Supplementary Materials

LASSO

Zubia Mansoor

November 27, 2019

Supplementary Materials I

ALGORITHMS FOR FINDING LASSO SOLUTIONS

- Least Angle Regression (LAR) provides an extremely efficient algorithm for computing the entire lasso path

Algorithm 3.2 *Least Angle Regression.*

1. Standardize the predictors to have mean zero and unit norm. Start with the residual $\mathbf{r} = \mathbf{y} - \bar{\mathbf{y}}$, $\beta_1, \beta_2, \ldots, \beta_p = 0$.
2. Find the predictor $\mathbf{x}_j$ most correlated with $\mathbf{r}$.
3. Move β_j from 0 towards its least-squares coefficient $\langle \mathbf{x}_j, \mathbf{r} \rangle$, until some other competitor $\mathbf{x}_k$ has as much correlation with the current residual as does $\mathbf{x}_j$.
4. Move β_j and β_k in the direction defined by their joint least squares coefficient of the current residual on $(\mathbf{x}_j, \mathbf{x}_k)$, until some other competitor $\mathbf{x}_l$ has as much correlation with the current residual.
5. Continue in this way until all p predictors have been entered. After $\min(N-1, p)$ steps, we arrive at the full least-squares solution.

Algorithm 3.2a *Least Angle Regression: Lasso Modification.*

4a. If a non-zero coefficient hits zero, drop its variable from the active set of variables and recompute the current joint least squares direction.

Reference: **Least Angle Regression**, Bradley Efron, Trevor Hastie, Iain Johnstone and Robert Tibshirani. The Annals of Statistics 2004, Vol. 32, No. 2, 407–499

- If there are grouped variables (highly correlated between each other), LASSO tends to select one variable from each group, ignoring the others
- To overcome this, Zou and Hastie introduced the **Elastic Net** in 2005:
  - encourages a grouping effect, where strongly correlated predictors tend to be in or out of the model together.

- Useful for high-dimensional sparse data:
    - the number of variables (p) is larger than the number of observations (n) but are also sparse $\rightarrow$ true coefficients has only few non-zero entries
- Computationally feasible method

.

Supplementary Materials IV

Limitations of LASSO

- When $p > n$ (the number of covariates is greater than the sample size) lasso can select only n covariates (even when more are associated with the outcome)
- If there are grouped variables (highly correlated between each other), LASSO tends to select one variable from each group, ignoring the others
- Additionally, even when $n > p$, if the covariates are strongly correlated, ridge regression tends to perform better

To address several shortcomings of LASSO, Zou and Hastie introduced the **Elastic Net** in 2005:

- convex combination of ridge and lasso
- enjoys a similar sparsity of representation
- encourages a grouping effect, where strongly correlated predictors tend to be in or out of the model together.
- particularly useful when the number of predictors (p) is much bigger than the number of observations (n)

K-fold Cross Validation

- Divide the set $\{1, 2, \ldots, n\}$ into K subsets (i.e.,folds) of roughly equal size, $F_1, \ldots, F_K$
- For $k = 1, \ldots, K$:
  - Consider training on $(x_i, y_i), i \notin F_k$, and validating on $(x_i, y_i), i \in F_k$
  - For each value of the tuning parameter $\theta \in \{\theta_1, \ldots, \theta_m\}$, compute the estimates on the training set, and record the total error on the validation set
- For each tuning parameter value θ, compute the average error over all folds
- Having done this, we get a cross-validation error curve and choose the value of tuning parameter that minimizes this curve

The effective degrees of freedom of the lasso in the framework of Stein's unbiased risk estimation (SURE)

- The number of non-zero coefficients is an unbiased estimate of the degrees of freedom of the lasso
- The unbiased estimator is asymptotically consistent
- Requires no special assumption on the predictors
- **Reference:** Zou, H., Hastie, T. and Tibshirani, R. (2007), 'On the "degrees of freedom" of the lasso', Annals of Statistics 35(5), 2173–2192

- We can generalize ridge regression and the lasso, and view them as Bayes estimates
- Ridge: $\hat{\beta}_R$ is the posterior mean, with a Normal prior on β
- Lasso: $\hat{\beta}_L$ is the posterior mode, with a Laplace prior on β