Lecture 9: Cross-validation

Statistical Learning (BST 263)

Jeffrey W. Miller

Department of Biostatistics
Harvard T.H. Chan School of Public Health

Outline

Choosing the amount of flexibility

Cross-validation (CV)

Choosing model settings via CV

Choosing the number of folds in CV

Outline

How to choose a good flexibility level in practice?

- Many methods have knobs that control the amount of flexibility.
  - E.g., # of neighbors in KNN.

- Theory (bias-variance decomposition) and simulations show us that the flexibility needs to be chosen well in order to obtain good test performance.

- How can we know what degree of flexibility will yield good performance on future test data?

How to choose a good flexibility level in practice?

- Train/test splits:
  1. split the data into pseudo-training and pseudo-test sets,
  2. fit the model on the pseudo-training set, and
  3. measure performance on the pseudo-test set.

- This provides an estimate of the test performance of the method on the data generating process of interest.

- The accuracy of this estimate can be improved by repeating the process with multiple train/test splits, and then averaging the test performance estimates.

How to choose a good flexibility level in practice?

- Cross-validation (CV)
  - CV is a particular way of defining a collection of train/test splits to estimate test performance.
  - Flexibility knobs (as well as other settings) can be chosen by optimizing the CV-estimated test performance.

- Model selection criteria and Bayesian methods
  - Another approach is to optimize a criterion such as AIC or BIC, which balance fit and complexity/flexibility.
  - Bayesian methods are similar, but use a prior distribution to penalize complexity/flexibility.
  - More on this later. . .

Outline

Train/test splits

- Suppose we want to estimate test MSE,

$$\text{test MSE} = \mathrm{E}\big((\hat{Y}_0 - Y_0)^2\big).$$

- Training MSE tends to underestimate test MSE because we used the training data to fit the model.

- Idea: Split the training data into pseudo-test and pseudo-training sets.
    - Pseudo-test set = random subset of the training data.
    - Pseudo-training set = the rest of the training data.
    - Fit model on pseudo-train and measure MSE on pseudo-test.
    - This provides an estimate of test MSE.

- Why only one train/test split?
- Can improve the accuracy of this estimate by repeating over multiple splits, and averaging the pseudo-test MSEs.

K-fold cross-validation

Suppose we have $n = 100$ training data examples:

1	2	3	4	$\cdots$	99	100

Choose a random permutation of $1, \ldots, n$:

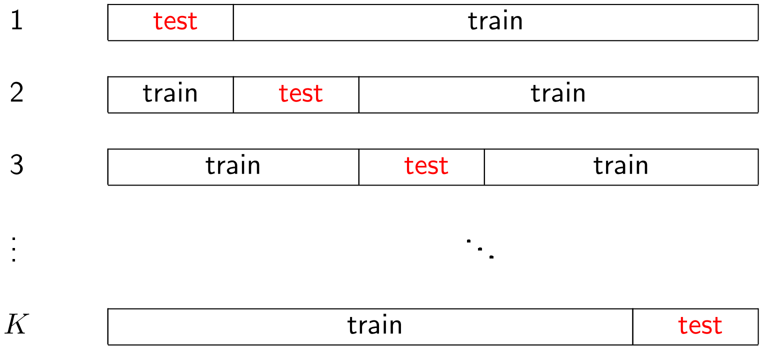
37	14	86	3	$\cdots$	62	21

Divide into K blocks ("folds") of size $\approx n/K$:

fold 1	fold 2	$\cdots$	fold K

K-fold cross-validation

For $k = 1, \ldots, K$, pseudo-test is fold k, pseudo-train is the rest:



K-fold cross-validation

- For each fold k, we get an estimate of test MSE by fitting on pseudo-train set k and measuring MSE on pseudo-test set k:

$$\widehat{\mathrm{MSE}}_k$$

- The K-fold cross-validation estimate of test MSE is obtained by averaging:

$$\widehat{\mathrm{MSE}}_{\mathsf{CV}} = \frac{1}{K} \sum_{k=1}^{K} \widehat{\mathrm{MSE}}_k.$$

- If $K = n$ then this is called *leave-one-out cross-validation* (LOO-CV).

Implementing cross-validation

(R code illustration)

Cross-validation with other loss functions

- More generally, cross-validation can be used to estimate expected loss for other loss functions:

$$\widehat{\text{loss}}_{\mathsf{CV}} = \frac{1}{K}\sum_{k=1}^{K}\widehat{\text{loss}}_k.$$

- E.g., for classification, we can estimate the test error rate.

- Careful! Is this formula a good way to estimate test RMSE?

$$\text{test RMSE} \overset{?}{\approx} \frac{1}{K}\sum_{k=1}^{K}\sqrt{\widehat{\text{MSE}}_k}.$$

("RMSE" = root mean squared error = square root of MSE)

Outline

Choosing model settings via CV

- CV is often used to choose model settings, e.g., flexibility.
  - Example: Choosing the # of neighbors in KNN.

- Suppose we want to choose some model setting α in order to obtain good test performance.
- For each α in some range, do CV and compute $\widehat{\text{loss}}_{\text{CV}}(\alpha)$.
- Choose the α with the smallest CV estimate of expected loss:

$$\alpha_{\text{CV}} = \operatorname*{argmin}_{\alpha} \widehat{\text{loss}}_{\text{CV}}(\alpha).$$

- Careful! This introduces a downward bias in $\widehat{\text{loss}}_{\text{CV}}(\alpha_{\text{CV}})$ as an estimate of the expected loss of α_{CV}. (Why?)

Choosing model settings via CV

(R code illustration)

Outline

Choosing the number of folds in CV

- How should we choose the number of folds K?
- Want CV estimate of expected loss to be accurate as possible.
- Meta-problem! Minimize MSE of the CV estimate itself.
- Need to choose # of folds K to balance bias and variance!
- Where does the bias come from???
- CV estimates of expected loss are biased upward since the pseudo-training set is smaller than the training set.
  - It's harder to learn from fewer examples, so test performance tends to be worse when training on a smaller set.

Choosing the number of folds in CV

(R code illustration)

Choosing the number of folds in CV

- Two possible objectives: Estimate expected loss when...
  (a) fitting on the training set we actually have:

  $$\text{expected loss given training set} = \mathrm{E}\big(L(\hat{Y}_0, Y_0) \,\big|\, x_{1:n}, y_{1:n}\big)$$

  (b) fitting on a random training set of the same size:

  $$\text{expected loss with random training set} = \mathrm{E}\big(L(\hat{Y}_0, Y_0)\big).$$

- Recall that the randomness in $\hat{Y}_0$ can come from the test point X_0, the training x's, and/or the training y's.

- Usually, we are interested in objective (a).
- But (b) is of interest when comparing methods in general.

Choosing the number of folds in CV

- Theory to the rescue: For objective (a), more folds is better!
  - For (a), Burman (1989) showed that the accuracy of CV is better when using more folds.
  - Accuracy is quantified in terms of MSE of the CV estimate.
- So, should we always use LOO-CV (i.e., $K = n$ folds)?
- In practice, computation is another consideration.
  - LOO-CV requires fitting n times, which may take too long.
  - The accuracy of CV may be sufficient with fewer folds.
- Recommended default choice in practice?
  - 10 folds is often a good balance of accuracy and computation.

References

P. Burman. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76(3):503–514, 1989.