One-Hot-Encoding and Cross-Validation

TM Quest

Overview

What Will we Learn in This Module?

- What is one-hot-encoding?
  - How to handle categorical data as features.
  - What are the benefits with one-hot-encoding?
  - How to fit one-hot-encoding into a scikit-learn pipeline.
- What is cross-validation?
  - How to not waste data.
  - How to compare models and choose the best.
  - How to ensure that the performance of the final model is representative.

One-Hot-Encoding

Motivation

```
# Converting the categories into numerical values

tips["is_weekend"] = tips["day"].replace({"Thur": 0,
    "Fri": 0, "Sat": 1, "Sun": 1})

tips["is_dinner"] = tips["time"].replace({"Lunch":
   0, "Dinner": 1})
```

Problems with our previous method

- Can not handle an increase in categories.
- A lot of code.
- Difficult to fit into a pipeline.

One-Hot-Encoding

Old Feature	New Features			
Weekday	Thursday	Friday	Saturday	Sunday
Thur	1	0	0	0
Sat	0	0	1	0
Fri	0	1	0	0
Thur	1	0	0	0
Sun	0	0	0	1
Sun	0	0	0	1

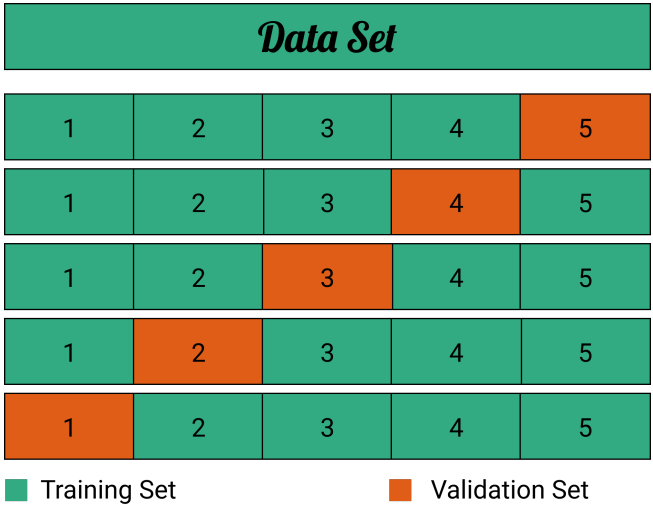
Cross-Validation

Cross-Validation

Old-method

- We are not using the testing set in our training!

Cross-Validation



- Divide the dataset into equal parts.
- Choose a part and train on the rest.
- Predict the final part and find the error.
- Choose another part and repeat.
- Go through all the parts.
- The final error is given by averaging.

Validation and Test Set

Validation and Test Set

Validation Set

Validation set is used to estimate the error of one model.

Problem when comparing models

- Let us say that we have several different models, and we choose the one with the lowest error.
- The more models we compare, the higher the probability that the chosen model just got lucky.
- Hence the probability of the estimated error being much smaller than the actual error becomes high.
- Hence we introduce the test set.

Validation and Test Set

Test Set

We set off some part of the dataset in the beginning to find the final error. This part is called the test set.