CASI 7.3 and Ch 16 Highlights Solution

Some concepts to clarify before discussing the new techniques themselves.

What is a shrinkage estimator? (Ch. 7 intro)

An estimator where deliberate bias is introduced in order to improve overall performance (@ a possible danger to individual estimates). (Neat quote)

What does regularization refer to? (7.3)

"Generically, regularization describes almost any method that tamps down statistical variability in high-dimensional estimation or prediction problems." pg. 106 student edition, CASI (ex. regression)

Describe the use of a training/test set in analysis.

Start from original data set. Split in 2 parts (70/30, 80/20, etc.). Fit model (or do analysis) on training data set. Then assess performance (such as model fit) on the test set. This way the test set is like "new" data for the model.

The new techniques will involve tuning parameters L^2 that must be chosen, often using cross-validation. (CV)

What is your understanding of CV? For example, describe what happens in a 10-fold CV. (Ch 16, 16.1)

k-fold CV $\Rightarrow$ $k = 10$. We split the data into 10 parts (a officially partition) usually of roughly equal size. Suppose we are trying to estimate MSE for a regression. We use 9 of the 10 parts to fit the model, and get MSE using it on the 10^{th} piece. I.E. 9 of 10 pieces act as training, last as test. We repeat this, rotating each piece in the role of the test data. This will give us 10 MSE estimates, which are then combined (often averaged) to get an overall estimate.

The jackknife approach, also called Leave-One-Out-CV (LOOCV) is a special case of CV.

What makes it special?

k = # of folds = n = # of observations

Next is highlights about new techniques Will implement in lab.

Ridge Regression

Matrix notation. ? ?

Review OLS

Regression setting. $y = X\beta + \varepsilon$ $\varepsilon \sim \underline{N}(\underline{0}, \sigma^2 I)$

Find $\hat{\beta}$ as $\hat{\beta} = (X'X)^{-1}X'Y$. The $\hat{\beta}$ is found to **minimize** the residual sum of squares, RSS.

$$RSS = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \Rightarrow \|y - X\beta\|^2$$

Ridge is designed to improve on the OLS solution.

Ridge **shrinks** β's towards 0, using a penalty term, with **tuning** parameter λ.

Ridge minimizes $RSS + \lambda \sum_{j=1}^{p}\beta_j^2$. ← This term is a shrinkage **penalty**.

matrix form: $\|y - X\beta\|^2 + \lambda\|\beta\|^2$ Have to pick λ.

$\lambda = 0$ is OLS. As you ↑ λ, what happens to the β estimates? They get closer and closer to 0. At $\lambda = \infty$, they would be 0.

Can ridge set any β's to 0? No.

The ridge penalty is an $\underline{l_2}$ penalty.

Remember to use standardized variables.

What is the Bayesian rationale for Ridge regression? (pg. 100 in pdf version of CASI) L4

The Math is shown in the text. A prior on β could be used ($N_p(0, \frac{\sigma^2}{\lambda} I)$) such that the resulting posterior mean is the same as the ridge estimate when evaluated. Ridge "amounts to an increased prior belief that β lies near 0." (pg. 100)

How do we pick λ? Assume you have a train/test data split, and a set of possible λ's.

Hint: Cross-validation. Explain in a few sentences.

Leave test set alone. Use CV on training data.

Compute an estimate of error for the regression using CV (i.e. 10-fold ⇒ 10 parts → 10 estimates averaged)

for each potential λ.

Then, choose the λ with the lowest CV error. / Extra - What Next?

Fit the model using that λ and all training data.

Finally, evaluate performance on test set.

Ch 16 - Sparse Modeling and the LASSO

L5

Key concepts - Regularization is required if $p > n$, b/c typical OLS will fail.

Techniques mentioned in chapter - Best subsets, forward (stepwise) regression, LASSO, LARS, elastic net

---

In Stat 230, you learned about best subsets, forward selection (forward stepwise), backward elimination, and stepwise regression. Briefly, recap how each of these variable selection procedures work, conceptually.

Best subsets - Fit models and find the "best" one (based on your choice of criteria) for each possible model size. Then choose one of the "best" models.

Forward selection - Start model with an intercept (or smallest desired model). Add the single "best" predictor to the model and refit. Add the next best predictor if the criteria to add is met, etc.

Backward Elimination - Start with largest desired model. Then drop predictors 1@ a time (refit after each drop) if chosen criteria is met. Drops worst each time.

Stepwise - Starts like forward but allows a drop like backward after each addition if conditions met.

LASSO - Least Absolute Shrinkage and Selection Operator

- Want to do better than forward selection @ picking variables
- Want to do better than ridge in terms of shrinkage.

Ridge added on l_2-penalty, $\lambda \sum_{j=1}^{p} \beta_j^2$.

Lasso adds on l_1-penalty, $\lambda \sum_{j=1}^{p} |\beta_j|$.

This means LASSO can set β's to 0.

(Again, use standardized variables).

Often view the entire collection of solutions, over different values of t, where $\lambda \sum |\beta_j| \leq t$. Look @ Figure 16.5 (the LASSO path ex.) and make sure you can explain what it shows in your own words. Explanation: Each variable has a line in the plot that shows how it's coef ($\hat{\beta}$) value changes as $t \uparrow$. The idea is t sets the bound on the $\sum_{j=1}^{p} |\beta_j|$. I.E. $t \geq \|\hat{\beta}\|_1$. At $t = 0$, all coefs are 0 (LHS), and as $t \uparrow$, the estimates approach the OLS estimates. You see more non-zero coefs as $t \uparrow$ and their overall magnitude $\uparrow$ also.

Fitting LASSO Models

L7

There is a lot of neat math explained here. I want to make sure a few points are clear.

What is the "active set" of variables for a fixed λ?

The active set is just the set of variables that have non-zero coefficients ($\hat{\beta}$'s) for that λ

If the active set of variables does not change (including the signs of their $\hat{\beta}$'s) between $\lambda_1 < \lambda_2$, then $\hat{\beta}(\lambda)$ is **linear** for $\lambda \in [\lambda_1, \lambda_2]$.

This means the coefficient profiles are continuous and **piecewise** linear over the range of λ.

The "knots" occur when the active set (or a sign of a $\hat{\beta}$) changes.

LAR - Least **Angle** Regression is an algorithm which capitalizes on the linearity properties above to fit the entire lasso regularization path.

The LAR algorithm needs an adjustment L8
to get the LASSO path to deal with
coefficient sign changes.

16.5 goes over some topics related to computing LASSO solutions in other settings
Some topics here could be future paper topics, as they are not covered in our courses.

16.6 touches on some inference ideas.

16.7 has connections to techniques we will see as we learn the other techniques

---

Last technique - Elastic net (in 16.5)

Elastic net bridges the gap between
Ridge and the LASSO b/c it uses a
penalty term of the form:

$$P_\alpha(\beta) = \frac{1}{2}(1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1$$

Your text uses the glmnet package for fitting and we will see it in lab.