Lecture 7: Model Validation

Statistical Learning and Data Mining

Xingye Qiao

Department of Mathematical Sciences
Binghamton University

E-mail: qiao@math.binghamton.edu

Outline

The next section would be . . . . . .

1. Two Goals of Statistical Learning
2. Best Practices in Data Analysis

Goal 1: Prediction

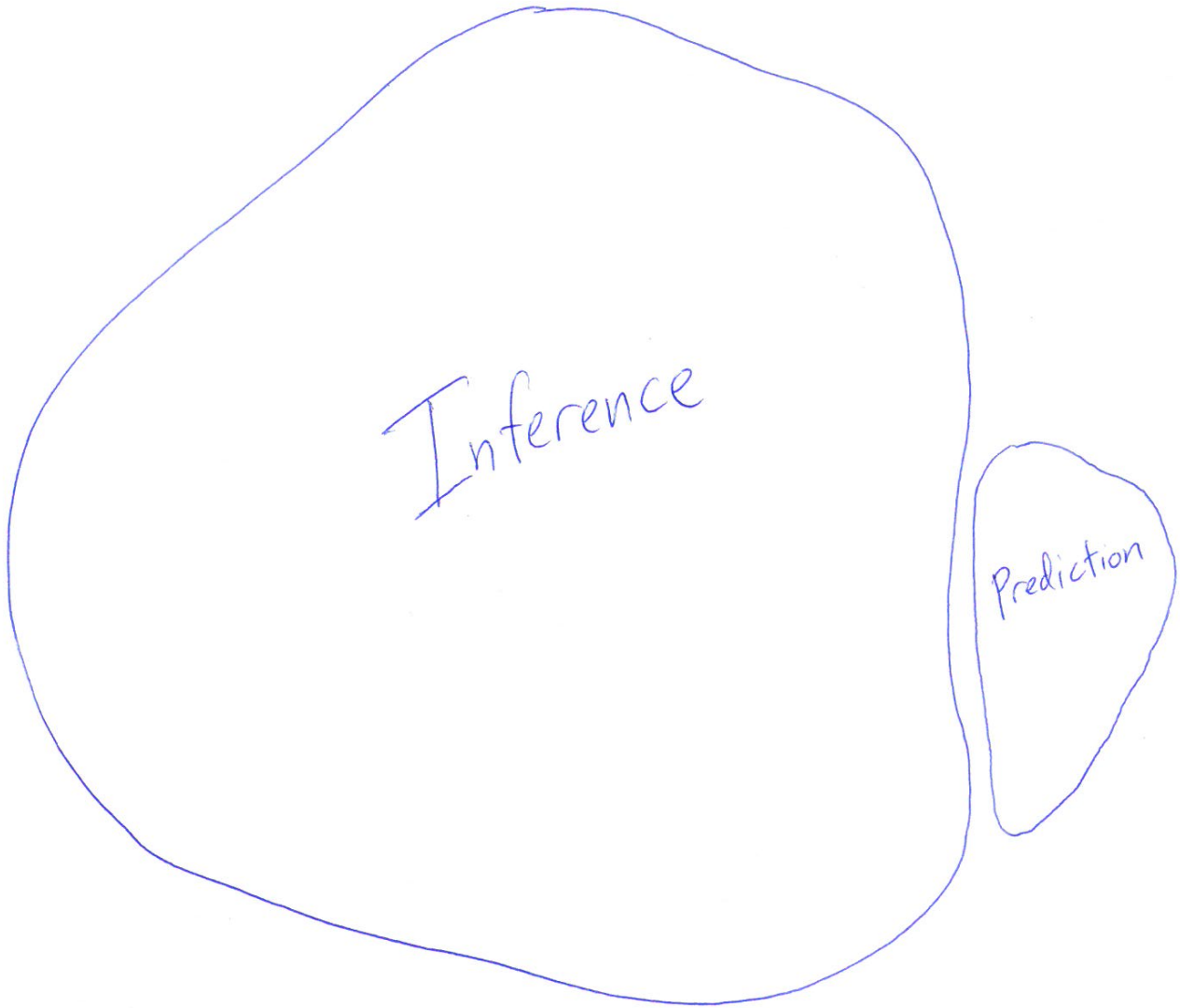
- If we have prediction accuracy as the ultimate goal in mind, then we conduct tuning parameter selection, model family selection, and finally make a model assessment.
- This corresponds to the
  - training, validation (or tuning) and test sets pipeline, or
  - training (cross-validation) and test set pipeline.

Goal 2: Data-driven Discovery

Examples:

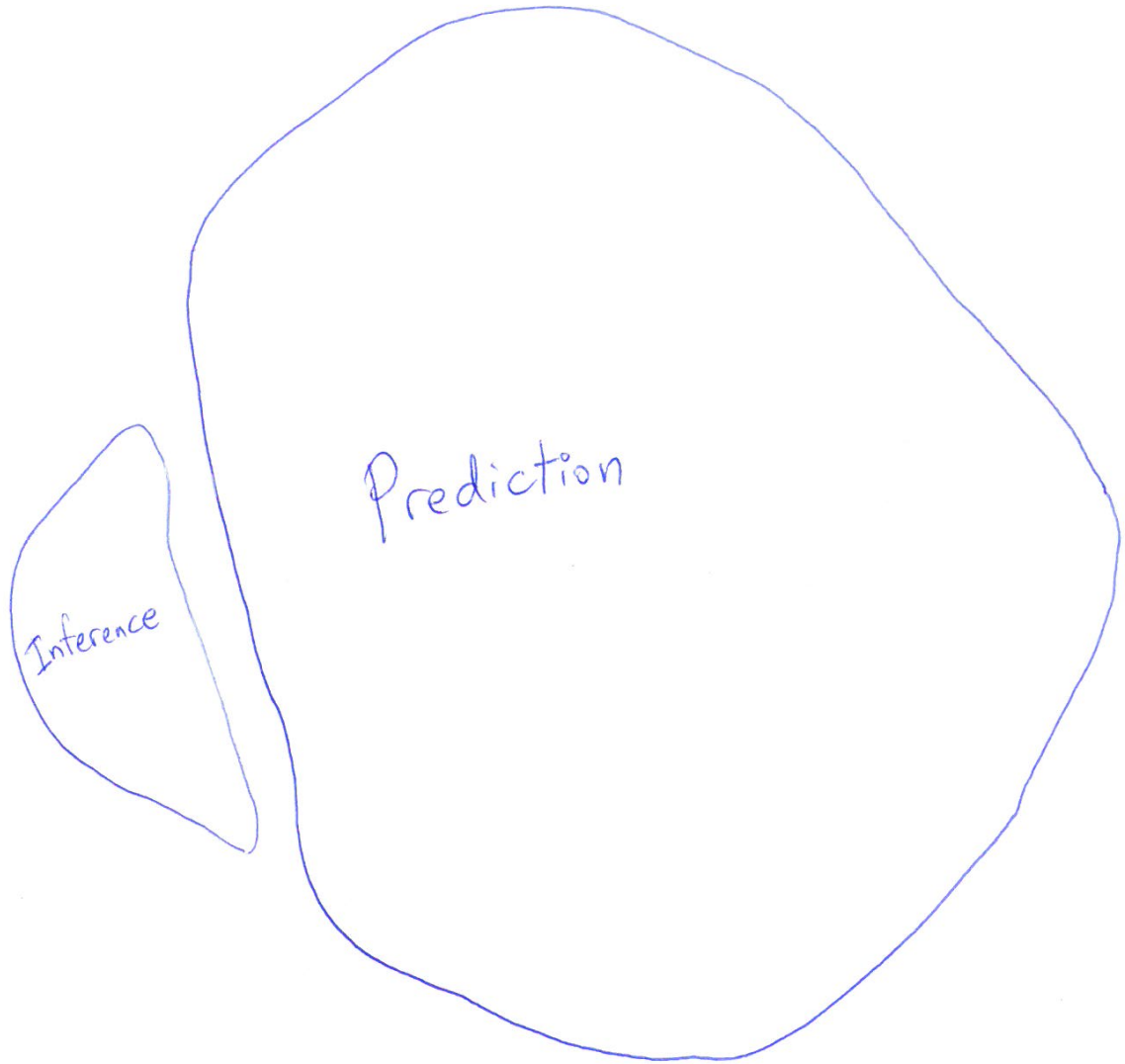
- Which genes cause or signify the presence of cancer?
- Are these identified genes really the cause? - Causal inference - Perhaps **corroborate via external information**.
- Are these genes identified because of random noise (error)? (In other words, are we in good luck.)
  - **Confirm via a separate test set**
  - **Statistical Inference**
- Is the **discovery stable**?
  - Multiple approaches discover the same thing.
  - Small changes to the data, algorithm, parameters, and etc. yield the same results.

Statistics versus Machine Learning



How statisticians see the world?

Statistics versus Machine Learning



How machine learners see the world?

Why inference is important?

