Cross-Validation

By:
Huaicheng Liu
Jiaxin Deng

Overviews

- 1.Model Assessment and Selection
- 2.Cross-Validation
- 3.K-Fold Cross-Validation

- The Application of Cross-Validation
- (1)What value should we choose for K
- (2)The wrong and right way to do Cross-Validation

Background: Model Assessment and Selection

Introduction

- The generalization performance of a learning method relates to its prediction capability on independent test data.
- Assessment of this performance is extremely important in practice.

Formulas:

$$(1).L(Y,\hat{f}(x))=\begin{cases}(Y-\hat{f}(x))^2\\ \left|Y-\hat{f}(x)\right|\end{cases}$$

$$(2).Err_{\tau}=E\left[L(Y,\hat{f}(x))\middle|\tau\right]$$

$$(3).Err=E\left[L(Y,\hat{f}(x))\right]=E\left[Err_{\tau}\right]$$

$$(4).\overline{err}=\frac{1}{N}\sum_{i=1}^{N}L(y_i,\hat{f}(x_i))$$

It is important to note that there are in fact two separate goals that we might have in mind:

- Model selection: estimating the performance of different models in order to choose the best one.

- Model assessment: having chosen a final model, estimating its prediction error (generalization error) on new data.

- The training set is used to fit the models
- The validation set is used to estimate prediction error for model selection
- The test set is used for assessment of the generalization error of the final chosen model.

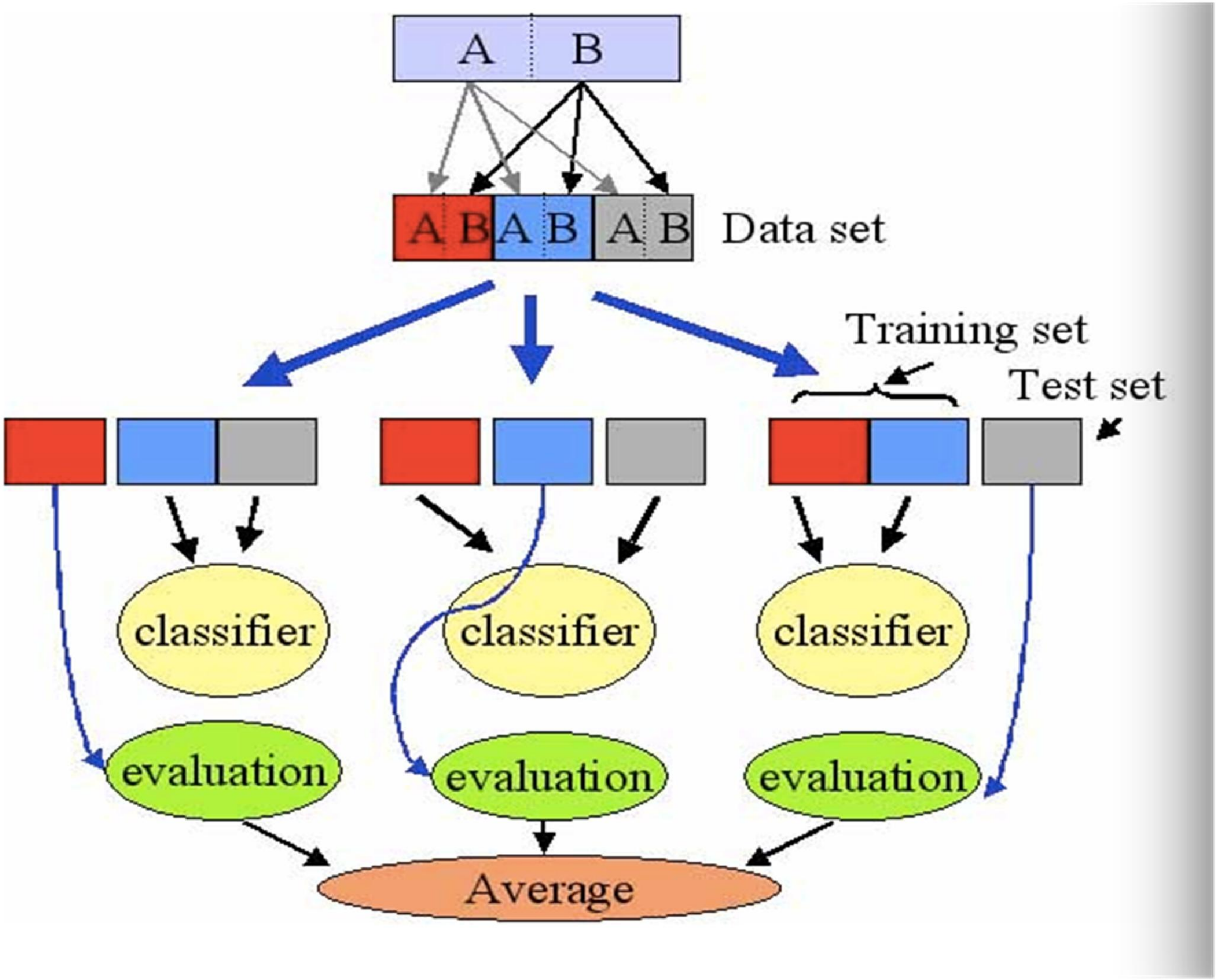
Main Contents:

Cross-Validation

- What is the Cross-Validation?

