MATH5470 Statistical Machine Learning: Homework 1

Cyril de Lavergne (20109216)

March 2020

Question 1

Ridge Regression

Let's build on what we know, which is that whenever the $n \times p$ model matrix is X, the response n-vector is y, and the parameter p-vector is β, the objective function.

$$f(\beta) = (y - X\beta)'(y - X\beta)$$

(which is the sum of squares of residuals) is minimized when β solves the Normal equations

$$(X'X)\beta = X'y.$$

Ridge regression adds another term to the objective function (usually after standardizing all variables in order to put them on a common footing), asking to minimize

$$(y - X\beta)'(y - X\beta) + \lambda\beta'\beta$$

for some non-negative constant λ. It is the sum of squares of the residuals plus a multiple of the sum of squares of the coefficients themselves (making it obvious that it has a global minimum). Because $\lambda \geq 0$, it has a positive square root $\nu^2 = \lambda$.

Consider the matrix X augmented with rows corresponding to ν times the $p \times p$ identity matrix I:

$$X_* = \begin{pmatrix} X \\ \nu I \end{pmatrix} \tag{1}$$

When the vector y is similarly extended with p zeros at the end to y_*, the matrix product in the objective function adds p additional terms of the form $(0 - \nu\beta_i)^2 = \lambda\beta_i^2$ to the original objective. Therefore

$$(y_* - X_*\beta)'(y_* - X_*\beta) = (y - X\beta)'(y - X\beta) + \lambda\beta'\beta.$$

From the form of the left hand expression it is immediate that the Normal equations are

$$(X'_*X_*)\beta = X'_*y_*.$$

Because we adjoined zeros to the end of y, the right hand side is the same as $X'y$. On the left hand side $\nu^2 I = \lambda I$ is added to the original $X'X$. Therefore the new Normal equations simplify to

$$(X'X + \lambda I)\beta = X'y.$$

Lasso Regression

This can be attacked in a number of ways, including fairly economical approaches via the Karush Tucker conditions.

Below is a quite elementary alternative argument.

The least squares solution for an orthogonal design

Suppose X is composed of orthogonal columns. Then, the least-squares solution is

$$\hat{\beta}^{\mathrm{LS}} = (X^T X)^{-1} X^T y = X^T y.$$

Some equivalent problems

Via the Lagrangian form, it is straightforward to see that an equivalent problem to that considered in the question is

$$\min_{\beta} \frac{1}{2}\|y - X\beta\|_2^2 + \gamma\|\beta\|_1.$$

Expanding out the first term we get $\frac{1}{2}y^T y - y^T X\beta + \frac{1}{2}\beta^T\beta$ and since $y^T y$ does not contain any of the variables of interest, we can discard it and consider yet another equivalent problem,

$$\min_{\beta}(-y^T X\beta + \frac{1}{2}\|\beta\|^2) + \gamma\|\beta\|_1.$$

Noting that $\hat{\beta}^{\mathrm{LS}} = X^T y$, the previous problem can be rewritten as

$$\min_{\beta} \sum_{i=1}^{p} -\hat{\beta}_i^{\mathrm{LS}}\beta_i + \frac{1}{2}\beta_i^2 + \gamma|\beta_i|.$$

Our objective function is now a sum of objectives, each corresponding to a separate variable β_i, so they may each be solved individually.

The whole is equal to the sum of its parts

Fix a certain i. Then, we want to minimize

$$\mathcal{L}_i = -\hat{\beta}_i^{\mathrm{LS}}\beta_i + \frac{1}{2}\beta_i^2 + \gamma|\beta_i|.$$

If $\hat{\beta}_i^{\text{LS}} > 0$, then we must have $\beta_i \geq 0$ since otherwise we could flip its sign and get a lower value for the objective function. Likewise if $\hat{\beta}_i^{\text{LS}} < 0$, then we must choose $\beta_i \leq 0$.

