3. Model Selection : Cross-validation

Bookmark this page

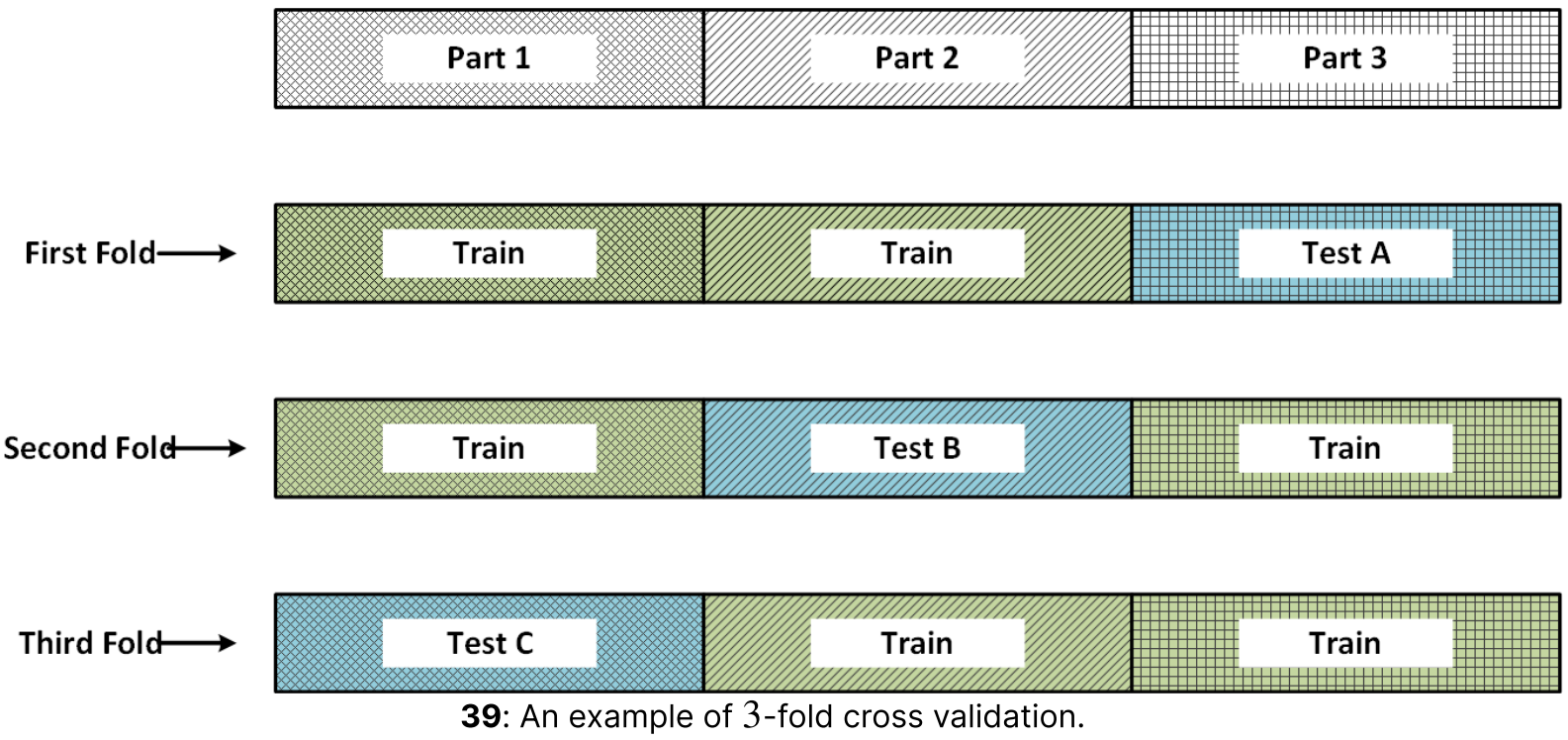
Exercises due May 21, 2021 19:59 EDT

Cross-validation

Definition 3.1 **Cross-validation** is a statistical technique for validating the generalization of a model when applied to independent data. This kind of validation consists of splitting the available data in various disjoint train and test subsets. Then, the model is used to form a prediction of the test data using the train data, and the performance evaluated. The goal is to validate the model's ability to predict new data that was not part of the training set.

Definition 3.2 **K-fold Cross-validation** assumes we partition our data set into K disjoint subsets. Then, we train K models, each one with one of the partitions selected as the test set and the remaining subsets are used to form the training set.

The below figure shows an example for 3-fold validation, where the data was partitioned into three disjoint subsets. At each fold, two of the partitions shown in green are used for training and the remaining partition shown in blue are used for testing.



39: An example of 3-fold cross validation.

The testing step on each of the folds produces a score, which will be defined according to the particular requirements of the problem. Recall that following the conditional distribution approach we have studied in the previous lecture, testing will produce the marginal distribution on the set of random variables marked for testing, from which we need to make a prediction and generate a score.

One immediate prediction could be the mean of the conditional distribution, and the score could be the mean square error to the true data point. The average score over all of the folds is computed and used to evaluate the performance of that particular choice of model. This process is repeated for each of the models being tested. Finally, one can select the model with the best performance according to the chosen score.

Definition 3.3 **Leave-one-out Cross-validation** refers to the case where the testing dataset is composed of a single data point.

Going back to the 3-fold example the above figure. Assume we have observed $\mathbf{X}_2 = \mathbf{x}_2$, and we have at our disposal $\boldsymbol{\mu_2}$. Moreover, let us assume we are working with the kernel function

$$k(z_i, z_j) = \exp\left(-\frac{\|z_i - z_j\|^2}{2\ell^2}\right), \tag{7.7}$$

with parameter $\theta = \{\ell\}$.

The validation then proceeds as follows:

1. We divide the observed random variables $\mathbf{X}_2 = \mathbf{x}_2 \in \mathbb{R}^{N-d}$ into K disjoint subsets such that

$$\mathbf{X}_2 = \begin{bmatrix} \mathbf{X}_2^1 \\ \mathbf{X}_2^2 \\ \vdots \\ \mathbf{X}_2^k \end{bmatrix}$$

2. We then select a kernel function parameterized by θ.

3. We then select the values of θ that will be tested. Recall that, for example, this will correspond to various values of ℓ in the kernel function above. In this example, we can fix a grid search and chose a finite number of values to test. Let us denote the set of possible parameter values as $\Theta = \{\theta_1, \cdots, \theta_p\}$.

4. For each of the possible values of the parameter in the set Θ:

   - Compute the covariance matrix $\boldsymbol{\Sigma}_{22}$ using the selected kernel function:

$$\boldsymbol{\Sigma}_{22} = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{X}_2^i, \mathbf{X}_2^i} & \boldsymbol{\Sigma}_{\mathbf{X}_2^i, \mathbf{X}_2^{-i}} \\ \boldsymbol{\Sigma}_{\mathbf{X}_2^{-i}, \mathbf{X}_2^i} & \boldsymbol{\Sigma}_{\mathbf{X}_2^{-i}, \mathbf{X}_2^{-i}} \end{bmatrix}.$$

     Note: Here we denote $\mathbf{X}_2^{-i}$ as the set of all random variables except for the i-th subset from the original partition into K subsets. Similarly, $\mathbf{X}_2^i$ denotes the i-th subset of the variables. This is a usual notation in combinatorial analysis, it does not mean that we are taking the $-i$-th power of the random variable.

