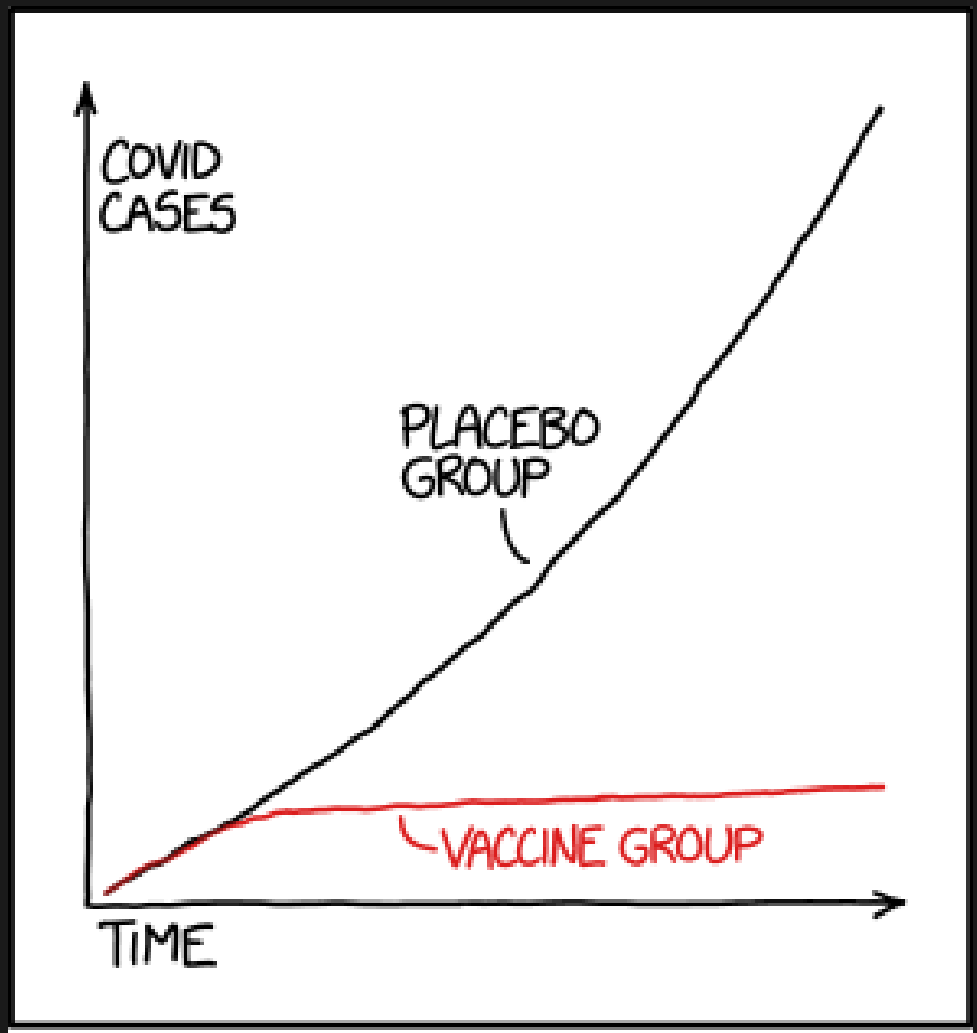
Cross Validation

Daniel Lawson University of Bristol

Lecture 02.2 (v4.0.0)