Case 1: $\hat{\beta}_i^{\text{LS}}0$. Since $\beta_i \geq 0$,

$$\mathcal{L}_i = -\hat{\beta}_i^{\text{LS}}\beta_i + \frac{1}{2}\beta_i^2 + \gamma\beta_i,$$

and differentiating this with respect to β_i and setting equal to zero, we get $\beta_i = \hat{\beta}_i^{\text{LS}} - \gamma$ and this is only feasible if the right-hand side is nonnegative, so in this case the actual solution is

$$\hat{\beta}_i^{\text{lasso}} = (\hat{\beta}_i^{\text{LS}} - \gamma)^+ = \text{sgn}(\hat{\beta}_i^{\text{LS}})(|\hat{\beta}_i^{\text{LS}}| - \gamma)^+.$$

Case 2: $\hat{\beta}_i^{\text{LS}} \leq 0$. This implies we must have $\beta_i \leq 0$ and so

$$\mathcal{L}_i = -\hat{\beta}_i^{\text{LS}}\beta_i + \frac{1}{2}\beta_i^2 - \gamma\beta_i.$$

Differentiating with respect to β_i and setting equal to zero, we get $\beta_i = \hat{\beta}_i^{\text{LS}} + \gamma = \text{sgn}(\hat{\beta}_i^{\text{LS}})(|\hat{\beta}_i^{\text{LS}}| - \gamma)$. But, again, to ensure this is feasible, we need $\beta_i \leq 0$, which is achieved by taking

$$\hat{\beta}_i^{\text{lasso}} = \text{sgn}(\hat{\beta}_i^{\text{LS}})(|\hat{\beta}_i^{\text{LS}}| - \gamma)^+.$$

In both cases, we get the desired form, and so we are done.

Final remarks

Note that as γ increases, then each of the $|\hat{\beta}_i^{\text{lasso}}|$ necessarily decreases, hence so does $\|\hat{\beta}^{\text{lasso}}\|_1$. When $\gamma = 0$, we recover the OLS solutions, and, for $\gamma \max_i |\hat{\beta}_i^{\text{LS}}|$, we obtain $\hat{\beta}_i^{\text{lasso}} = 0$ for all i

Question 2: Prostate Cancer data

```
> cor(Prostate)
           lcavol      lweight       age         lbph         svi          lcp      gleason      pgg45      lpsa
lcavol  1.0000000  0.194128286 0.2249999  0.027349703  0.53884500  0.675310484  0.432417056 0.43365225 0.7344603
lweight 0.1941283  1.000000000 0.3075286  0.434934636  0.10877851  0.100237795 -0.001275658 0.05084682 0.3541204
age     0.2249999  0.307528614 1.0000000  0.350185896  0.11765804  0.127667752  0.268891599 0.27611245 0.1695928
lbph    0.0273497  0.434934636 0.3501859  1.000000000 -0.08584324 -0.006999431  0.077820447 0.07846002 0.1798094
svi     0.5388450  0.108778505 0.1176580 -0.085843238  1.00000000  0.673111185  0.320412221 0.45764762 0.5662182
lcp     0.6753105  0.100237795 0.1276678 -0.006999431  0.67311118  1.000000000  0.514830063 0.63152825 0.5488132
gleason 0.4324171 -0.001275658 0.2688916  0.077820447  0.32041222  0.514830063  1.000000000 0.75190451 0.3689868
pgg45   0.4336522  0.050846821 0.2761124  0.078460018  0.45764762  0.631528245  0.751904512 1.00000000 0.4223159
lpsa    0.7344603  0.354120390 0.1695928  0.179809410  0.56621822  0.548813169  0.368986803 0.42231586 1.0000000
```

Figure 1: Correlation

```
Call:
lm(formula = lpsa ~ ., data = Prostate)

Residuals:
     Min       1Q   Median       3Q      Max
-1.73316 -0.37133 -0.01702  0.41414  1.63811

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.669399   1.296381   0.516  0.60690
lcavol       0.587023   0.087920   6.677 2.11e-09 ***
lweight      0.454461   0.170012   2.673  0.00896 **
age         -0.019637   0.011173  -1.758  0.08229 .
lbph         0.107054   0.058449   1.832  0.07040 .
svi          0.766156   0.244309   3.136  0.00233 **
lcp         -0.105474   0.091013  -1.159  0.24964
gleason      0.045136   0.157464   0.287  0.77506
pgg45        0.004525   0.004421   1.024  0.30885
---
```

Figure 2: Linear Model regression

Question 3: What is multi linearity and how about its influence?