   - Compute the conditional distribution for each of the folds where at fold i the variables $\mathbf{X}_2^i$ are used for testing, the remaining variables are denoted $\mathbf{X}_2^{-i}$ and made part of the training set. The parameters of the conditional distribution can then be computed as

$$\begin{aligned} \mu_{\mathbf{X}_2^i | \mathbf{X}_2^{-i}} &= \mu_2^i + \Sigma_{X_2^i, X_2^{-i}} \Sigma^{-1}_{\mathbf{X}_2^{-i}, \mathbf{X}_2^{-i}} (\mathbf{x}_2^{-i} - \mu_2^{-i}) \\ \Sigma_{\mathbf{X}_2^i | \mathbf{X}_2^{-i}} &= \Sigma_{\mathbf{X}_2^i, \mathbf{X}_2^i} - \Sigma_{\mathbf{X}_2^i, \mathbf{X}_2^{-i}} \Sigma^{-1}_{\mathbf{X}_2^{-i}, \mathbf{X}_2^{-i}} \Sigma_{\mathbf{X}_2^{-i}, \mathbf{X}_2^i}. \end{aligned}$$

   - With the parameters of the conditional distribution computed above, one can define a new distribution. In this case, one way to compute the error or performance of such estimation is to compute the predictive probability of the generated model on the testing set $\mathbf{X}_2^i$. Recall that a multivariate Normal distribution can be written as:

$$p(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^\intercal \Sigma^{-1} (x - \mu)\right),$$

     In our case, by taking the logarithm we have:

$$\begin{aligned} &\log p\left(\mathbf{X}_2^i \mid \mathbf{Y}_2, \mathbf{X}_2^{-i}, \theta\right) \\ &\qquad = \log\left(\frac{1}{(2\pi)^{((N-d)/k)/2} |\Sigma_{\mathbf{X}_2^i | \mathbf{X}_2^{-i}}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{X}_2^i - \mu_{\mathbf{X}_2^i | \mathbf{X}_2^{-i}})^\intercal \Sigma^{-1}_{\mathbf{X}_2^i | \mathbf{X}_2^{-i}} (\mathbf{X}_2^i - \mu_{\mathbf{X}_2^i | \mathbf{X}_2^{-i}})\right)\right) \end{aligned}$$

$$= -\log\left((2\pi)^{((N-d)/k)/2}|\Sigma_{\mathbf{X}_2^i|\mathbf{X}_2^{-i}}|^{1/2}\right) - \frac{1}{2}(\mathbf{X}_2^i - \mu_{\mathbf{X}_2^i|\mathbf{X}_2^{-i}})^\intercal \Sigma^{-1}\,(\mathbf{X}_2^i - \mu_{\mathbf{X}_2^i|\mathbf{X}_2^{-i}})\,,$$

5. For each of the parameters θ, we compute the predictive probability

$$\sum_{i=1}^{k} \log p\left(\mathbf{X}_2^i \mid \mathbf{Y}_2, \mathbf{X}_2^{-i}, \theta\right).$$

Then, select the θ^* that maximizes this predictive probability.

Leave-one-out vs. K-fold: bias

1 point possible (graded)
Recall that in K-fold cross-validation, one Kth of the data is reserved for testing. While in leave-one-out cross-validation, only a single data point is reserved for testing.

Let us define the bias of the cross-validation procedure as the mean of the difference between the actual performance of the model, and the performance estimated using the cross-validation procedure. Which method will generally produce an estimate with higher bias?

- Leave-one-out
- K-fold

Submit You have used 0 of 1 attempt

Leave-one-out vs. K-fold: variance

1 point possible (graded)
Let us define the variance of the cross-validation procedure as the variance of the performance estimate. Which method will generally produce an estimate with higher variance?

- Leave-one-out
- K-fold

Submit You have used 0 of 1 attempt

Discussion

Hide Discussion

Topic: Module 5: Environmental Data and Gaussian Processes:Sensing and Analyzing global patterns of dependence / 3. Model Selection : Cross-validation

Add a Post

Show all posts by recent activity

There are no posts in this topic yet.

✖

‹ Previous Next ›