $\beta^* := (\beta_0^*, \dots, \beta_p^*)$ (in ML we "learn").
. How? Minimize the residual sum of squares: $\hat{\beta}_{LS} \in \operatorname{argmin} \|Y - X\beta\|_2^2$.
. $RSS(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$. How to minimize? Matrix notation!
. $X \in \mathbb{R}^{n \times (p+1)}$, matrix with each row being an input vector x_i (with 1 in first position).
. $y \in \mathbb{R}^n$ is the vector of outputs and $\beta \in \mathbb{R}^{p+1}$ is the parameter to "learn".
↳ $RSS(\beta) = (y - X\beta)^T (y - X\beta)$ and $\frac{\partial RSS}{\partial \beta}(\beta) = -2X^T(y - X\beta)$ and $\frac{\partial^2 RSS}{\partial \beta^2}(\beta) = 2X^TX$.
↳ If X is full rank (ie $\operatorname{rank}(X) = p+1$) then $\operatorname{rank}(X^TX) = p+1$ and X^TX invertible.
Moreover X^TX invertible ensures X^TX is positive definite ($\forall u \in \mathbb{R}^{p+1}$, $u^TX^TXu = \|Xu\|^2 > 0$).
Hessian positive-definite ensures convexity of the function. Hence RSS is convex in β.
This implies that any critical point is the global minimizer. Hence to minimize RSS it suffices to find $\hat{\beta}$ s.t. $RSS'(\hat{\beta}_{LS}) = 0$ ie $X^TX\hat{\beta}_{LS} = X^TY$.
↳ $\hat{\beta}_{LS} = (X^TX)^{-1}X^TY$. where $Y = X\beta^* + \varepsilon$.
↳ The predicted value at an input vector x_{n+1} is $(1, x_{n+1,1}, \dots, x_{n+1,p})^T \hat{\beta}$.
↳ The fitted values at the training inputs are $\hat{y} = X\hat{\beta} = X(X^TX)^{-1}X^TY$.
. What happens if X^TX is not invertible?
↳ perfect Multicolinearity: one predictor is a linear combination of the others → remove it.
↳ High-dimension: $p+1 > n$ → Regularization techniques.
. What happens when mild multicolinearity: predictors have close to exact linear relationship.
↳ LS estimates for β_j is well defined but have large variance → Regularization.
↳ Ridge regression for example.

1. Gauss-Markov theorem (why $\hat{\beta}^{LS}$ and not minimizing another criterion?)

. The least squares estimate $\hat{\beta}^{LS}$ has the smallest variance among all linear unbiased estimators.
. Estimation of any linear combination of the parameters: $\theta = a^T\beta^*$.
↳ $\hat{\theta}_{LS} = a^T\hat{\beta}_{LS} = a^T(X^TX)^{-1}X^Ty$. → $\mathbb{E}[a^T\hat{\beta}_{LS}] = a^T(X^TX)^{-1}X^T\mathbb{E}[y] = a^T(X^TX)^{-1}X^TX\beta = a^T\beta^*$.
↳ If we have any other linear estimator $\tilde{\theta} = c^Ty$ that is unbiased ($\mathbb{E}[\tilde{\theta}] = a^T\beta^*$):

$$\mathbb{V}(\hat{\theta}_{LS}) \leq \mathbb{V}(\tilde{\theta})$$

Date:

For any estimator $\tilde{\theta}$ of θ^* : $MSE(\tilde{\theta}) = \mathbb{E}_{\theta^*}[(\tilde{\theta}-\theta^*)^2] = \mathbb{V}_{\theta^*}(\tilde{\theta}) + (\mathbb{E}[\tilde{\theta}]-\theta^*)^2$.

↳ Variance + Squared bias. Gauss-Markov → The LS estimator has the smallest MSE among all unbiased linear estimators.

↳ However we may find a biased estimator with smaller MSE.

↳ Add a little bias for a huge reduction in variance.

↳ Any estimator that shrinks the coefficients of the LS estimator is biased.

III. Shrinkage Methods

1. Ridge Regression.

$\hat{\beta}^R_\lambda \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|Y - X\beta\|_2^2 + \lambda \|\beta_{1:}\|_2^2$. (We omit β_0 in the penalty).

Idea: When there is multicolinearity, the coefficients of the parameter are poorly determined by the OLS estimator. A wildly large positive coefficient on one variable can be canceled by a similarly large negative coefficient on its correlated cousin. Ridge imposes a size constraint on the coefficients → reduces this problem.

How to find $\hat{\beta}^R_\lambda$? → Objective is differentiable.

↳ $RSS(\beta) = (y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta \rightarrow \frac{\partial RSS}{\partial \beta}(\beta) = -2X^T(y - X\beta) + 2\lambda\beta$.

↳ $\frac{\partial^2 RSS}{\partial \beta^2}(\beta) = 2(X^TX + \lambda I_p)$ is positive definite because $\forall u \in \mathbb{R}^p : 2u^TX^TXu + 2\lambda u^Tu \geq 0$.

↳ $\beta \mapsto RSS(\beta)$ is convex and thus $\hat{\beta}^R_\lambda$ satisfies $\frac{\partial RSS}{\partial \beta}(\hat{\beta}^R_\lambda) = 0 \rightarrow \hat{\beta}^R_\lambda = (X^TX + \lambda I_p)^{-1}X^Ty$.

2. Lasso.

$\hat{\beta}^L_\lambda \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|Y - X\beta\|_2^2 + \lambda \|\beta_{1:}\|_1$ (we omit β_0 in the penalty).

↳ No closed form. Quadratic programming problem. Efficient algorithms with same computational cost as for ridge.