Multicollinearity generally occurs when there are high correlations between two or more predictor variables. In other words, one predictor variable can be used to predict the other. This creates redundant information, skewing the results in a regression model.

We have seen in our prostate cancer dataset for subset regression that certain features are dropped due to multicollinearity.

Question 4

1. Lasso Tibshirani

Lasso is similar to shrinking variables towards zero hence applying feature selection and give interpretable results. The paper approaches different ways to select λ the only parameter in the model. Generalized cross validation, cross validation and analytical unbiased estimate. Several experiments on small numbers of large effects, small to to moderate number of moderated sized effects and large numbers of small effects.

It would be interesting to simulate bayesian posterior distribution with RStan (https://mc-stan.org/users/interfaces/rstan).

	LASSO	LM	PCR	PLS	RIDGE	SUBSET
mean	-0.784710	-0.804869	-0.832168	-0.783552	-0.784625	-0.679310
std	0.542782	0.894232	0.575198	0.537526	0.546569	0.551321

	LASSO	LM	PCR	PLS	RIDGE	SUBSET
Intercept	2.47839	2.47839	2.47839	[2.478386878805867]	2.47839	2.47839
lcavol	0.688303	0.688304	0.560421	[0.6918203198270823]	0.681	0.794422
lweight	0.224532	0.224533	0.242393	[0.22467832512678135]	0.224223	0.252439
age	-0.145445	-0.145446	-0.111775	[-0.1382505382411826]	-0.142757	NaN
lbph	0.154512	0.154512	0.142471	[0.15093294927132833]	0.153126	NaN
svi	0.315545	0.315545	0.298958	[0.3225949063569298]	0.312904	NaN
lcp	-0.146714	-0.146716	0.0546817	[-0.14732096782852105]	-0.138338	NaN
gleason	0.0324255	0.0324258	0.204531	[0.04754343472714366]	0.0336278	NaN
pgg45	0.126972	0.126973	-0.125177	[0.10628951871166427]	0.123923	NaN

Figure 3: Different models and metrics

2. Elastic Nets Zou Hastie

Elastic net according to the paper outperforms Lasso while having great interpretability properties and also have grouping effects. The elastic net is particularly useful when the number of predictors (p) is much bigger than the number of observations (n). By contrast, the lasso is not a very satisfactory variable selection method in the p ¿ n case. Prostate cancer, Microarray classification and simulations experiments are held. Authors compare elastic nets to boosting.

3. Adaptive lasso Zou 2006

Authors offers an adaptive laso where adaptive weights are used for penalizing different coefficients in the L1 penalty. It appears to be near-minimax optimal. It outperforms lasso as sometimes lasso can be inconsistent while having equal computational advantage as Lasso.

4. Group lasso Yuan 2006

Group lasso grouped certain features all together to prevent multicollinearity and provide superior performance to the traditional stepwise backward elimination method in factor selection problems. AUthors provide Cp-criterion for parameter selection which is generally comparable to cross-validation and appears to be better sometimes. It is important to acknowledge that the solution path is not piecewise linear. Authors claim that the algorithm perform better that the non-negative garrotte.

5. Fused lasso Tibshirani

The fused lasso can be computationally expensive however it appears to be a promising method for both regression and classification for ordered dataset. The method is indeed sparse and has asymptotic properties hence it is better to use fused lasso when p ¿ N. The solution path of the method is piecewise linear.

It would be interesting to see fusing a diverse set of variables with fused lasso to obtain better computational speed.