Supervised Learning - Selection and Regularization for Regression Models

Mark Asch - IMU/VLP/CSU

2023

Intro to ML

Program

1. Data Analysis
   (a) Introduction: the 4 identifiers of “big data” and “data science”
   (b) Supervised learning methods: regression—advanced, k-NN, linear classification methods, SVM, NN, decision trees.
   (c) Unsupervised learning methods: k-means, principal component analysis, clustering.

Limits of the Regression Model

- Recall the standard linear model for a regression,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

  that describes the relationship between the response Y and the p explanatory variables $X_1, \ldots, X_p$ and is fit by least-squares

- Properties:

  $\Rightarrow$ easy to interpret
  $\Rightarrow$ inference is possible and rigorous
  $\Rightarrow$ can take into account non-linearities (on condition to know, or to guess, them...)
  $\Rightarrow$ robustness for real problems

- Before passing to nonlinear models (k-nn, SVM, trees, NN, etc.), is it possible to improve the linear model?

- Why should we use other fitting methods based on least-squares?

- ⇒ Forecast Precision
  - → if the true relation is approximately linear, then the bias of SLR will be small
  - → if $n \gg p$, then the variance of SLR will be small and we obtain good predictive performance on (unseen) test data
  - → if $n \approx p$, then SLR will have a tendency to overfit and the variance will be high and give bad predictions
  - → if $p > n$, then the variance is infinite and SLR cannot be used...
  - → by constraining or shrinking the estimated coefficients, we can reduce the variance without too much increase in the bias—this produces a clear improvement of the predictive precision
- ⇒ Model Interpretation
  - → often, in multiple regression, several explanatory variables are not associated with the response
  - → including such non-pertinent variables leads to complex models—see bias-variance tradeoff
  - → the selection of attributes/variables can eliminate these nuisance variables...

Three classes of methods

1. Selection of subsets: we identify a subset of the p predictors and we apply least-squares to this reduced set

2. Regularization/penalization/shrinking: we fit on all p variables, but we shrink the coefficients towards zero, thus reducing the variance (in the limit, we perform attribute selection...). The two common approaches are:

   - ridge regression
   - LASSO regression

3. Reduction of dimension: we project the p predictors onto a subspace of dimension M with $M < p$. The two common approaches are:

   - PCR—principal component regression
   - PLS—partial least-squares

Subset Selection

- Recall:

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

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

- we fit a SLR for each possible combination of p predictors

  ⇒ p models with a single predictor
  ⇒ $\binom{p}{2} = p(p-1)/2$ models with two predictors, etc., etc.

- we then choose the best model (see criteria below)

- Algorithm:

```
Let $M_0$ l be the null model,  without predictors
for $k = 1, 2, \ldots, p$
  fit all  $\binom{p}{k}$ models of $k$ predictors
  choose the best model $M_k$ of minimal RSS
     or maximal $R^2$
next $k$
Select the best model among $M_0, \ldots, M_p$
  by cross-validation,  AIC,  BIC or adjusted $R^2$
```

- when p is large, the method of subset selection can be too expensive

- we can proceed stepwise, by adding (forward) or removing (backward) predictors one-by-one (stepwise selection)

Criteria of Choice

- the selection methods produce an ensemble of models

$\Rightarrow$ each model contains a subset of p explanatory variables (predictors)

$\Rightarrow$ which model is the best???

$\Rightarrow$ the model that contains all the predictors will always have the smallest value of RSS and the greatest value of $R^2 = 1 - \mathrm{RSS}/\mathrm{TSS}$

→ Conclusion: RSS and R^2 are **NOT** good criteria for choosing a model among those with a different number of predictors

- We must estimate the test error in order to best select among the models:

1. Indirect estimation by refitting on the training error
2. Direct estimation by a validation set or cross-validation (see below)

Criteria for Refitting

There are four criteria that can be used for selecting a model among models with different numbers of variables:

1. C_p

2. AIC (Akaike Information Criterion)

3. BIC (Bayesian Information Criterion)

4. Adjusted R^2

Criterion C_p

- for a model fitted by least-squares with d explanatory variables (predictors), define

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

  where $\hat{\sigma}^2$ is an estimation of the variance of the error ϵ associated to each measurement/observation in the multilinear regression formula

- we estimate $\hat{\sigma}^2$ using the complete model with all the explanatory variables

- the C_p statistic adds a penalization of $2d\hat{\sigma}^2$ to the training RSS in order to compensate for the fact that the training error always underestimates the test error.