Conception:

- Cross-Validation is used to verify the classifier performance of a statistical analysis method.
- The data sets is divided into two parts, one part as a training set, another part as the test set.
- The classifier is trained with training set, test set is used to test the model obtained from the training. Be used to evaluate the classifier performance.


A
B
A B A B A B
Data set
Training set
Test set
classifier
classifier
classifier
evaluation
evaluation
evaluation
Average

Cross-Validation methods are:

- Hold-Out Method
- Leave-One-Out Cross-Validation
- K-Fold Cross-Validation
- K*2-Fold Cross-Validation

Hold-Out Method

- Data sets are randomly divided into two groups, one group as the training set, another group as the test set.
- Using the training set training classifier, and then using the test set checking the model.

Leave-One-Out Cross-Validation

- We assume data sets have N samples, each samples separately as a test set, the rest samples as the training set, it can get N models.
- Finally, we can get the average of the prediction error about the models of test set.

K-Fold Cross-Validation

- The data sets into k groups on average, each subset data for a test set respectively, the remaining subset k-1 as the training set.
- For the Kth part, we fit the model to the other K-1 parts of the data, and calculate the prediction error of the fitted model when predicting the Kth part of data. We do this for k=1 , … , K and combine the K estimates of prediction error.

Details:

- Denote by $\hat{f}^{-k}(x)$ the fitted function, computed with the kth part of the data removed.
- Then the cross-validation estimate of prediction error is

$$CV(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}^{-k(i)}(x_i))$$

- Given a set of models $f(x,\alpha)$ indexed by a tuning parameter α, denote by $\hat{f}^{-k}(x,\alpha)$ the α th model fit with the kth part of the data removed. Then for this set of models we define

$$CV(\hat{f},\alpha) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}^{-k(i)}(x_i,\alpha))$$

- The function $CV(\hat{f},\alpha)$ provides an estimate of the test error curve, and we find the tuning parameter $\hat{\alpha}$ that minimizes it. Our final chosen model is $f(x,\hat{\alpha})$ which we then fit to all the data.

K*2-Fold Cross-Validation

- The change of K-Fold Cross-Validation method, for each group of k, to average is divided into two sets: S1, S.

- Training with S1, and S test; then use S training, S1 test.

The quantity of K-Fold Cross-Validation estimates

- With k=5 or 10, we might guess that it estimates the expected error Err.

- If K=N we might guess that cross-validation estimates the conditional error Err_{τ}.

- What value should we choose for K?

The Application of Cross-Validation

1. What value should we choose for K?

2.The wrong and right way to do Cross-Validation

Section 1:

What value should we choose for K?

Section 1:What value should we choose for K?

In cross-validation with given K, we consider:

- Err: the average prediction error;
- Variance of estimation;
- Computational burden etc.

Section 1:What value should we choose for K?

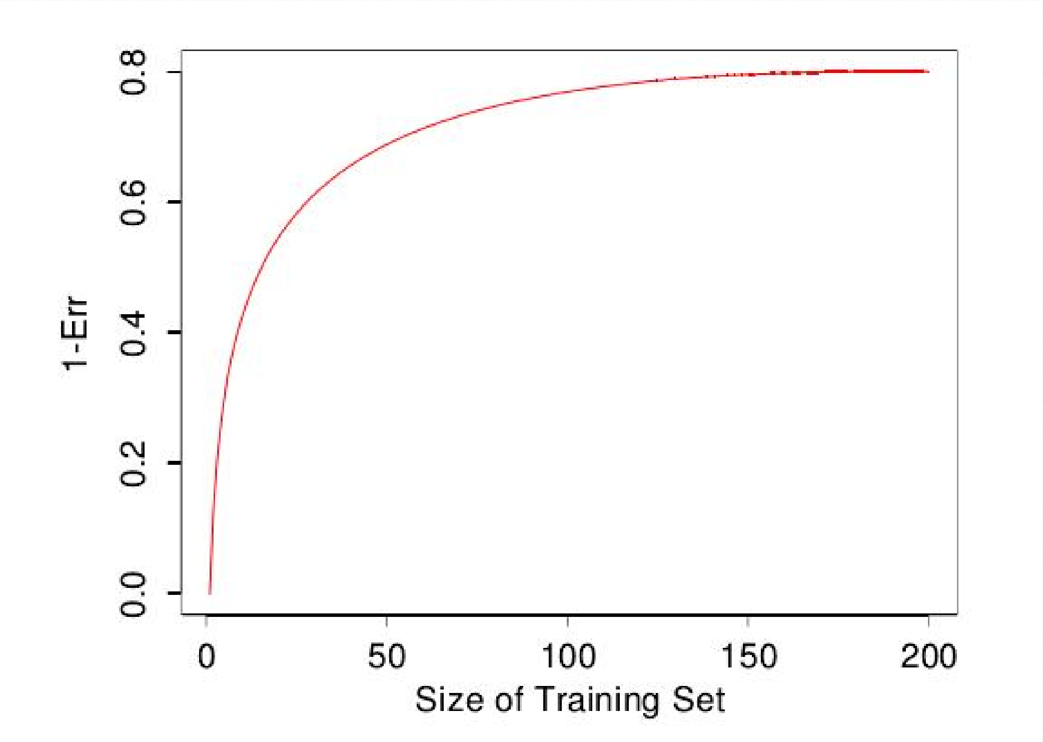
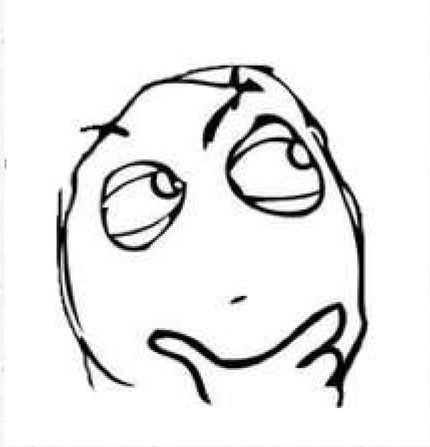


FIGURE 1. Hypothetical learning curve for a classifier on a given task: a plot of 1-Err versus the size of the training set N.

Section 1:What value should we choose for K?

Another situation:

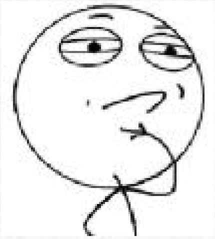
What if we only have 50 samples in the model?



Section 1:What value should we choose for K?

- If the learning curve has a considerable slope at the given training set size, five or tenfold cross-validation will estimate the true prediction error effectively.

Section 1:What value should we choose for K?



Section 2:

The Wrong and Right Way to Do Cross-Validation

Section 2: The wrong and right way to do cross-validation

The predictor: a variable of our classifier

Section 2: The wrong and right way to do cross-validation

Example:

Consider a classification problem with N=50 samples in two equal-sized classes, and p=5000 predictors that are independent of the class labels. The true error rate of any classifier is 50%.

Section 2: The wrong and right way to do cross-validation

A typical strategy for analysis might be as follows:

- 1. Screen the predictors: find a subset of predictors that show fairly strong correlation with the class labels.
- 2. Using just this subset of predictors, build a multivariate classifier.
- 3. Use cross-validation to estimate the prediction error of the final model.

Section 2: The wrong and right way to do cross-validation

- Firstly we choose the 100 predictors having highest correlation with the class labels over the 50 samples.
- Then we use a 1-nearest neighbor classifier based on just these 100 predictors.
- Over 50 simulations from this setting, we build a multivariate classifier.
- Then we do cross-validation and find out the average CV error rate is 3% which is far lower than the true error rate of 50%.

Section 2: The wrong and right way to do cross-validation

Let's do the cross-validation.
OK mom, but the CV error is 3% not 50%.
You have done it in the wrong way!
How can I make it right?...

Section 2: The wrong and right way to do cross-validation

Review what we have done:

- We selected the 100 predictors having largest correlation with the class labels over all 50 samples.
- Then we leave samples out to do the cross-validation.

Here comes the problem:

- The classifier is not completely independent to the test set ,these predictors “have already seen” the left out samples.

