itsl-601

Best Subset Selection

To perform best subset selection, we fit {a seperate least squares regression for each possible combination of the p predictors}.

The procedure is (where $\mathcal{M}_0$ is the null model with no predictors):

```
1. For $k = 1, \ldots, p$: {
    (a) Fit all $\binom{p}{k}$ models that contain exactly $k$ predictors;
    (b) Pick the "best" model, $\mathcal{M}_k$, such that e.g. RSS is minimised, or $R^2$ maximised.
   }
2. {Select the single best model from $\mathcal{M}_0, \ldots, \mathcal{M}_p$, using e.g. the prediction error on a validation set, adjusted $R^2$, or cross validation.}
```

Back:

Best Subset Selection

Best subset selection suffers from {computational limitations}. In general, there are {2^p models that involve subsets of p predictors}.

Back:

Forward Stepwise Selection

Forward stepwise selection begins with {a model containing no predictors, and then adds predictors to the model, one-at-a-time, until all of the predictors are in the model}. At each step {the variable that gives the greatest additional improvement to the fit is added to the model}.

Back:

itsl-604

Forward Stepwise Selection

Forward stepwise selection involves fitting a total of $\{\sum_{i=0}^{p-1}(p-k)=1+\frac{p(p+1)}{2}\}$ models.

Back:

Forward Stepwise Selection

Forward stepwise selection can be applied even in the scenario {$n < p$, although here only the first n stepwise models, $\mathcal{M}_0, \ldots, \mathcal{M}_{n-1}$, can be found}. Beyond this, {least squares does not give a unique solution}.

Back:

itsl-606

Adjusting the training error

Four approaches to selecting among a set of models with different number of variables are: {C_p, Akaike information criterion (AIC), Bayesian information criterion (BIC), and adjusted R^2}.

Back:

C_p estimate of test MSE

For a fitted least squares model containing d predictors, the C_p estimate of test MSE is given by, {

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2),$$

} where {$\hat{\sigma}^2$ is an estimate of the variance of the error ε associated with each response measurement}. Typically, {$\hat{\sigma}^2$ is estimated using the full model containing all predictors}.

Back:

C_p estimate of test MSE

Essentially, the C_p statistic {adds a penalty of $2d\hat{\sigma}^2$ to the training RSS in order to adjust for the fact that the training error tends to underestimate the test error}.

Back:

Akaike Information Criterion

For a standard multiple regression model, least squares and maximum likelihood are the same. In this case AIC is given by {

$$\text{AIC} \propto \frac{1}{n}(RSS + 2d\hat{\sigma}^2).$$

} Hence, {for least squares models, C_p and AIC are proportional to each other}.

Back:

Bayesian Information Criterion

For a least squares model with d predictors, the BIC is given by {

$$\text{BIC} \propto \frac{1}{n}(\text{RSS} + \log(n)d\hat{\sigma}^2).$$

}

Back:

Adjusted-R^2

For a least squares model with d predictors, the adjusted R^2 statistic is given by {

$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}.$$

} A {large value } of adjusted R^2 indicates a model with {small test error }.

Back:

Adjusted-R^2

The intuition behind the adjusted R^2 is that once all of the correct variables have been included in the model, adding {additional noise variables will lead to only a very small decrease in RSS}. Since adding {noise variables leads to an increase in d, such variables will lead to an increase in $RSS/(n-d-1)$, and consequently a decrease in the adjusted R^2}.

Back:

itsl-613

Validation and cross-validation

Using validation and cross-validation procedures to estimate test error has an advantage relative to AIC, BIC, C_p, and adjusted R^2, in that {it provides a direct estimate of the test error, and makes fewer assumptions about the true underlying model}. It can also be used in a wider range of model selection tasks, even in cases where {it is hard to pinpoint the model degrees of freedom (*e.g.* the number of predictors in the model) or hard to estimate the error variance σ^2}.

