Homework 8

36-465/665, Spring 2021

Due at 6 pm on Thursday, 1 April 2021

Agenda: Working with data splitting for model selection; inference after selection.

1. *Multiple optima* Lecture 15 defined $k^{\dagger}$ as the k which minimizes $r(s^*_k)$, and k^* as the k which minimizes $r(\hat{s}_k)$.

   a. (5) Is $k^{\dagger}$ random? Does it change with n?

   b. (5) Is k^* random? Does it change with n?

2. (14) *Generalization error bound for hold-out* In the notes for lecture 15, we established that when the data are IID and the loss ℓ is bounded between 0 and m,

$$\mathbb{P}\left(r(\hat{s}_{\hat{k}}) \leq r(\hat{s}_{k^*}) + m\sqrt{\frac{2\log{(2q/\alpha)}}{n_s}}\right) \geq 1-\alpha$$

   This is an example of an "oracle inequality", showing that some procedure (here, using holdout to pick $\hat{k}$) does almost as well as an "oracle" which supernaturally knows the right answer (here, the risk-minimizing model class k^*). For empirical risk minimization, oracle inequalities went along with generalization error bounds. Prove a generalization bound here, by finding a function $g(n_s, q, \alpha, m)$ such that

$$\mathbb{P}\left(r(\hat{s}_{\hat{k}}) \geq \tilde{r}(\hat{s}_{\hat{k}}) + g(n_s, q, \alpha, m)\right) \leq \alpha$$

   Hint: Look carefully at the proof in the slides, and at after-class exercise 7.

3. *Fitting an anomaly* The data file `anomaly.csv` contains a predictor variable (`t`) and a response (`anomaly`). Write code to randomly divide this into two equally sized parts, the training and the selection sets. Fit polynomials of order 0 to 20 to the training set.

   a. (5) Create a plot showing a scatter-plot of the training set, and 21 curves for the predictions of the polynomials.

   b. (5) Create a plot showing the MSE on the training set versus the order of the polynomials. What order of polynomial would you pick in order to minimize the empirical risk?

   c. (5) Create another plot which shows R^2 and "adjusted" R^2 for the polynomials as a function of the order of the polynomials. What order is favored by maximizing R^2? How is this related to the previous plot?

   d. (6) Calculate the mean squared error of each of the polynomials on the selection set. Plot these hold-out estimates of the risk against model order. What order is favored by hold-out?

   e. (6) Go back to the plot in Q3a and add the points from the selection set. Is it reasonable to pick the polynomial selected by hold-out over the one selected by R^2?

4. *The perils of post-selection inference, and data splitting to the rescue* Generate a 10000×101 array, where all the entries are IID standard Gaussian variables. Call the first column the response variable Y, and the others the predictors $X_1, \ldots X_{100}$. By design, there is no true relationship between the response and the predictors.

a. (5) Explain why all the assumptions about linear models commonly made in regression hold true here.

b. (5) The F-test for a linear regression is often said to check the over-all significance of the model. Explain how that relates to the improvement in predictive power or risk, as measured in terms of squared error. (This is not a request for deriving the F distribution, but more "improvement compared to what, exactly?" and "improvement measured how?")

c. (6) Estimate the model $Y = \beta_0 + \beta_1 X_1 + \beta_{50} X_{50} + \epsilon$. Find the p-value for the F test of the whole model. (In R, this is part of the output of `summary()` applied to an `lm` object.) Repeat the simulation, estimation and testing 100 times, and plot the histogram of the p-values. What does it look like? What should it look like?

d. (7) Use forward stepwise regression on your simulated data set to select a linear model specification. (In R, the `step()` function gives a convenient implementation.) Find the p-value for the F-test of the selected model. Repeat 100 times and plot the histogram of p-values. Explain what's going on.

e. (7) Again use `step` to select a model based on one random 5000 × 101 array. Now re-estimate the selected model on a new 5000 × 101 array, and find the new p-value. Repeat 100 times, with new selection and inference sets each time, and plot the histogram of p-values.

f. (8) Summarize what Q4c, Q4d and Q4e tell you about model selection and significance testing.

5. (1) How much time did you spend on this problem set?

Presentation rubric (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision.