Lecture 6

IDS575: Statistical Models and Methods
Theja Tulabandhula

Notes derived from the book titled "Elements of Statistical Learning [2nd edition] (Sections 8.2)

1 Beyond Supervised Learning

There are very interesting and realistic ways in which statistical models and methods influence several fields. We have seen such examples in previous lectures. Here is one more:

Example 1. A very visual example of the potential of using statistical models is illustrated in a webapp, made by a few companies (Stoj, UseAllFive, Google).

The setting is supervised learning. It happens that they use a certain family of models called neural networks (instead of the familiar LDA and logistic regression) to perform a classification task. The number of classes is 3. You need to generate visual input variable realizations (images!) using the browser/webcam. Once trained on examples from each class, the classifier can classify any new visual input to one of the three classes. Each prediction triggers a gif to be played, giving a cool overall effect.

We will now look at the general landscape of statistical modeling and a couple of key methods to model and *infer*.

Previously:

- We have been studying the supervised learning task. This involved minimizing the squared loss or the *cross-entropy* loss (this is defined for a classification model as $\sum_{i=1}^{N} \log Pr_\theta(G = g_i|X = x_i)$, where θ is the parameter of the model. Classification is made at a new point x_0 as $\arg\max_{k=1,..,K} Pr_\theta(G = k|X = x_0)$).
- We connected these loss functions the method of maximizing likelihood estimation (MLE).
- The MLE method (and others) are applicable to statistical problems beyond supervised learning.

So, we will look at a couple of these general statistical problems and methods that solve them.

In general, in a statistical modeling task, the inference problem is to estimate either the unknown parameters of the model or a distribution over them. This is called *inference*.

Before discussing inference, lets look at a tool called the *bootstrap*, a straightforward process to estimate prediction uncertainty.

2 The Bootstrap for Capturing Prediction Uncertainty

The bootstrap is a general process for capturing prediction uncertainty, used throughout statistics. Consider data $\mathbf{Z} = (z_1, ..., z_N)$. For instance, this can be training data ($z_i = (x_i, y_i)$). The idea is to sample new datasets with replacement from Z, where each new dataset is of size N (there are variations where this is not necessary). Let the number of new datasets is B. This is illustrated in Figure 1.

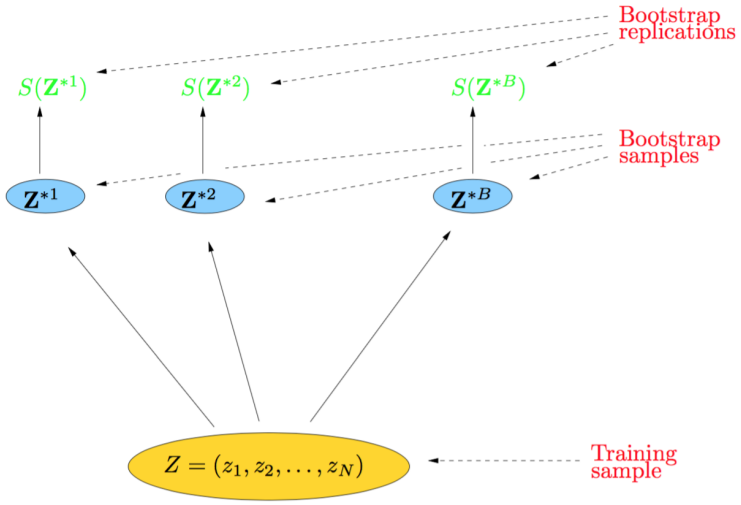

Figure 1: The bootstrap process. The new datasets are denoted $\mathbf{Z}^{*b}, b = 1, ..., B$.

We fit a model to each of the dataset and look at the performance of the models across the B replications. Let $S(\mathbf{Z})$ be a function of data (for example, it can be the linear regression model making a prediction at a fixed point x_0).

Using bootstrap, we can estimate any aspect of the distribution of $S(\mathbf{Z})$. It is like a *Monte-Carlo estimation*[1] of any function of $S(\mathbf{Z})$ under sampling from the empirical distribution function for data $\mathbf{Z} = (z_1, ..., z_N)$.

Below, we list two example ways to quantify prediction uncertainty.

Example 2. How do we use the bootstrap for say *estimating Err*? We can do the following:

- For each observation, we record the predictions from bootstrap samples that do not contain this observation.
- Then we average over the performance of these predictions as: $\frac{1}{N}\sum_{i=1}^{N}\frac{1}{C^{-i}}\sum_{b\in C^{-i}} L(y_i, \widehat{\mathrm{f}}^{*b}(x_i))$.

Here, C^{-i} is the set of bootstrap samples b that do not contain observation i, and $|C^{-i}|$ is their number. $\widehat{\mathrm{f}}^{*b}(x)$ is the model fit using the b^{th} bootstrap sample.

Just like cross-validation, training size may impact how well the bootstrap estimates *Err*.

Example 3. How do we use the bootstrap to get *confidence bands* around predictions? We can do the following:

- For each bootstrap sample, fit the prediction function.
- Get the 95% (in general $1-\alpha\%$) point-wise confidence band from the percentiles at each x. Say, for a 95% with $B = 200$, the $2.5\% \times 200$ percentile on each side would correspond to the fifth largest and smallest value at each x.

See Figures 2 and 3 for an illustration. Here, $N = 50$ and 7 pre-determined basis functions of the original 1-dimensional input variable were used. The model is nothing but a vector of coefficients $\widehat{\beta}$ that linearly combines the basis functions.

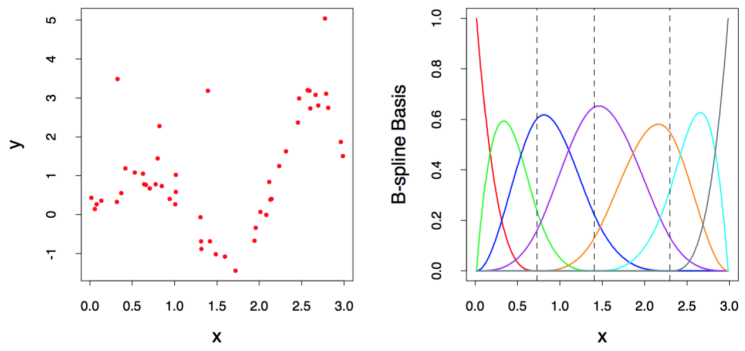

Figure 2: Dataset and the basis functions.

Thus, the bootstrap gives a direct computational way to assess prediction uncertainty.

[1]Monte-Carlo estimate corresponds to taking an average of a bunch of points. We will discuss this later.

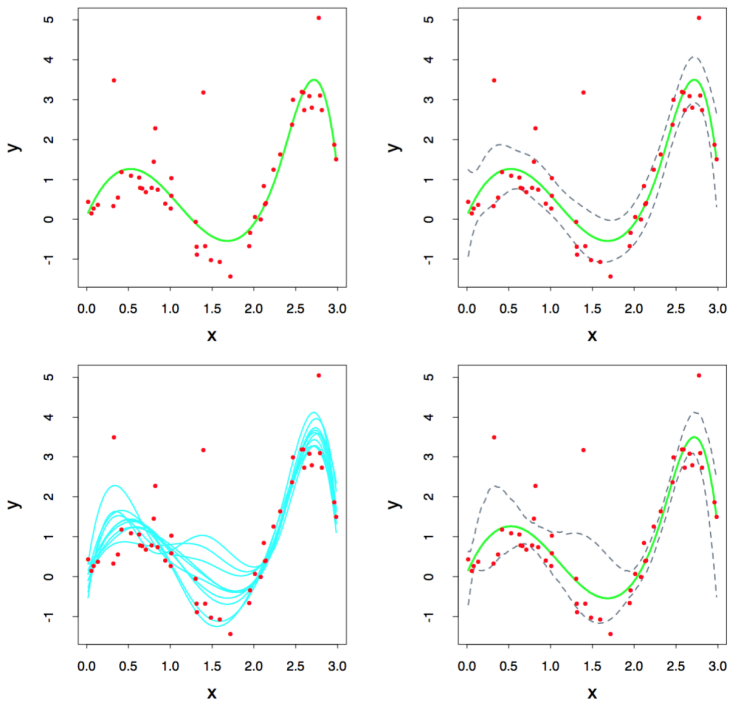

Figure 3: Confidence intervals ($\pm 1.96\times$ standard error) using linear regression (top right) and confidence bands using the bootstrap (bottom right).

3 Inference using MLE

Now, lets focus on a general statistical task: say $Z \sim g_\theta$. That is, there is a *parameter* vector θ that specifies a distribution function (more precisely, the density function) for the random variable Z.

Example 4. If $Z \sim N(\mu, \sigma^2)$, then $\theta = [\mu, \sigma^2]^T$.

The likelihood function is $L(\theta; \mathbf{Z}) = \prod_{i=1}^{N} g_\theta(z_i)$, which captures the probability of observed data under model g_θ. We maximize the likelihood, keeping $\mathbf{Z}$ fixed and varying θ. While maximizing a function, taking its log does not change the maximum point, so we maximize $l(\theta; \mathbf{Z}) = \sum_{i=1}^{N} \log g_\theta(z_i)$.

Say the maximum is attained at $\widehat{\theta}$. We can create a confidence bands for this estimate as well. How? With our familiar tool: the bootstrap!

Note 1. Just like in cross validation, choices should be made for each bootstrap sample separately, to get the right confidence bands.

4 Summary

We learned the following things:

- Learning problems that are broader in scope than supervised learning.
- Learning/inference in these setting can be achieved using maximum likelihood.

A Sample Exam Questions

1. What are the different ways to use the bootstrap in a statistical modeling task?

B List of Topics

1. Linear models
2. k-nearest neighbors
3. Comparison between nearest neighbor and linear methods
4. Statistical decision theory
5. Curse of dimensionality
6. Supervised learning
7. Linear regression
8. Bias-variance trade-off
9. Subset selection
10. Cross validation
11. Ridge regression
12. LASSO
13. Classification using linear regression
14. Linear discriminant analysis
15. Logistic regression
16. Model selection
17. Maximum likelihood method
18. Bootstrap