

Lecture 10: Cross validation

Reading: Sections 7.10, 7.11

GU4241/GR5241 Statistical Machine Learning

Linxi Liu

February 21, 2017

Validation set approach

Goal:

- ▶ Estimate the test error of a learning method.
- ▶ Select the best model from a given set of models. “Set of models” can simply mean “set of different parameter values”.

Elements of Statistical Learning (2nd Ed.) ©Hastie, Tibshirani & Friedman 2009 Chap 2

Elements of Statistical Learning (2nd Ed.) ©Hastie, Tibshirani & Friedman 2009 Chap 2

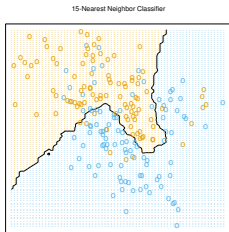


FIGURE 2.2. The same classification example in two dimensions as in Figure 2.1. The classes are coded as a binary variable (BLUE = 0, ORANGE = 1) and then fit by 15-nearest-neighbor averaging as in (2.8). The predicted class is hence chosen by majority vote amongst the 15-nearest neighbors.

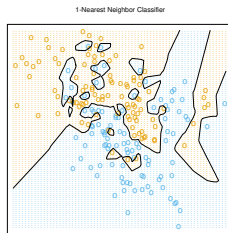


FIGURE 2.3. The same classification example in two dimensions as in Figure 2.1. The classes are coded as a binary variable (BLUE = 0, ORANGE = 1), and then predicted by 1-nearest-neighbor classification.

Training Vs. test error

Training error IS NOT a good estimate of the test error.

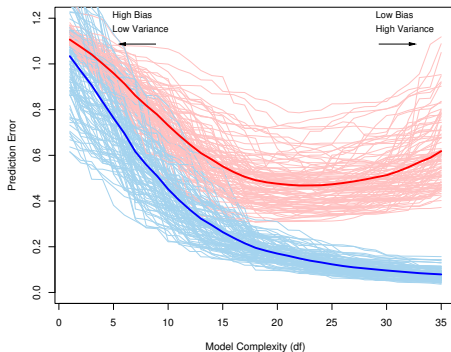


FIGURE 7.1. Behavior of test sample and training sample error as the model complexity is varied. The light blue curves show the training error $\overline{\text{err}}$, while the light red curves show the conditional test error $\text{Err}_{\mathcal{T}}$ for 100 training sets of size 50 each, as the model complexity is increased. The solid curves show the expected test error Err and the expected training error $\text{E}[\overline{\text{err}}]$.

Some concepts

$L(\cdot, \cdot)$ is the loss function. training error will underestimate the test error

Training error is the average loss over the training sample

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

Test error, also referred to as **generalization error**, is the prediction error over an independent test sample

test error has the 3 part


$$\text{Err}_{\mathcal{T}} = \mathbb{E}[L(Y, \hat{f}(X)) | \mathcal{T}].$$

Usually, it is more amenable to estimate the **expected prediction error** (or expected test error)

$$\text{Err} = \mathbb{E}[L(Y, \hat{f}(X))] = \mathbb{E}[\text{Err}_{\mathcal{T}}].$$

Model selection

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

- ▶ Randomly split data into three sets: a training set, a validation set and a test set.
calculate the average loss of the validation, because if it the data on the training set, then we have a good estimation. we choose a tuning parameter minimus he test error
- 
- ▶ Train different models on the training set.
 - ▶ Evaluate each trained model on the validation set (i.e. compute prediction error).
 - ▶ Select the model with lowest prediction error.

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

- ▶ Finally, estimate the prediction error of the selected model on the test set.

K -fold cross-validation

Each of the error estimates computed on validation set is computed from a single example of a trained classifier. Can we improve the estimate?

Strategy:

- ▶ Set aside the test set.
- ▶ Split the remaining data into K blocks.
- ▶ Use each block in turn as validation set. Perform cross validation and average the results over all K combinations.

This method is called **K -fold cross-validation**.

each time we use one block of data as training data, then we calculate the average loss

Example: $K=5$, step $k=3$

of the test error repeat this step 5 times we have 5 test errors, we

1	2	3	4	5
Train	Train	Validation	Train	Train

how to average
 $= 1/5 \sum L(y_i - f(x_i))$

K -fold cross-validation

Assume we have a set of models $f(\cdot, \alpha)$ indexed by a tuning parameter α .

