Lecturer: Prof. Homrighausen Scribe: Veronica Burt

1 Previous Lectures

We've been discussing

- Bias/variance trade off when selecting a model
- Splitting the model into training/testing components
- Non-convex optimization problems, which lack an efficient solution method

2 Regularization

- Consider having one response variable, and a large number of features or covariates.
- Classically, we use least squares to solve this problem (which minimizes squared error), and then drop unnecessary covariates from the model using marginal tests.
- Instead of removing dimensions (which is what happen when we set a $\hat{\beta} = 0$), consider shrinking them to an area around the origin.
- Regularization can always result in a smaller risk than not regularizing if we know which tuning parameters to pick.

3 Overview of Ridge Regression

- Ridge Regression is also referred to as Tikhonov Regularization.
- It is one of many methods proposed to solve problems that arise from multicollinearity. Multicollinearity causes our coefficient estimates to be unstable. To control this problem, Ridge Regression provides a biased estimate of $\hat{\beta}$ in hopes that these estimates are more precise than the unbiased estimates.
- It still uses minimized squared error criterion, but subject to $||\beta||_2^2 \leq t$, where t is fixed ≥ 0.
- When $t = 0$, all of the β values are set equal to zero.
- When $t = \infty$, we have ordinary least squares regression.
- This method confines $\hat{\beta}$ to a region around the origin.
- We know that regularization can always result in a smaller risk, but the question is now "Which t value results in a smaller risk than ordinary least squares?"
- λ and t are called tuning parameters, or hyper parameters. They are data dependent.
- Another way to write $\hat{\beta}$ is in the LaGrangian form, $\hat{\beta}_{ridge} = \text{argmin } (||\boldsymbol{Y} - \mathbb{X}\beta||_2^2 + \lambda||\beta||_2^2$.

- We have a plethora of solutions based on the value we choose for λ (or t). For every value of λ, there is a unique value of t, and vice-versa.
- We can choose λ via a risk estimation procedure.

4 Standardization

- Coefficient vectors are not invariant to rescaling.
- Do not penalize the intercept if it is included in the model.
- To address the invariance issue:

  Standardize covariates first by subtracting the mean and dividing by the standard deviation; this makes them unit-less. However, don't standardize indicator variables since they are already dimensionless.

  Standardize the response by subtracting the mean. This helps us stabilize numerical problems.

  Don't include the intercept since it would be equal to the mean of the response variable.

5 Uniqueness

- Note that in ordinary least squares, there are an infinite number of solutions for $\hat{\beta}$ if $\mathbb{X}$ is rank deficient (meaning that the rank($\mathbb{X}$) $<p$); if $\mathbb{X}\mathbf{b} = 0$ then $\hat{\beta} + \mathbf{b}$ is a valid least squares solution.
- However, as long as $\lambda > 0$, the solution for $\hat{\beta}_{ridge}$ is always unique. The solution is

  $\hat{\beta}_{ridge} = (\mathbb{X}^T\mathbb{X} + \lambda I)^{-1}\mathbb{X}^T\boldsymbol{Y}$
- Using the Singular Value Decomposition ($\mathbb{X} = UDV^T$), we can see that estimates of β can be written as:

  $\hat{\beta}_{LS} = VD^{-1}U^T\boldsymbol{Y} = \sum_{j=1}^{p} \boldsymbol{v}_j(\frac{1}{d_j})\boldsymbol{u}_j^T\boldsymbol{Y}$

  $\hat{\beta}_{ridge} = V(D^2 + \lambda I)^{-1}DU^T\boldsymbol{Y} = \sum_{j=1}^{p} \boldsymbol{v}_j(\frac{d_j}{d_j^2+\lambda})\boldsymbol{u}_j^T\boldsymbol{Y}$
- If we look at the predictions formed by these estimates, we see

  $\mathbb{X}\hat{\beta}_{LS} = (UDV^T)(VD^{-1}U^TY) = (UU^TY) = \sum_{j=1}^{p} \boldsymbol{u}_j\boldsymbol{u}_j^T\boldsymbol{Y}$

  $\mathbb{X}\hat{\beta}_{ridge} = (UDV^T)(V(D^2 + \lambda I)^{-1}DU^TY) = UD(D^2 + \lambda I)^{-1}DU^TY = \sum_{j=1}^{p} \boldsymbol{u}_j(\frac{d_j^2}{d_j^2+\lambda})\boldsymbol{u}_j^T\boldsymbol{Y}$
- This leads us to conclude that Ridge Regression shrinks the data by an additional factor of λ.

6 Bayes Approach

- Consider a hierarchical model where the likelihood of $Y_i \sim N({X_i}^T\beta, \sigma^2)$ and consider a prior of $\beta \sim N(0, \tau^2 I)$.
- Making some conditional independence assumptions, our posterior is $p(\beta|Y, X, \sigma^2, \tau^2) \propto p(Y|, \beta, \sigma^2)p(\beta|\tau^2)$.
- After kernel matching, we find that the posterior mean is $\lambda = \sigma^2/\tau^2$.

7 Computation

- The Woodbury Identity states that $(A - BC^{-1}E)^{-1}BC^{-1} = A^{-1}B(C - EA^{-1}B)^{-1}$.
- Applying this to the ridge solution for $\hat{\beta}$ shows us
  $\hat{\beta}_{ridge} = (\mathbb{X}^T\mathbb{X} + \lambda I)^{-1}\mathbb{X}^T\boldsymbol{Y} = \mathbb{X}^T(\mathbb{X}\mathbb{X}^T + \lambda I)^{-1}\boldsymbol{Y}$
- This results in only having to invert an $n \times n$ matrix as opposed to a $p \times p$ matrix, which can be much less expensive in terms of computation time in the big data setting.

8 Kernel Ridge Regression

- Suppose we want to make a prediction for X, then our prediction is
  $\hat{f}(X) = X^T\hat{\beta}_{ridge} = X^T\mathbb{X}^T(\mathbb{X}\mathbb{X}^T + \lambda I)^{-1}\boldsymbol{Y}$
- If we transform $X_i \mapsto \phi(X_i)$ and the range of ϕ is equipped with an inner product, we can use $\langle\phi(X_i), \phi(X_{i'})\rangle$. This is known as kernelization; a technique that makes algorithms more efficient by using a preprocessing stage to reduce the inputs to the algorithm.

9 The Tuning Parameter

- To select the tuning parameter λ, we can use a risk estimator based on degrees of freedom.
- For ridge regression, the degrees of freedom is df =trace $[\mathbb{X}(\mathbb{X}^T\mathbb{X} + \lambda I)^{-1}\mathbb{X}^T] = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2+\lambda}$
- Note that as λ approaches zero, we get the number of parameters in our model, p.
- A typical choice is Generalized Cross-Validation (GCV), which is calculated using $GCV(\hat{\beta}) = \frac{\mathbb{P}l_\beta}{(1-df(\hat{\beta})/n)^2}$.
- GCV has similar behavior to AIC, but with a different penalty: $log(GCV(\hat{\beta})) \propto log(\hat{R}_{train}) - 2log(1 - df(\hat{\beta})/n)$ compared to $AIC(\hat{\beta}) \propto log(\hat{R}_{train}) + 2n^{-1}df(\hat{\beta})$.
- Using K-fold cross-validation is also common.
- Think of CV_K as a function of λ. Choose the λ for which CV_K is minimized. Then use this λ value to compute $\hat{\beta}_{ridge}$.