- we can show that C_p diminishes for models with a small value of the test error

- Conclusion: we choose the model with the minimal C_p value

Criterion AIC

- defined for a broad class of models fitted by a maximum likelihood (ML) method

- the Akaike Information Criterion is defined as

$$\mathrm{AIC} = \frac{1}{n\hat{\sigma}^2}\left(\mathrm{RSS} + 2d\hat{\sigma}^2\right)$$

- in the case of SLR with Gaussian errors, ML and LS are identical!

Criterion BIC

- criterion obtained from a Bayesian analysis...

- for a LS model with d predictors, define the Bayesian Information Criterion

$$\mathrm{BIC} = \frac{1}{n\hat{\sigma}^2}\left(\mathrm{RSS} + d\hat{\sigma}^2 \log n\right)$$

- just as for C_p, the BIC will take a small value for a model with a low test error

- but, the factor $\log n$ will penalize models having many variables, and thus select smaller models (less complex ones) than C_p or AIC

- Conclusion: we choose the model with the minimal BIC value

Criterion Adjusted R^2

- this criterion modifies the coefficient of determination R^2, to compensate for the fact that the RSS always diminishes when we add variables, and hence $R^2 = 1 - \mathrm{RSS}/\mathrm{TSS}$ increases

- the definition takes into account both n and d,

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

- here, a large value implies a model with small test error

- adding nuisance variables gives an increase in d and an increase of $\mathrm{RSS}/(n-d-1)$, and thus a reduction of the $\text{Adjusted}\, R^2$

  ⇒ in theory, the model with the largest (best) $\text{Adjusted}\, R^2$ will only contain the good variables and no nuisance variables

Cross Validation

✔ use a validation/test set to directly estimate the test error, without additional hypotheses

✔ applicable in more general contexts...

✘ CPU time can become a practical limit...

Methods of Regularization/Shrinking

- we will now modify directly the least-squares minimization criterion

- Recall: least-squares regression estimates the coefficients $\beta_0, \ldots, \beta_p$ by the minimization of

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

- we obtain 2 alternative methods:

  ⇒ ridge regression
  ⇒ the LASSO (least absolute shrinkage and selection operator)

Ridge Regression (RR)

- Penalized Least-Squares are used to estimate the values of β_j

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

where $\lambda \geq 0$ is a tuning parameter, and the new term introduces a "shrinkage" by reducing the effects of terms whose squared values are small—this term will be small when $\beta_1, \ldots \beta_p$, are close to zero and its effect will be to shrink the estimates of the β_j towards zero

⇒ we use cross validation to estimate λ

⇒ the coefficient of the intercept, β_0, is not shrunk—it measures the average of the response when all the $X_i = 0$

⇒ advantage over SLR: less variance, more bias when λ increases

- influence of λ

- ⇒ when $\lambda = 0$, the ridge gives the least-squares estimates
- ⇒ when $\lambda \to \infty$, the ridge coefficients tend to zero
- ⇒ selecting a good value is critical, and we use cross-validation.

- advantages over SLR:
  - ⇒ bias-variance tradeoff...
  - ⇒ with increasing λ, the flexibility of the fit by ridge decreases, which implies less variance and more bias
  - ⇒ works well when SLR has a high variance, especially when $n \approx p$ and $n < p$.

LASSO Regression

- uses the l_1-norm to penalize the β_j,

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

- can, for λ large enough, cancel the coefficients and thus reduce the dimension of predictors, which facilitates interpretation of the regression obtained—this is also called "feature selection";

$\Rightarrow$ we use cross-validation on a set of values $\{\lambda_1, \ldots, \lambda_m\}$ to estimate λ

- Ridge or LASSO?

$\Rightarrow$ LASSO provides more interpretable models

$\Rightarrow$ LASSO has better performance when the response indeed depends on a subset of features

Cross-Validation for λ

- choose a grid of values for λ
- fix the number of folds for the cross-validation
- for each value of λ compute the CV error
- select the value of λ for which the CV error is minimal
- fit the model again,

  $\Rightarrow$ with all the observations
  $\Rightarrow$ with the optimal value of λ

- Beware of RNG initialization for reproducibility.

