Lec 11: Regularized Learning

Ailin Zhang

2022-10-21

Roadmap for Regularized Learning

- Ridge regression
- Lasso regression
- Coordinate descent
- Spline regression
- Least angle regression
- Stagewise regression / epsilon learning
- Bayesian regression
- Perceptron
- SVM
- Adaboost

Note: We will have midterm after regularized learning! (Est. Nov.7 - 11)

Stagewise Regression

The stagewise regression iterates the following steps:

1. Start with $\mathbf{R} = \mathbf{Y}$, $\beta_1, \beta_2, \cdots, \beta_p = 0$.
2. Find the predictor $\mathbf{X}_j$ most correlated with R: ind j with the maximal $|\langle \mathbf{R}, \mathbf{X}_j \rangle|$.
3. Then update $\beta_j \leftarrow \beta_j + \epsilon \langle \mathbf{R}, \mathbf{X}_j \rangle$
4. Set $\mathbf{R} = \mathbf{Y} - \sum_{j=1}^{p} \mathbf{X}_j \beta_j$, or $\mathbf{R} = \mathbf{R} - \epsilon \langle \mathbf{R}, \mathbf{X}_j \rangle \cdot \mathbf{X}_j$.

Repeat step 2-4

This is similar to the matching pursuit but is much less greedy. Such an update will change $\mathbf{R}$ and reduce $|\langle \mathbf{R}, \mathbf{X}_j \rangle|$, until another $\mathbf{X}_j$ catches up.

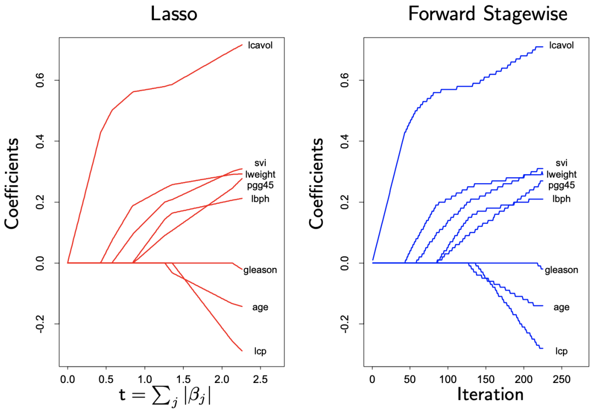
So overall, the algorithm ensures that all of the selected $\mathbf{X}_j$ to have the same $|\langle \mathbf{R}, \mathbf{X}_j \rangle|$.

The stagewise regression is also called ϵ-boosting.

R code for Stagewise Regression

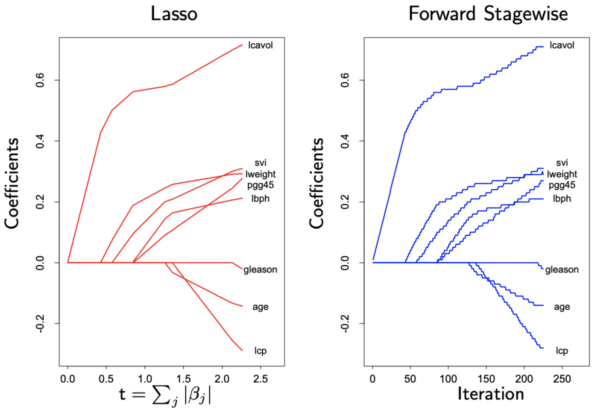
```
T = 3000
epsilon = .0001
beta = matrix(rep(0, p), nrow = p)
db = matrix(rep(0, p), nrow = p)
beta_all = matrix(rep(0, p*T), nrow = p)

R = Y
for (t in 1:T)
    {
      for (j in 1:p)
        db[j] = sum(R*X[, j])
      j = which.max(abs(db))
      beta[j] = beta[j]+db[j]*epsilon
      R = R - X[, j]*db[j]*epsilon
      beta_all[, t] = beta
}
matplot(t(matrix(rep(1, p), nrow = 1)%*%abs(beta_all)), t(beta_all), type = 'l')
```

Stagewise Regression vs Lasso Regression



Forward Stagewise and Lasso look similar. Are they Identical?

Stagewise Regression vs Lasso Regression



Forward Stagewise and Lasso look similar. Are they Identical?

- If X is orthogonal: yes
- A more general case: almost identical, not exactly same.

Relationship among Lasso, LAR, and stagewise regression

- LAR: uses least squares directions in the active set of variables
- LASSO: uses least square directions; if a variable crosses zero, it is removed from the active set.
- Forward stagewise: Move in the direction of maximum $Corr(\mathbf{R}, \mathbf{X}_j)$ in the active set.

Relationship among Lasso, LAR, and stagewise regression

- LAR: uses least squares directions in the active set of variables
- LASSO: uses least square directions; if a variable crosses zero, it is removed from the active set.
- Forward stagewise: Move in the direction of maximum $Corr(\mathbf{R}, \mathbf{X}_j)$ in the active set.

In forward stagewise, if $\epsilon \to 0$, it converges to LAR.