Estimating prediction error

- ▶ Split data into K equally sized blocks.
- ▶ Train an instance $\hat{f}^{-k}(\cdot, \alpha)$ of the model, using all blocks except block k as training data.
- ▶ Compute the cross validation estimate

$$CV(\hat{f}, \alpha) := \frac{1}{K} \sum_{k=1}^K \frac{1}{|\text{block } k|} \sum_{(x,y) \in \text{block } k} L(y, \hat{f}^{-k}(x, \alpha))$$

Repeat this for each tuning parameter α .

Selecting a model

Choose the parameter value α^* according to certain criteria.

Model assessment

Report risk estimate for $f(\cdot, \alpha^*)$ computed on *test* data.

cv error

increase due to overfitting

1/k

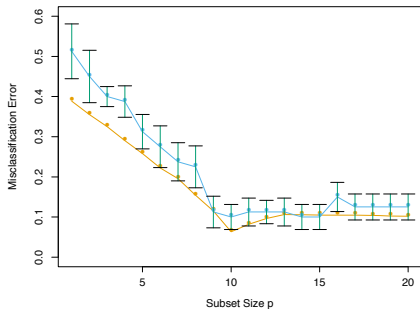
for each k , we get one point 'on the plot, and there is a trade off between bias and variance

The one standard error rule

point on the blue curve: cross validation error

yellow: one sd from cross validation error

Forward stepwise selection



Blue: 10-fold cross validation

Yellow: True test error

- ▶ A number of models with $10 \leq p \leq 15$ have the same CV error.
- ▶ The vertical bars represent 1 standard error in the test error from the 10 folds.
- ▶ **Rule of thumb:** Choose the simplest model whose CV error is no more than one standard error above the model with the lowest CV error.

How to choose K

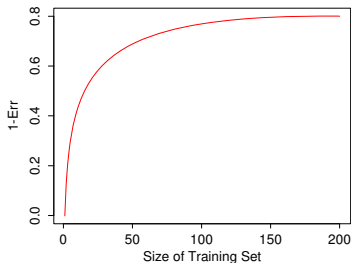


FIGURE 7.8. *Hypothetical learning curve for a classifier on a given task: a plot of $1 - \text{Err}$ versus the size of the training set N . With a dataset of 200 observations, 5-fold cross-validation would use training sets of size 160, which would behave much like the full set. However, with a dataset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

Leave one out cross-validation

- ▶ For every $i = 1, \dots, n$:
 - ▶ train the model on every point except i ,
 - ▶ compute the test error on the held out point.
- ▶ Average the test errors.



Leave one out cross-validation

- ▶ For every $i = 1, \dots, n$:
 - ▶ train the model on every point except i ,
 - ▶ compute the test error on the held out point.
- ▶ Average the test errors.

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i^{(-i)})^2$$

Prediction for the i sample without using the i th sample.

Leave one out cross-validation

- ▶ For every $i = 1, \dots, n$:
 - ▶ train the model on every point except i ,
 - ▶ compute the test error on the held out point.
- ▶ Average the test errors.

$$\text{CV}_{(n)} = \frac{1}{n} \sum_{i=1}^n \mathbf{1}(y_i \neq \hat{y}_i^{(-i)})$$

... for a classification problem.

Leave one out cross-validation

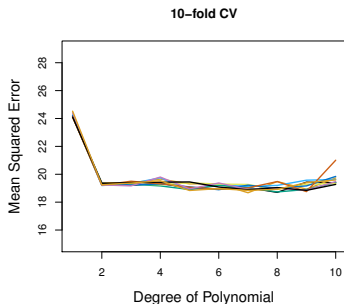
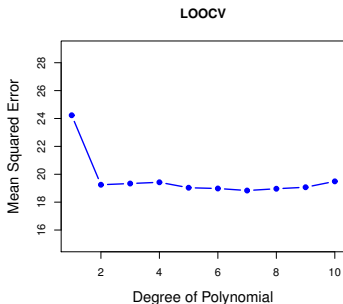
Computing $CV_{(n)}$ can be computationally expensive, since it involves fitting the model n times.

For linear regression, there is a shortcut:

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \hat{y}_i}{1 - h_{ii}} \right)^2$$

where h_{ii} is the leverage statistic.

LOOCV vs. K -fold cross-validation



- ▶ K -fold CV depends on the chosen split.
- ▶ In K -fold CV, we train the model on less data than what is available. This introduces **bias** into the estimates of test error.
- ▶ In LOOCV, the training samples highly resemble each other. This increases the **variance** of the test error estimate.