STATISTICS TIP: ALWAYS TRY TO GET DATA THAT'S GOOD ENOUGH THAT YOU DON'T NEED TO DO STATISTICS ON IT

https://xkcd.com/2400/

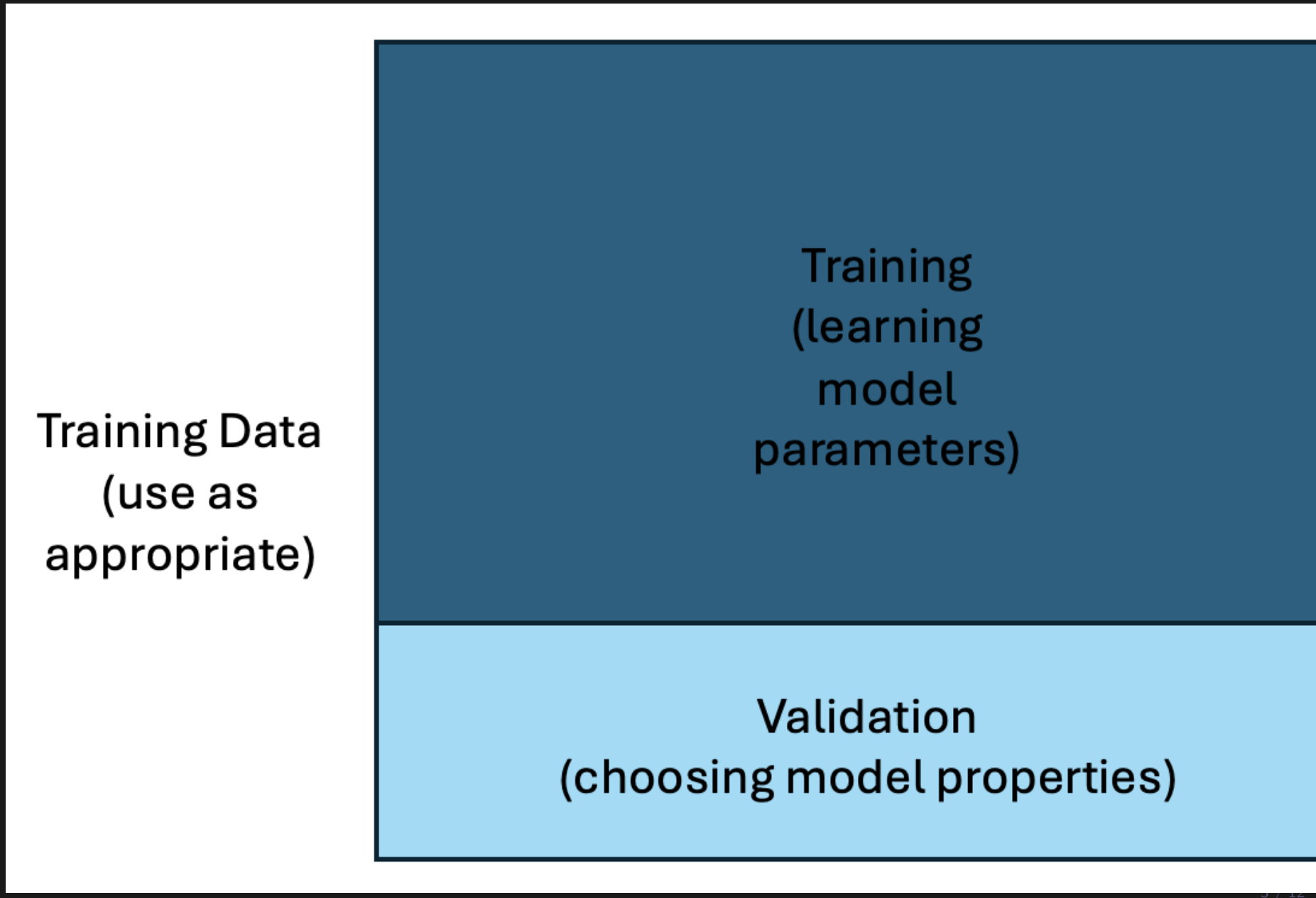
Signposting

- This lecture covers Cross Validation:
  - why you should always do it
  - What it looks like

Model Selection

- Imagine that we have run two different inference procedures (models) on our data.
- We want to decide which of these gives the **best** description of the data.
  - (we might pretend we want to know which one is **right**...)
- Model selection formalises how to make this assessment.
  - Focus on cross-validation: performance in held out data

Data distribution



K-fold CV

- **Split** the training data into k "folds" $f(i)$, that is, **random non-overlapping samples** of the data of size n/k. Then:
- **For each fold** i:
  - Call $X^{-(f(i))}$ the "training" dataset and $X^{(f(i))}$ the "validation" dataset
  - Learn parameters $\hat{\theta}_i$ with data $X^{-(f(i))}$
  - Evaluate $l_i = \text{Loss}(X^{(f(i))}|\hat{\theta}_i)$
- And report $\frac{1}{n}\sum_{i=1}^{k} l_i$

How to use the data?

- We can make many different CV splits of our **training** data
- To evaluate variance etc
- We can choose model structure, even do model choice with it
- When we look at the test data, that is it: report performance and stop
- Test data can be used to score, not to influence, the algorithm

How many folds?

- k-fold CV loses a fraction of the data
  - A **bias-variance tradeoff**: we introduce a bias (towards simpler models) but also significantly reduce the variance of the MSE estimation.
  - This means that (under the assumption that the **true model is not in the model space**) k-fold CV will choose a **simpler model** with less predictive power than was possible.
- However, smaller k can make the inference more consistent across different data.
- For **larger data**, $k = 10$ is often chosen:
  - **cost**. k defines the minimum number of times you need to run the models. If you can afford to run a model once, you can probably afford 10 times.
  - **practicality**. If you had only 10% more data you might expect to get the same performance as using validation data for training. We frequently lose this amount of data to quality control, etc.

Handling correlation

- **Correlation** structures can be handled in k-fold CV by **careful sampling**:
  - a-priori there is a correlation in time or space expected. we can therefore **remove windows**.
  - the data have some associated covariate, which can be removed en-masse.
  - empirical correlation structures can be used to select a point i and all points correlated with it above some **correlation threshold**.
- Some of these can be used in other contexts. Examples include:
  - **block bootstrap**
  - Using a different definition of a "datapoint" in a leave-one-out context, for example: datapoints are countries instead of countries at timepoints

General considerations

- To make Cross-Validation work, we need to be able to define our inference goal cleanly. Some scenarios:
  - **Same source, single datapoint**: Within a single datastream, how well can we predict the **next** point?
  - **Same source, segment of data**: Within a single datastream, how well could we predict everything that happens within an hour?
  - **New but understood source**: We have multiple datastreams, each of which might be different but all are generated by a similar process. How well can we predict a new such datasource?
  - **Unexpected source**: We have many classes of datastream. How well can we predict what would happen on a new class of datastream?

Reflection

- You should understand how to:
  - Define and use a null hypothesis significance test,
  - Contrast classical and resampling tests, and judge appropriate uses,
  - Use statistical testing appropriately in projects.

Further reading

- Cross Validation
  - Chapters 2.3 and 7.10 of The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Friedman, Hastie and Tibshirani).
  - Cosma Shalizi's Modern Regression Lectures (Lectures 20, 26)