Exercise Set 3

Please write the name(s) of the student(s) to the answer sheet! Submission DL 14 May 9:15. The solutions will be presented in a session at noon 14 May, where you should be present (you may be asked to present your solutions).

Problem 1

Learning objective: validation procedure.

Well continue from Problem 3 of Exercise Set 2, see E2P3 for context and a more detailed description.

Tasks

- Do E2P3.
- Now do the validation and testing properly: yake your 100 data points and split them in random into training+validation set and a new test set. Use cross-validation on training+validation set to find the best model family (your model families: OLS linear regression, random forest etc.). Train the chosen model family on full training+validation set and estimate the out-sample loss on new test set. Then train the model on all 100 points. Report the training/validation/in-sample/out-sample losses during the process. How do you losses compare with the "true loss" of your model on the remaining 9096 data items in the full CO2 flux dataset?

Problem 2

Learning objective: the tragedy of many models.

We will continue here with the CO2 flux dataset of E2P3 and Problem 1 above. The purpose of this problem is to study experimentally the properties of validation loss, i.e., the loss you obtain with validation procedure (e.g., with k-fold cross validation) and the error you incur because you select a model with the smallest validation loss.

You can think validation loss (e.g., estimate of loss given by cross-validation) as a random variable: for a different (random) selection of training+validation data you will get a slightly different validation loss. The validation loss is an estimate of the "true loss" on newly sampled data not used in training+validation. As all real valued random variables that are used to estimate other parameters, also this random variable has bias and variance.

Tasks

- Pick one of your regression models from E2P3, resample a new 100 points and compute the validation loss as you did in Problem 1 above. Repeat the procedure, e.g., 100 times, thereby obtaining 100 validation losses (each for newly sampled training+vaidation data). Because the validation losses are computed for different datasets they should differ slightly from each other. Compute and report mean and standard deviation of your valiation losses. Compare the mean to the loss on the out-sample data. Does the validation loss seem to be an unbiased estimate of the loss on data not used in the trainig? Is the variance of your estimator smaller or larger than you would expect?
- The real purpose of validation procedure is to choose out of k models the one with the smallest loss on the test data. In the validation set approach, you choose the model with the smallest validation set loss.

Now consider a situation where your k models all have the same expected loss on the test data, i.e., they are "equally good". Due to the variance in the validation loss estimate you will incur negative bias to your estimate when you select the smallest of k losses: when you select the smallest of k numbers you will probably select smaller than average number. Your task here is to simulate the bias due to multiple selection between k models in the case. More specifically: assume that you have k models and that the validation loss of each of the models obeys normal distribution with the mean and variance given by your result in the item above. Draw k samples from this normal distribution (simulating validation losses of k different models) and pick the smallest number. Repeat this procedure, e.g., 1000 times and find the mean and standard deviation of this simulated estimator. Plot the bias of this estimator as a function of the number of models k.
- Compare this bias to the differences between validation losses and test set losses in Problem 1.

Problem 3

Learning objectives: bias and variance and model flexibiility.

Read Section 2.2 of James et al.

Consider the bias variance decomposition in the context of model selection.

Tasks

- Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be five curves. Make sure to label each one.
- Explain why each of the five curves has the shape displayed.

Problem 4

Learning objectives: properties of (cross-)validation.

Consider "normal" case where data is sampled i.i.d. from unknown but fixed distribution and cross-validation procedure is applied to estimate loss on new data (not used in training). Consider a simple validation procedure where you split the n data items in random into a training set of size m and a validation set of size $n - m$.

Tasks

- Show that the loss on the validation set is unbiased estimate of the loss on test data.
- Read about the bias-variance trade-off for k-fold cross-validation. (James et al., Section 5.1.4)