The wrong way to do cross validation

Reading: Section 7.10.2 of The Elements of Statistical Learning.

We want to classify 200 individuals according to whether they have cancer or not. We use logistic regression onto 1000 measurements of gene expression.

Proposed strategy: [to estimate the test data](#)

- ▶ Using all the data, select the 20 most significant genes using z -tests.
- ▶ Estimate the test error of logistic regression with these 20 predictors via 10-fold cross validation.

The wrong way to do cross validation

To see how that works, let's use the following simulated data:

- ▶ Each gene expression is standard normal and independent of all others.
- ▶ The response (cancer or not) is sampled from a coin flip — no correlation to any of the “genes”.

What should the misclassification rate be for any classification method using these predictors?

Roughly 50%.

The wrong way to do cross validation

We run this simulation, and obtain a CV error rate of 3%!

Why is this?

- ▶ Since we only have 200 individuals in total, among 1000 variables, at least some will be correlated with the response.
- ▶ We do variable selection using *all the data*, so the variables we select have some correlation with the response in every subset or fold in the cross validation.

The **right** way to do cross validation

- ▶ Divide the data into 10 folds.
- ▶ For $i = 1, \dots, 10$:
 - ▶ Using every fold except i , perform the variable selection and fit the model with the selected variables.
 - ▶ Compute the error on fold i .
- ▶ Average the 10 test errors obtained.

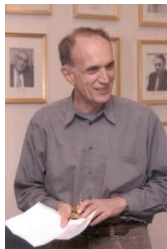
In our simulation, this produces an error estimate of close to 50%.

Moral of the story: Every aspect of the learning method that involves using the data — variable selection, for example — must be cross-validated.

Cross-validation vs. the Bootstrap

Cross-validation: provides **estimates** of the (test) **error**.

The Bootstrap: provides the (standard) **error** of **estimates**.



- ▶ One of the most important techniques in all of Statistics.
- ▶ Computer intensive method.
- ▶ Popularized by Brad Efron.

Standard errors in linear regression

Standard error: SD of an estimate from a sample of size n .

```
Residuals:
    Min       1Q   Median       3Q      Max
-15.594   -2.730   -0.518    1.777   26.199

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
crim         -1.080e-01  3.286e-02  -3.287 0.001087 **
zn           4.642e-02  1.373e-02   3.382 0.000778 ***
indus        2.056e-02  6.150e-02   0.334 0.738288
chas         2.687e+00  8.616e-01   3.118 0.001925 **
nox          -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
rm           3.810e+00  4.179e-01   9.116 < 2e-16 ***
age          6.922e-04  1.321e-02   0.052 0.958229
dis          -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
tax          -1.233e-02  3.761e-03  -3.280 0.001112 **
ptratio      -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
black        9.312e-03  2.686e-03   3.467 0.000573 ***
lstat        -5.248e-01  5.072e-02 -10.347 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.745 on 492 degrees of freedom
Multiple R-Squared:  0.7406,    Adjusted R-squared:  0.7338
F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

Classical way to compute Standard Errors

Example: Estimate the variance of a sample x_1, x_2, \dots, x_n :

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

What is the Standard Error of $\hat{\sigma}^2$?

- ▶ Assume that x_1, \dots, x_n are normally distributed.
- ▶ Assume that the true variance is close to $\hat{\sigma}^2$ and the true mean is close to \bar{x} .
- ▶ Then $\hat{\sigma}^2(n-1)$ has a χ -squared distribution with n degrees of freedom.
- ▶ The SD of this *sampling distribution* is the Standard Error.

Limitations of the classical approach

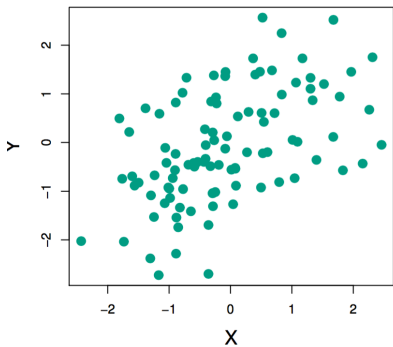
This approach has served statisticians well for 90 years; however, what happens if:

- ▶ The distributional assumption — for example, x_1, \dots, x_n being normal — breaks down?
- ▶ The estimator does not have a simple form and its sampling distribution cannot be derived analytically?

Example. Investing in two assets

Suppose that X and Y are the returns of two assets.

These returns are observed every day: $(x_1, y_1), \dots, (x_n, y_n)$.



Example. Investing in two assets

We have a fixed amount of money to invest and we will invest a fraction α on X and a fraction $(1 - \alpha)$ on Y . Therefore, our return will be

$$\alpha X + (1 - \alpha)Y.$$

Our goal will be to minimize the variance of our return as a function of α . One can show that the optimal α is:

$$\alpha = \frac{\sigma_Y^2 - \text{Cov}(X, Y)}{\sigma_X^2 + \sigma_Y^2 - 2\text{Cov}(X, Y)}.$$

Proposal: Use an estimate:

$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\text{Cov}}(X, Y)}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\text{Cov}}(X, Y)}.$$

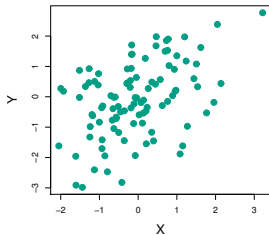
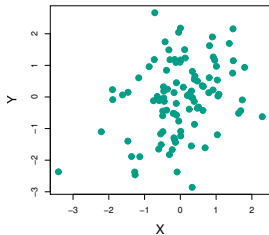
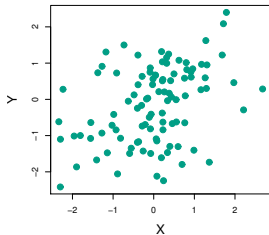
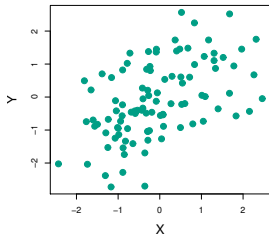
Example. Investing in two assets

Suppose we compute the estimate $\hat{\alpha} = 0.6$ using the samples $(x_1, y_1), \dots, (x_n, y_n)$.

- ▶ How sure can we be of this value?
- ▶ If we resampled the observations, would we get a wildly different $\hat{\alpha}$?

In this thought experiment, we know the actual joint distribution $P(X, Y)$, so we can resample the n observations to our hearts' content.

Resampling the data from the true distribution



Computing the standard error of $\hat{\alpha}$

For each resampling of the data,

$$(x_1^{(1)}, \dots, x_n^{(1)})$$

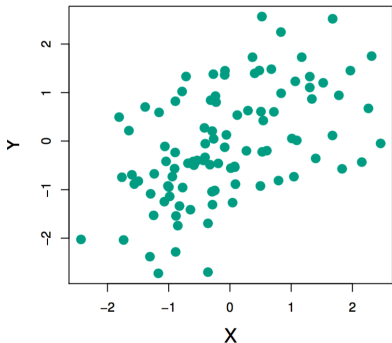
$$(x_1^{(2)}, \dots, x_n^{(2)})$$

...

we can compute a value of the estimate $\hat{\alpha}^{(1)}, \hat{\alpha}^{(2)}, \dots$

The Standard Error of $\hat{\alpha}$ is approximated by the standard deviation of these values.

In reality, we only have n samples

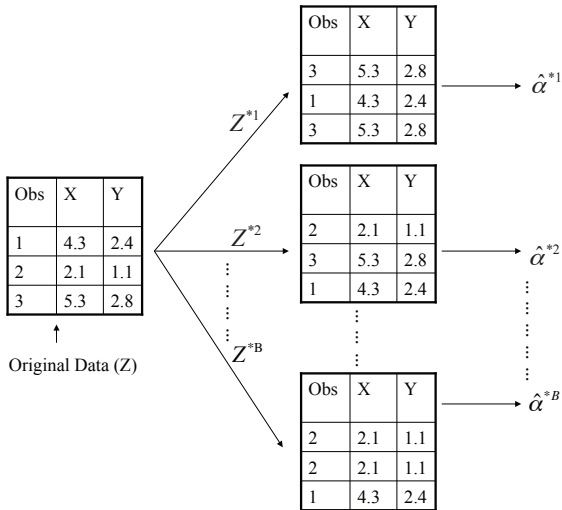


- ▶ However, these samples can be used to approximate the joint distribution of X and Y .
- ▶ **The Bootstrap:** Resample from the *empirical distribution*:

$$\hat{P}(X, Y) = \frac{1}{n} \sum_{i=1}^n \delta(x_i, y_i).$$

- ▶ Equivalently, resample the data by drawing n samples *with replacement* from the actual observations.

A schematic of the Bootstrap



Comparing Bootstrap resamplings to resamplings from the true distribution