Back:

Linear Model Selection –

itsl-614

The one-standard-error rule

When choosing between models with different degrees of freedoms, the one-standard-error rule can be applied, where the {simplest model out of “equally good” models is chosen}. We first calculate {the one-standard error of the estimated test MSE for each model size, and then select the smallest model for which the estimated test error is within one standard error of the lowest point on the curve}.

Back:

itsl-615

Ridge Regression

In linear models, ridge regression coefficients β^R are chosen to minimise, {

$$\text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2,$$

} where {$\lambda \geq 0$ is some tuning parameter}. Note that the second term, the {shrinkage penalyty}, is only {applied to $\beta_1, \ldots, \beta_p$, not the intercept β_0}.

Back:

Linear Model Least Squares

The standard linear model least squares coefficient estimates are {scale equivariant}: multiplying X_j by a constant c leads to {a scaling of the least squares coefficient estimates by a factor of $1/c$, such that $X_j\hat{\beta}_j$ will remain the same}.

Back:

itsl-617

Ridge Regression

Ridge regression coefficient estimates can change sub- stantially when {multiplying a given predictor by a constant (*i.e.* they are not scale equivariant), and therefore it is best to apply ridge regression after standardising the predictors}, {using the formula,

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}}.$$

}

Back:

itsl-618

Ridge Regression Advantages

Ridge regression's advantage over least squares is rooted in the {bias-variance trade-off}. As {λ increases, the flexibility of the ridge regression fit decreases, leading to decreased variance but increased bias}.
Ridge regression works best in scenarios {where the least squares solution has high variance, *e.g.* if $p \approx n$}. Ridge regression can also perform in the {$p > n$ scenario, even though least squares does not have a unique solution}.

Back:

Shinkage Methods –

Ridge Regression Advantages

Ridge regression has substantial computational advantages {over best subset selection. For any fixed value of λ, ridge regression only fits a single model, and the model-fitting procedure can be performed quite quickly}.
The computations required {for solving,

$$\mathrm{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2,$$

simultaneously for all values of λ, are almost identical to those for fitting a model using least squares}.

Back:

itsl-620

The Lasso

The lasso coefficients, $\hat{\beta}_\lambda^L$, minimises the quantity, {

$$\mathrm{RSS} + \lambda \sum_{j=1}^{p} |\beta_j| = \sum_{i=1}^{n} \Big(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \Big)^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

}

Back:

itsl-621

Ridge vs. Lasso

The only difference {between the quantity to minimise in ridge and lasso is the penalty term – ridge uses β_j^2 and lasso $|\beta_j|$}. That is, ridge regression uses an {l_2 penatly term, and lasso uses l_1}.

Back:

The Lasso

The l_1 penalty has the effect of {forcing some coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large. Hence, lasso performs variable selection}.
The lasso yields {sparse models which are often much easier to interprest than those produced using ridge regression}.

Back:

itsl-623

Another formulation for Ridge Regression and the Lasso

The lasso and ridge regression coefficients estimates solve the problems, {

$$\min_{\beta}\left(\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right) \quad \text{s.t.} \quad \sum_{j=1}^{p}|\beta_j|\le s$$

} and, {

$$\min_{\beta}\left(\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right) \quad \text{s.t.} \quad \sum_{j=1}^{p}\beta_j^2\le s,$$

} respectively.

Back:

itsl-624

The Lasso

Consider the formulation

$$\min_{\beta}(\text{RSS}) \quad \text{s.t.} \quad \sum_{j=1}^{p} |\beta_j| \leq s.$$

When $p = 2$, {this indicates that the lasso coefficient estimates have the smallest RSS out of all points that lie within the diamond defined by $|\beta_1| + |\beta_2| \leq s$}.
Therefore, the lasso regression coefficient estimates are given by the {first point at which an ellipse centered around $\hat{\beta}$ contacts the constraint region}. As the lasso constraint has {corners at each of the axes, therefore the ellipse will often intersect the constraint region at an axis (where some of the coefficients will equal zero)}.

Back: