Introduction to Machine Learning

Day 4: Cross Validation for Real

Joanna Bieri DATA301

Day 4

Important Information

- Email: joanna_bieri@redlands.edu
- Office Hours: Duke 209, Click Here for Joanna's Schedule
- Class Website

The Four Things From Today

1. **One split is not enough.** The same data, split ten different ways, chose alphas from 0.018 to 17.783.
2. **K-fold** gives you k scores, so you can see the **spread**, not just one number.
3. **Leakage** is anything that learns outside the pipeline. It does not look like a mistake.
4. **Stratify** when the thing you care about is rare.

Get out your handwritten notes!

One Split Lies

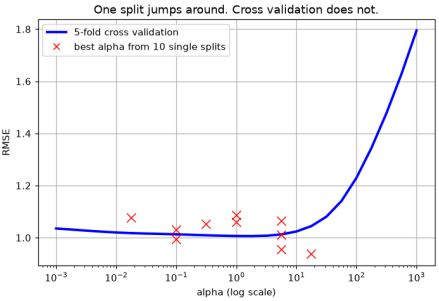
We Split The Same 200 Points Ten Ways

Each time we ran the whole alpha search from scratch. The only thing that changed was which 60 points landed in the validation set.

smallest alpha chosen	**0.018**
largest alpha chosen	**17.783**
RMSE across all ten	0.936 to 1.086

The **choice** moved by a factor of 1000. The **score** barely moved at all.

Discuss

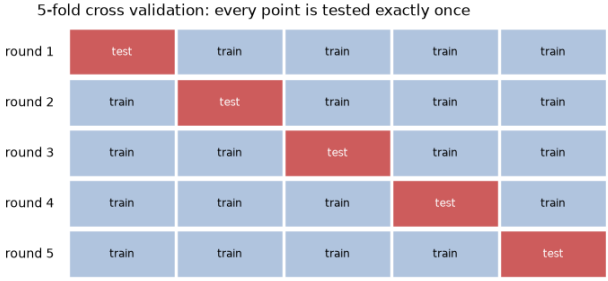


Blue is 5-fold cross validation. Each red x is where one single split thought the best alpha was.

- **Q4b.** The blue curve is nearly flat from alpha 0.01 to 10. What does that tell you about how much your exact choice of alpha matters?

K-Fold

Every Point Gets Tested Exactly Once



Our Five Folds, Same Model, Same Data

fold 1	fold 2	fold 3	fold 4	fold 5
0.860	1.094	0.987	1.082	1.010

mean **1.007**, standard deviation **0.084**

Best fold to worst fold is **0.23 apart.**

Discuss

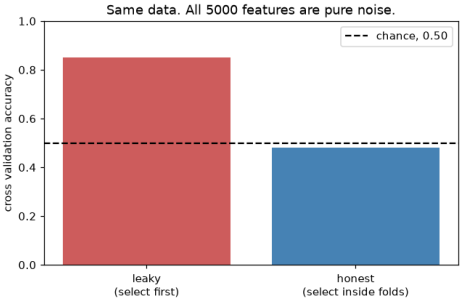
Those five folds ran from **0.860 to 1.094**, with a mean of **1.007**.

On Day 3, two different Elastic Net settings scored **0.962** and **0.970**, and we treated that as a difference.

- **Q3d.** Write one sentence you could honestly say to a client about this model, using both the mean and the range.
- Was that Day 3 difference of 0.008 real?

Leakage

5000 Features Of Pure Noise, Coin Flip Labels



We picked the 20 features most correlated with the labels, then cross validated. **0.850 against 0.480**, on data containing nothing at all.

The Rule

If a step looks at `y`, or computes any statistic from `X`, it **learns**.

The tell is `.fit()`. If a step has one, it belongs inside the pipeline.

Discuss

You have 50 patients. Each one had 10 scans taken, so your table has 500 rows. You split those 500 rows randomly into training and test.

Nothing in your code touches the test set. Your scaler is inside the pipeline. Your cross validation is set up correctly.

- **Q6d.** Does this leak? If so, what can the model do that it should not be able to do, and how would you split it instead?

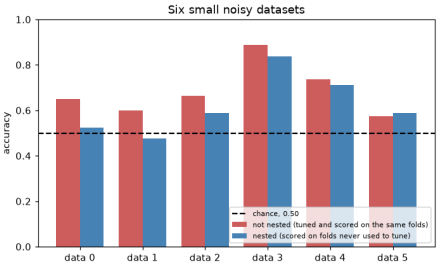
Stratification

10 Positives Out Of 200, Split Five Ways

Plain `KFold` gave **[3, 0, 3, 3, 1]**. Fold 2 got **zero positives**, so the model is tested on 40 patients and none of them have the condition. Answering "no" every time scores 100 percent there.

- **Q7a.** Explain to a doctor, in one sentence, why that 100 percent should worry them.

Discuss



Red tunes and scores on the same folds. Blue scores on folds the tuning never saw. Dataset 1 claims **0.6000** and scores **0.4750**, worse than a coin flip.

- **Q8b.** What should you do with that model?

Wrap-Up

The Workflow From Here On

1. Split off a test set **first**. Stratify it. Leave it alone.
2. Build a **pipeline**, with every step that learns inside it.
3. Tune with `GridSearchCV`, and `StratifiedKFold` for classification.
4. Look at the **spread**, not just the mean.
5. Score on the test set **once**, at the very end.

Before Next Class

1. **Weekly Homework 2, due Sunday 9/13 at 11:59pm.** Covers Day 3 and Day 4.
2. Read Geron ch. 3, **Performance Measures**.
3. Watch the Day 5 video.

Day 5

A model that answers "no condition" to every patient scored **100 percent** on that zero-positive fold.

86.5 percent of the wines in your homework are "not good", so always answering "not good" is 86.5 percent accurate.

Accuracy can hide almost anything. Next class we fix that.