STATS 769

Resampling Methods

Yong Wang
Department of Statistics
The University of Auckland

2021

Predictive Performance

- It is often not a problem to build one or many models for a given data set, but how to assess their predictive performance in an objective manner?
- Model selection criteria may be unreliable or even unavailable.
  - E.g., the tuning parameter value for ridge regression.
- In the cases where theory can not help, we would desire to have another independent data set to compute the prediction errors (PE) of these models.
- The data set that is used to build models is often known as a training set, while the other data set that is used to evaluate the models is known as the test set (or validation set).

Splitting Data

- We can, e.g., split all data at hand into two subsets, one for training and the other for testing.
- A conflict:
  - We want the training set to be as large as possible so that each model is better fitted.
  - We want the test set to be as large as possible so that the PEs can be more accurately obtained.
- We really want to use all the data to build the best model for future prediction.

Training *vs.* Test Errors

- One may apply the fitted model to the training set itself.
- The resulting error is known as the training error (or resubstitution error).
  - For regression, this is $\mathrm{MSE}(\widehat{f}; \mathbf{X}, \mathbf{y})$.
- One can not use it (directly) as PE, since it is downwardly biased (i.e., smaller in expectation than the true PE).
- It is the test error that should be used.
  - For regression, this is $\mathrm{MSE}(\widehat{f}; \mathbf{X}', \mathbf{y}')$.

Evaluating a Method

- It is critically important to realise that we are not really assessing how accurate a fitted model is, because of the randomness of data.
- Instead, we would like to assess a method that is used to build a model, or an estimator $\widehat{f}$ of f.
- For example, we may want to decide between the first-order polynomial

$$\widehat{f}^{(1)}(x) = \widehat{\beta}_0^{(1)} + \widehat{\beta}_1^{(1)} x$$

  and the second-order polynomial

$$\widehat{f}^{(2)}(x) = \widehat{\beta}_0^{(2)} + \widehat{\beta}_1^{(2)} x + \widehat{\beta}_2^{(2)} x^2,$$

  where both are fitted by least squares to a given data set.
- Another example is to determine an appropriate value for a tuning parameter λ, or an estimator $\widehat{f}^{(\lambda)}$ which corresponds to a unique λ-value.

Data Resampling

- It is often reasonable to treat observations in a given data set as a random (i.i.d.) sample of some distribution.
- Hence, any random subsample of the observations is also a random sample of the distribution.
- We can thus create training and test sets, using random subsets of the data, for evaluating any estimator $\widehat{f}$.
- Data resampling techniques
  - Jackknifing
  - Leave-one-out cross-validation
  - K-fold cross-validation
  - (Bootstrapping)
- Typically, it is a finite number of estimators that are included in comparison: $\widehat{f}^{(1)}, \ldots, \widehat{f}^{(m)}$.
  - For a continuous tuning parameter λ, one may consider a fine grid of λ-values.

Jackknifing

- Delete-d Jackknifing (for each $\widehat{f}^{(j)}$):
  1. Delete d observations randomly from the entire data set.
  2. Fit the model to the remaining data.
  3. Calculate the PE of the fitted model using the deleted observations.
  4. Repeat the above steps a number of times and compute the mean PE.
- Find j^*, the optimal j that gives the smallest PE.

Purpose of Data Resampling

What is the purpose of using a data resampling method?

- If one is to compare the performance of different methods, report the above PEs.
  - E.g., compare your new method against others in the literature.
- If finding j^* is just part of model selection, then compute $\widehat{f}^{(j^*)}$ using the entire data set. This is the final fitted model, built from all observations.
  - E.g., determine an appropriate value for a tuning parameter λ.

Leave-one-out Cross-validation

- It is just delete-1 Jackknifing.
- Generally, computationally expensive for n large.
- For some special models, there exist fast evaluations.

K-fold Cross-validation

- K-fold cross-validation (for each $\widehat{f}^{(j)}$):
  1. Split the data into K (roughly) equal-sized parts.
  2. Fit the model to all the data except the kth part.
  3. Calculate the PE of the fitted model over the kth part.
  4. Repeat the above steps for $k = 1, \ldots, K$ and compute the mean PE.
- Find j_{opt}, the optimal j that gives the smallest PE.

K-fold Cross-validation II

- Common choices: $K = 10$, 5, or 2.
- Cross-validation (CV) makes a more efficient use of data than jackknifing.
  - At the same computational cost, CV gives more accurate PE estimation.
- To be more accurate, one may shuffle (randomly) the data, repeat the K-fold cross-validation a number of times, and compute the overall mean PE.

Using Same Subsamples

- Data resampling methods are computationally intensive, so it is very important to make them as efficient as possible.
- One very useful technique is to use the same subsamples for different methods included in comparison, i.e, the same training and test sets.
- This does not change the variation of the estimated PE of any method, but it helps greatly when comparing their relative performance.
- The technical reason behind is known as correlated sampling.
    - It is for the same reason that CV is more efficient than Jackknifing.

Computing Issues

- It is often better to use random seeds so that your random subsamples and results are repreducible.

```
> set.seed(769)   # set a random seed
> rnorm(5)
[1]  0.4679440  0.2219677 -1.2712991 -0.8091662
[5] -1.5271119
> set.seed(769)   # set the same random seed
> rnorm(5)
[1]  0.4679440  0.2219677 -1.2712991 -0.8091662
[5] -1.5271119
```

- Good to use parallel computing.
  - Make sure you can still create the same subsamples for different methods.
  - Make sure your subsamples and results are reproducible

Recommended Readings

ISLv2 (basics):

- Sections 5.1.1–5.1.3
- Labs: Sections 5.3.1–5.3.3

ESL (advanced):

- ESL: Section 7.10