Sampling & Resampling

Handout 3 of Introduction to Machine Learning

January 2020

Problem #1:

You cannot measure all the things (in the population).

Solution #1:

Take a _____.

Problem #2:

Your sample could be biased.

Solution #2:

Take a _____ sample.

Problem #3:

But now you have to worry about sampling _____.

Why worry? Because it affects how well your model can make accurate predictions.

Solutions:

1.
2.
3.

Let's say *more flexible models* is our only option. We cannot start over; we cannot get more data. Now we have two new problems!

Problems:

1. Less flexible models tend to have more _____ (they may miss more noise, but at the cost of missing some signal too)

2. More flexible models tend have more _____ (they may catch more signal, but at the cost of picking up too much noise)

Why worry?

1. It is really hard to do anything with model _____ (once the signal is gone, it is gone)...

2. Model _____, on the other hand, we have methods for (because the signal is there, and the noise is estimable)

Solution:

Bootstrapping: resampling from our _____ data with _____.

Problem:

But now you have many (re-)samples, and many models.

Solution:

Model _____.