Methods of Dimension Reduction

- 2 steps

  ⇒ transform the explanatory variables $X_1, X_2, \ldots, X_p$, then

  ⇒ fit a least-squares model to the transformed variables

- 2 approaches

  ⇒ non-supervised: PCR—principal component regression

  ⇒ supervised: PLS—partial least-squares

Methods of Dimension Reduction II

- let $Z_1, Z_2, \ldots, Z_M$ with $M < p$, be linear combinations of the original p predictors

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j, \quad m = 1, \ldots, M$$

- fit a linear regression model

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i, \quad i = 1, \ldots, n$$

- if the coefficients $\phi_{1m}, \phi_{2m}, \ldots, \phi_{pm}$ are well-chosen, the a dimension reduction approach can attain a better performance than SLR

$\Rightarrow$ the dimension is reduced from $p + 1$ to $M + 1$

$\Rightarrow$ the fit is special case of SLR

$$\sum_{m=1}^{M} \theta_m z_{im} = \sum_{j=1}^{p} \beta_j x_{ij}$$

with

$$\beta_j = \sum_{m=1}^{M} \theta_m \phi_{jm}$$

- the new coefficients are constrained and thus biased, but when $p > n$ and $M \ll p$ the reduction of variance can be consequential

- the 2 steps of any variance reduction method are:

  1. obtain transformed predictors $Z_1, Z_2, \ldots, Z_M$
  2. fit a model to these M predictors

Principal Component Regression (PCR)

- PCA (see below) is an established method for obtaining a low-dimensional set of attributes from a large set of variables

  ⇒ PCA is an unsupervised approach

  ⇒ the first principal component gives the direction in which the observations vary the most, etc.

- strong hypothesis:

  ⇒ the principal components, that are calculated from X, are indeed representative of Y

  ⇒ if yes, they can detect causality by considerably reducing the dimension of the parameter space

- 2 steps:

  1. Calculate the first Mprincipal components
  2. Use these M components in a linear regression model that we fit by least-squares

- Conclusions
  - ⇒ since $M \ll p$, any overfitting is automatically attenuated
  - ⇒ PCR does not perform feature selection—the original p predictors are still there, though in the form of linear combinations
  - ⇒ there is thus a link between PCR and ridge regression...
  - ⇒ the choice of M is made by cross-validation
  - ⇒ it is strongly recommended to normalize the explanatory variables, unless they are of the same units

Partial Least-Squares (PLS)

- This is a supervised alternative to PCR:

  ⇒ find the components of X that are also pertinent for Y

  ⇒ calculate an ensemble of latent vectors that execute simultaneously a decomposition of X and of Y, under the constraint that these components describe as much as possible of the covariance between X and Y

  ⇒ the algorithms are quite complex...

- Steps of the computation:

  ⇒ normalize the p explanatory variables

  ⇒ calculate Z_1 by setting each ϕ_{j1} in

$$Z_1 = \sum_{j=1}^{p} \phi_{j1} X_j$$

equal to the linear regression coefficient of Y on X_j which will place the most weight on the variables having the strongest correlation with the response

⇒ calculate Z_2

  → fit each of the variables for Z_1 by computing the regression on the residues...

⇒ etc.

- Remarks:

  ⇒ PLS is often used in industrial applications where p is big and n is small

  ⇒ PLS is rarely better than LASSO, but does not require any tuning...

Take-Home Lessons

- The linear regression method, in its numerous guises, shows how we quantify uncertainty as best as possible.

- Recall that the methods reduce the known part of the uncertainty since they are optimal estimates, but that the unknown, irreducible part remains.

- Our job is then to inform the decision-maker on how risky this is. For this, we carefully modeled the "noise" in the system, and we proposed five methods for quantifying its effects:

1. Use the R-squared value.

2. Use the p-values resulting from a hypothesis test, based on the t-statistic.

3. Check the normality of the residues.

4. Use cross-validation.

5. Check whether adding variables or transforming variables has an effect on the previous four.

- If we perform all the above, rigorously, then we have fulfilled our responsibilities as modelers, engineers, data scientists, and applied mathematicians.

Examples

1. Ridge Regression and LASSO for baseball data `reg-ridge-lasso.html`

2. PCR and PLS for baseball `reg_PCR_PLS.html`

3. Subset Selection for baseball `reg-subset-sel.html`

References

1. G. James, D. Witten, T. Hastie, R. Tibshirani. *An Introduction to Statistical Learning with Applications in R.* Springer. 2013.

2. T. Hastie, R. Tibshirani, J. Friedman. *The Elements of Statistical Learning*. Springer. 2009.

3. Rachel Schutt and Cathy O'Neil. *Doing Data Science.* O'Reilly. 2014.