Section 2: The wrong and right way to do cross-validation

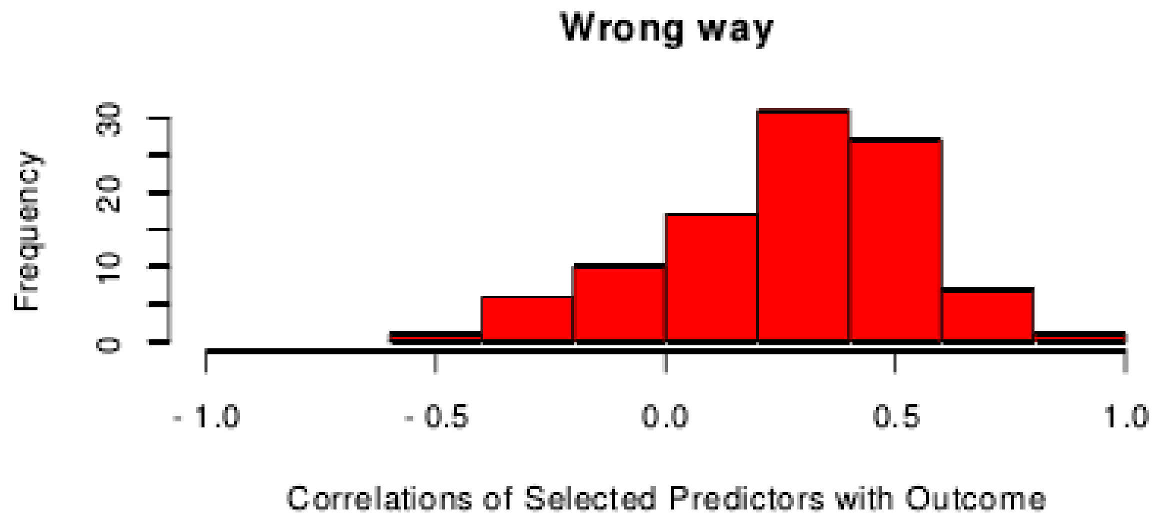


FIGURE 2: Histograms shows the correlation of class labels, in 10 randomly chosen samples, with the 100 predictors chosen using the incorrect version of cross-validation.

Section 2: The wrong and right way to do cross-validation

Here is the correct way to carry out cross-validation in this example:

- 1.Divide the samples into K cross-validation folds at random.
- 2. (a) Find a subset of “good” predictors, using all of the samples except those in fold K

  (b) Build a multivariate classifier

  (c) Use the classifier to predict the class labels for the samples in fold k.

Section 2: The wrong and right way to do cross-validation

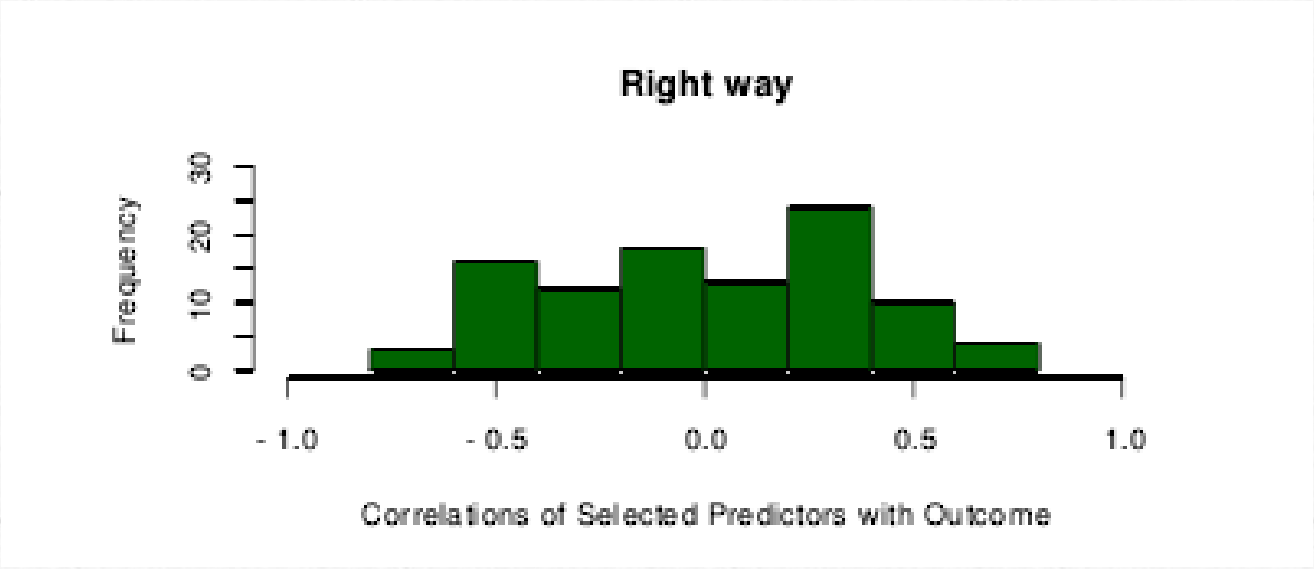


FIGURE 3. Histograms shows the correlation of class labels, in 10 randomly chosen samples, with the 100 predictors chosen using the correct version of cross-validation.

Are you intrested in the cross-validation?
I think maybe some of you are !
Just learn deeper !

Thanks for listening!