- In many situations we care about the identity of the features e.g. biomarker studies: which genes are related to cancer?
- There is a crisis in reproducibility in Science: Ioannidis (2005) "Why Most Published Research Findings Are False"
- But part of the problem is statistical - we search through large number of models to find the best one; we don't have good ways of assessing the strength of the evidence

Classical vs modern-world inference

- Classical inference assumes the model is chosen independently of the data.
- Modern world practice: Use the data to select the model.
- This introduces additional uncertainty. Violates the assumption in classical inference.

A real data example

To illustrate with an example, we consider the `candy_rankings` data from the `fivethirtyeight` package. The outcome variable is how often a given candy won in popularity matchups against other candies, and the predictor variables are various properties like whether or not the candy has chocolate, whether or not its fruit flavored, how sugary it is relative to other candies, and so on. There are 85 candies and 11 predictor variables in the dataset.

We use model selection to pick a subset of the 11 predictor variables. Then we'll report significance tests for the selected variables to show they really are important. This is a fairly common sequence of steps for analyzing data in linear regression.

The result

```
> library(fivethirtyeight)
> data(candy_rankings)
> # Forward stepwise with AIC
> model <- step(lm(winpercent ~ ., candy), k = 2, trace = (
> # Significance tests for selected model
> print(summary(model)$coefficients, digits = 2)
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)          -0.25       0.31   -0.82   0.4169
chocolateTRUE         0.62       0.31    2.03   0.0460
fruityTRUE            0.52       0.29    1.78   0.0784
hardTRUE              0.41       0.27    1.48   0.1419
barTRUE              -0.98       0.32   -3.04   0.0032
pluribusTRUE         -0.47       0.24   -1.99   0.0502
```

* example given by Joshua Loftus

What really happened...

```
> library(fivethirtyeight)
> data(candy_rankings)
> candy_rankings$winpercent = rnorm(85)
> # Forward stepwise with AIC
> model <- step(lm(winpercent ~ ., candy), k = 2, trace = (
```

Instead of using the actual outcome variable from the data, I generated a random normal variable, totally independent of the predictors. And yet, we found a model with predictor variables that are significant.

Post-selection inference

- Classical tools cannot be used post-selection, because they do not yield valid inferences **(generally, too optimistic)**
- Selection bias can make noise look like signal.
- This is in the same spirit as "cross-validation cannot be used post selection (rather, selection should be embedded in cross validation.)"
- There is a clear incentive for scientific researchers to do this!
- Leo Breiman referred to the use of classical tools for post-selection inference as a "quiet scandal" in the statistical community.
- It's not often Statisticians are involved in such scandals.
- Post-selection inference: **Inference after having arrived at a statistical model adaptively, through examination of observed data.**

Example: Forward Stepwise Regression[1]

Classical inference for linear regression: 2 fixed nexted models.

- Model A: $M \in \{1, \ldots, p\}$
- Model B: $M \cup \{j\}$

Goal: test significance of the jth predictor. note: j is given

Compute RSS drop:

$$R_j = \frac{RSS_M - RSS_{M \cup \{j\}}}{\sigma^2} \text{ versus } \chi^2_1$$

[1] R. Lockhart, J. Taylor, Ryan Tibshirani, Rob Tibshirani (2014), "A significance test for the lasso", Annals of Statistics.

What will FS do?

- FS will choose variable j at each step to maximize R_j
- For any arbitrary j, $R_j \sim \chi^2_1$ under null .
- But the maximal possible R_j is stochastically larger than χ^2_1, even under null.

The naive (scandalous) approach in FS:

$$\max_{j=1}^{p} R_j \text{ versus } \chi^2_1$$

In general, for large p, this naive approach will yield significant result, whether or not the selected variable is important.

Why should we care?

We hate false discovery.

Crude Classification of PS Inference

- data splitting and data carving
  - root cause: same data used for selection and inference
  - solution: use some data for selection, the rest for inference
- high-dimensional inference
- simultaneous inference
- selective inference

Post-selection inference is an exciting new area. Lots of potential research problems and generalizations.

Data splitting: Cox (1975)

- Transparent justification. Even a nonexpert can appreciate.
- For example, it is customary in genetics to use one cohort to identify loci of interest and a separate cohort to confirm them.
- If imagine Y_1 is observed "first," then we proceed to analyze Y_2 as though model selection took place "ahead of time."
- Key assumption:

$$P_{M,H_0}(\text{reject } H_0 | (M, H_0) \text{ selected}) = P_{M,H_0}(\text{reject } H_0)$$

- Drawbacks:
  - half of the data for selection and half of the data for inference. Reduced power.
  - Sometimes, difficult to split data into completely independent parts (spatial and time series data.)

