Model Selection, Validation and Ensembles

Alvaro J. Riascos Villegas
University of los Andes and Quantil

July 6 2018

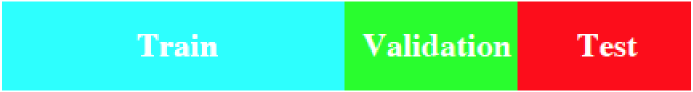
Contenido

Objetives

- Choose hyper parameters: regularization parameter, number of nearest neighbors, layers or neurons of artificial neural net, etc.
- Estimating test error.

Best of the worlds

- In reach data environments it is possible to estimate correctly the hyper parameters of a model as well as the text errro.



Contenido

Cross-Validation

- When data is not so rich an alternative is to do cross validation. Cross validation is another key concept form ML.
- It is a technique to estimate hyper parameters and the expected test error.
- K- fold cross validation:
  1. Divide in K random samples the originals data set. Given sample k, train a model with the rest of the samples ($K - 1$ samples). Test the model with the choosen sample. Repeat and take an average of the K cross validation errors.
  2. When $K = N$, the size of the training set, it is called leave one out cross validation.

Cross-Validation: K vrs. 1

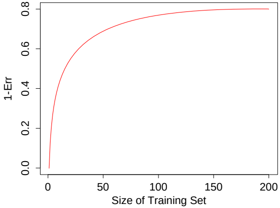


FIGURE 7.8. *Hypothetical learning curve for a classifier on a given task: a plot of* $1-\text{Err}$ *versus the size of the training set* N*. With a dataset of* 200 *observations,* 5*-fold cross-validation would use training sets of size* 160*, which would behave much like the full set. However, with a dataset of* 50 *observations fivefold cross–validation would use training sets of size* 40*, and this would result in a considerable overestimate of prediction error.*

- The optimal k depends on the size of the data. A large K with few data, overestimates the test error. A low K underestimates the test error. $K = 5, 10$ are standard.

Cross-Validation: Correct use

- Consider a problem with many predictors.
- Reduce the number of predictors using any of the studied techniques.
- Use cross validation to estimate hyper parameters and test error.
- Is this a good use of cross-validations?

Contenido

Introduction

- These are general methodologies for combining a collection of simpler models.
- We give a very brief overview of:
    1. Bagging and Sub-bagging.
    2. Boosting.
    3. Stacking.

In a nutshell

- Bagging is a general methodology for averaging models and reducing variance.
- Bagging is a bootstrap of the prediction.
- Sub-bagging is a special case, it is also a boostrap of the prediction and balances clases when disproportionately unbalanced.

In a nutshell

- Bagging is a general methodology for averaging models and reducing variance.
- Bagging is a bootstrap of the prediction.
- Sub-bagging is a special case, it is also a boostrap of the prediction and balances clases when disproportionately unbalanced.

In a nutshell

- Bagging is a general methodology for averaging models and reducing variance.
- Bagging is a bootstrap of the prediction.
- Sub-bagging is a special case, it is also a boostrap of the prediction and balances clases when disproportionately unbalanced.

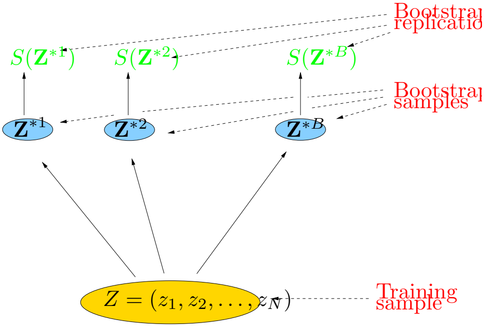


FIGURE 7.12. *Schematic of the bootstrap process. We wish to assess the statistical accuracy of a quantity* $S(\mathbf{Z})$ *computed from our dataset.* B *training sets* $\mathbf{Z}^{*b}$, $b = 1, \ldots, B$ *each of size* N *are drawn with replacement from the original dataset. The quantity of interest* $S(\mathbf{Z})$ *is computed from each bootstrap training set, and the values* $S(\mathbf{Z}^{*1}), \ldots, S(\mathbf{Z}^{*B})$ *are used to assess the statistical accuracy of* $S(\mathbf{Z})$.

Setup

- Let $\tau_n = \{(x_1, y_1), ..., (x_m, y_m)\}$ donde $y_i \in \{-1, 1\}$.
- Define an initial set of weights for each obervation$i$: $D_1(i) = \frac{1}{m}$. D_t will denote a distribution on the m observations.

AdaBoost

- For each $t = 1, \ldots, T$
- Construct a weak classifier $h_t$that minimizes the loss function:
  1. Define the error e_t as:

$$e_t = \sum_{i=1}^{m} D_t(i) I(y_i \neq h_t(x_i)) \tag{1}$$

  2. Let $\alpha_t = \frac{1}{2} \log(\frac{1-e_t}{e_t})$
  3. Modify weights:

$$D_{t+1}(i) \to \frac{D_t(i) \exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$

where $Z_t = \sum_{i=1}^{m} D_t(i)$

- $H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x))$

Optimal combination of models

- Take the prediction of many models as features in a regression problem.
- A simple example is to use regularization techniques to make combined model (e.g., Ridge, Lasso, etc)