Stepwise Regression

- Stepwise regression is a **variable selection** procedure for independent variables (**X**)
- Consists of a series of steps designed to find the most features to include in a regression model
- Basis for selection:
  - Choose a variable that satisfies the criterion
  - Remove a variable that least satisfies the criterion

Stepwise Regression: Example

At each step, we either enter or remove a predictor based on the partial F-tests — the t-tests for the slope parameters.

We stop when no more predictors can be justifiably entered or removed from our stepwise model, thereby leading us to a “final model.”

Stepwise Regression: Example

Regress y on x_1, y on x_2, y on x_3, y on x_4. Choose significance level as 0.15.

Predictor	Coef	SE Coef	T	P
Constant	81.479	4.927	16.54	0.000
x1	1.8687	0.5264	**3.55**	**0.005**

Predictor	Coef	SE Coef	T	P
Constant	57.424	8.491	6.76	0.000
x2	0.7891	0.1684	**4.69**	**0.001**

Predictor	Coef	SE Coef	T	P
Constant	110.203	7.948	13.87	0.000
x3	-1.2558	0.5984	**-2.10**	**0.060**

Predictor	Coef	SE Coef	T	P
Constant	117.568	5.262	22.34	0.000
x4	-0.7382	0.1546	**-4.77**	**0.001**

Stepwise Regression: Example

Regress y on x_1, y on x_2, y on x_3, y on x_4. Choose significance level as 0.15.

Predictor	Coef	SE Coef	T	P
Constant	81.479	4.927	16.54	0.000
x1	1.8687	0.5264	**3.55**	**0.005**

Predictor	Coef	SE Coef	T	P
Constant	57.424	8.491	6.76	0.000
x2	0.7891	0.1684	**4.69**	**0.001**

Predictor	Coef	SE Coef	T	P
Constant	110.203	7.948	13.87	0.000
x3	-1.2558	0.5984	**-2.10**	**0.060**

Predictor	Coef	SE Coef	T	P
Constant	117.568	5.262	22.34	0.000
x4	-0.7382	0.1546	**-4.77**	**0.001**

As a result of the first step, we enter x_4 into our stepwise model.

Stepwise Regression: Example

we fit the next two-predictor model that includes x_4 as a predictor — that is, we regress y on x_4 and x_1, y on x_4 and x_2 , and y on x_4 and x_3

Predictor	Coef	SE Coef	T	P
Constant	103.097	2.124	48.54	0.000
x4	-0.61395	0.04864	-12.62	0.000
x1	1.4400	0.1384	**10.40**	**0.000**

Predictor	Coef	SE Coef	T	P
Constant	94.16	56.63	1.66	0.127
x4	-0.4569	0.6960	-0.66	0.526
x2	0.3109	0.7486	**0.42**	**0.687**

Predictor	Coef	SE Coef	T	P
Constant	131.282	3.275	40.09	0.000
x4	-0.72460	0.07233	-10.02	0.000
x3	-1.1999	0.1890	**-6.35**	**0.000**

Stepwise Regression: Example

we fit the next two-predictor model that includes x_4 as a predictor — that is, we regress y on x_4 and x_1, y on x_4 and x_2 , and y on x_4 and x_3

Predictor	Coef	SE Coef	T	P
Constant	103.097	2.124	48.54	0.000
x4	-0.61395	0.04864	-12.62	0.000
x1	1.4400	0.1384	**10.40**	**0.000**

Predictor	Coef	SE Coef	T	P
Constant	94.16	56.63	1.66	0.127
x4	-0.4569	0.6960	-0.66	0.526
x2	0.3109	0.7486	**0.42**	**0.687**

Predictor	Coef	SE Coef	T	P
Constant	131.282	3.275	40.09	0.000
x4	-0.72460	0.07233	-10.02	0.000
x3	-1.1999	0.1890	**-6.35**	**0.000**

As a result of the second step, we enter x_1 into our stepwise model.

Stepwise Regression: Example

Regress y on x_4, x_1,and x_2, and y on x_4, x_1 and x_3

Predictor	Coef	SE Coef	T	P
Constant	71.65	14.14	5.07	0.001
x4	-0.2365	0.1733	-1.37	0.205
x1	1.4519	0.1170	12.41	0.000
x2	0.4161	0.1856	**2.24**	**0.052**

Predictor	Coef	SE Coef	T	P
Constant	111.684	4.562	24.48	0.000
x4	-0.64280	0.04454	-14.43	0.000
x1	1.0519	0.2237	4.70	0.001
x3	-0.4100	0.1992	**-2.06**	**0.070**

Stepwise Regression: Example

Regress y on x_4, x_1,and x_2, and y on x_4, x_1 and x_3

Predictor	Coef	SE Coef	T	P
Constant	71.65	14.14	5.07	0.001
x4	-0.2365	0.1733	-1.37	0.205
x1	1.4519	0.1170	12.41	0.000
x2	0.4161	0.1856	**2.24**	**0.052**

Predictor	Coef	SE Coef	T	P
Constant	111.684	4.562	24.48	0.000
x4	-0.64280	0.04454	-14.43	0.000
x1	1.0519	0.2237	4.70	0.001
x3	-0.4100	0.1992	**-2.06**	**0.070**

As a result of the third step, we enter x_1 into our stepwise.

At the same time, remove x_4.

Stepwise Regression: Example

Proceed fitting each of the three-predictor models that include x_1 and x_2 as predictors — that is, we regress y on x_1, x_2, and x_3; y on x_1, x_2, and x_4:

Predictor	Coef	SE Coef	T	P
Constant	71.65	14.14	5.07	0.001
x1	1.4519	0.1170	12.41	0.000
x2	0.4161	0.1856	2.24	0.052
x4	-0.2365	0.1733	**-1.37**	**0.205**

Predictor	Coef	SE Coef	T	P
Constant	48.194	3.913	12.32	0.000
x1	1.6959	0.2046	8.29	0.000
x2	0.65691	0.04423	14.85	0.000
x3	0.2500	0.1847	**1.35**	**0.209**

Stepwise Regression: Example

Proceed fitting each of the three-predictor models that include x_1 and x_2 as predictors — that is, we regress y on x_1, x_2, and x_3; y on x_1, x_2, and x_4:

Predictor	Coef	SE Coef	T	P
Constant	71.65	14.14	5.07	0.001
x1	1.4519	0.1170	12.41	0.000
x2	0.4161	0.1856	2.24	0.052
x4	-0.2365	0.1733	**-1.37**	**0.205**

Predictor	Coef	SE Coef	T	P
Constant	48.194	3.913	12.32	0.000
x1	1.6959	0.2046	8.29	0.000
x2	0.65691	0.04423	14.85	0.000
x3	0.2500	0.1847	**1.35**	**0.209**

We stop our stepwise regression procedure. Our final regression model, based on the stepwise procedure contains only the predictors x_1 and x_2.

Stepwise Regression: Example

Final model:

Predictor	Coef	SE Coef	T	P
Constant	52.577	2.286	23.00	0.000
x1	1.4683	0.1213	12.10	0.000
x2	0.66225	0.04585	14.44	0.000

Not only can you use t-test, you can also consider R^2, AIC, BIC,etc. . .

Summary for Stepwise Regression

1. The final model is not guaranteed to be optimal in any specified sense.
2. The procedure yields a single final model, although there are often several equally good models.
3. Stepwise regression does not take into account domain knowledge about the predictors. It may be necessary to force the procedure to include important predictors.
4. One should not over-interpret the order in which predictors are entered into the model.
5. It is possible that we may have committed a Type I or Type II error along the way.

Bayesian Regression

Let $\beta \sim \mathrm{N}(0, \tau^2 \mathbf{I}_p)$ be the prior distribution of β. The joint log probability density of β and $\mathbf{Y}$ is

$$-\frac{1}{2\sigma^2}\|\mathbf{Y} - \mathbf{X}\beta\|_{\ell_2}^2 - \frac{1}{2\tau^2}\|\beta\|_{\ell_2}^2,$$

up to an additive constant.

The above function is quadratic in β. By setting the first derivative to 0, we get the mode of β,

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X}/\sigma^2 + \mathbf{I}_p/\tau^2)^{-1}\mathbf{X}^\top \mathbf{Y}/\sigma^2.$$

Bayesian Regression

Let $\beta \sim \mathrm{N}(0, \tau^2 \mathbf{I}_p)$ be the prior distribution of β. The joint log probability density of β and $\mathbf{Y}$ is

$$-\frac{1}{2\sigma^2}\|\mathbf{Y} - \mathbf{X}\beta\|_{\ell_2}^2 - \frac{1}{2\tau^2}\|\beta\|_{\ell_2}^2,$$

up to an additive constant.

The above function is quadratic in β. By setting the first derivative to 0, we get the mode of β,

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X}/\sigma^2 + \mathbf{I}_p/\tau^2)^{-1}\mathbf{X}^\top \mathbf{Y}/\sigma^2.$$

which corresponds to the ridge regression with $\lambda = \sigma^2/\tau^2$.

Bayesian Regression

The joint log probability density of β and $\mathbf{Y}$ is

$$-\frac{1}{2\sigma^2}\|\mathbf{Y}-\mathbf{X}\beta\|_{\ell_2}^2-\frac{1}{2\tau^2}\|\beta\|_{\ell_2}^2,$$

The second derivative or the Hessian matrix is $H=\mathbf{X}^\top\mathbf{X}/\sigma^2+\mathbf{I}_p/\tau^2$. The inverse is the variance-covariance matrix $V=H^{-1}$. So the posterior distribution of β given $\mathbf{X}$ and $\mathbf{Y}$ is

$$[\beta|\mathbf{X},\mathbf{Y}]\sim \mathrm{N}\left(\hat{\beta},V\right).$$

Both $\hat{\beta}$ and V can be obtained by the sweep operator, very much like the original linear regression.

Bayesian Regression

Prior: $\beta \sim \mathit{Laplace}(\gamma)$

$p(\beta) = (\frac{\gamma}{2})^p \exp(-\gamma||\beta||_1)$

Bayesian Regression

Prior: $\beta \sim \frac{I_p}{\sigma^2}$