What does stable discovery entail?

- Multiple approaches discover the same thing.
- Small changes to the data, algorithm, parameters, and etc. yield the same results.

Example 1: Bootstrap

- Bootstrap is a tool for inference.
- Bootstrap = sampling n observations from a sample of n observations with replacement.
- Key idea: the newly sampled data sets are identically distributed (following the empirical distribution, which is consistent with the true distribution.)
- For each bootstrapped sample, compute the coefficient estimates, then we have an idea about the distribution of the coefficient estimate. This can be used for inference.

Example 2: Stability selection

Meinshausen, N. and Buhlmann, P. (2010), "Stability selection". Journal of the Royal Statistical Society: Series B (Statistical Methodology)

- Tuning parameter selection from cross-validation seems to be flawed, or troublesome at least.
- Lasso and others have stringent conditions for sparsistency.
- Instead, select those variables which are constantly selected.
  - Subsample with $n/2$ obs. Do Lasso. Compute the whole solution path. Repeat M times.
  - For each variable, each λ, compute the proportion of times the variable is selected, denoted as $P_{j,\lambda}$.
  - For each variable, compute $Q_j = \max_\lambda P_{j,\lambda}$ along the path.
  - Find a cutoff point. Select variables with Q_j large enough.
- Q_j is independent of the value of the tuning parameter λ.

Example 3: Classification stability

Sun, W., Qiao, X. and Cheng, G. (2016), "Stabilized Nearest Neighbor Classifier and Its Statistical Properties," Journal of the American Statistical Association.

- Classification Instability (CIS):

$$\mathsf{E}_{\mathcal{D}_1,\mathcal{D}_2}\mathsf{E}_X(\hat{f}_{\mathcal{D}_1}(X) \neq \hat{f}_{\mathcal{D}_2}(X))$$

- Incorporate CIS as a second goal (to classification accuracy)
- Closed form solution for stabilized nearest neighbor classifier.

The next section would be . . . . . .

1. Two Goals of Statistical Learning
2. Best Practices in Data Analysis

Best Practices

- Plan ahead (& think validation) - how many observations - how to split - inference needed?
- Always visualize to help shape effort - do techniques make sense - assumptions?
- Use multiple techniques - no single method works for all scenario - stable discovery.
- Communicate uncertainty - where is it - what are the source - could the results be artifact - how to mitigate - how to quantify uncertainty
- Make your analysis reproducible - save random seeds - good documents

Case 1

A scientist has measured expression levels for 10,000 genes on 180 patients. He is interested in building a regression model based upon a subset of the genes to predict a continuous clinical response and decides to use the lasso. He implements the following procedure to build his model:

1. Filter the genes down to the top 2,000 that exhibit the highest correlation with the response.
2. Employ 5-fold cross-validation on the 2,000 genes to select and report the prediction error.

Questions:

a. Will this procedure result in an unbiased estimate of the prediction error?
b. What would you change in this procedure? Why?
c. Write out the steps of a new procedure.
d. Suppose the scientist wants to work with only 2,000 genes. What is another way to filter the genes?

Case 2

A researcher is illustrating a new prediction method (the method has one tuning parameter) on proteomic data with $n = 35$ observations and $p = 3200$ features. He is comparing his method to the lasso, adaptive lasso, and ridge.

a. Devise a scheme that will allow the researcher to fairly compare these methods in terms of prediction error and that also makes the best use of the limited data. Write this scheme out explicitly and explain why this will result in correct estimation of the prediction error and a fair comparison between methods.

b. Suppose the researcher was also interested in assessing the accuracy of the variables selected by each method. What would you suggest? Explain.

Case 3

A scientist running a big cohort study is interested in discovering new biomarkers predictive of disease status. The cohort has n = 1500 subjects and collects measurements on p = 400,000 biomarkers. The scientist wants to try all the state-of-the-art machine learning and feature selection methods to make the best possible predictions.

a. Devise a scheme to set up the analysis pipeline. Write out this scheme explicitly.

b. The scientists knows in advance that age, gender, and education are very important predictors of disease status. But, the cohort has very uneven distributions of especially age and education levels. Should this knowledge change your pipeline form part (a)? How?

c. After following the proposed pipeline, the scientist settles on an optimal model family and tuning parameter and reports the prediction error. Then, she goes back and re-fits the optimal model to the entire cohort. But in doing so, she notices that the biomarkers selected by following the pipeline and the biomarkers selected on the whole cohort are different. Why? How can she reconcile this? Are there other procedures that you would recommend?