Introduction to Machine Learning
Lecture 7: Bias-Variance Decomposition
Train-Validation-Test Framework

Harbour.Space University
February 2021

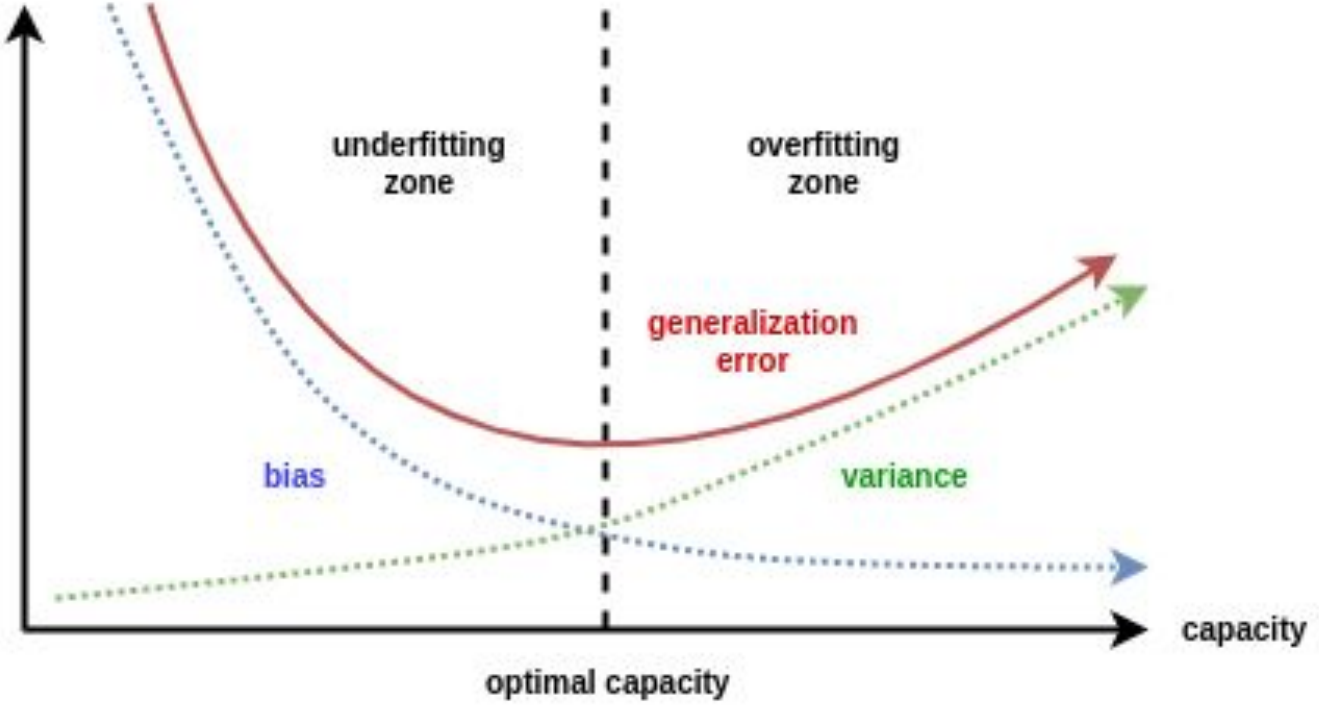
Iurii Efimov

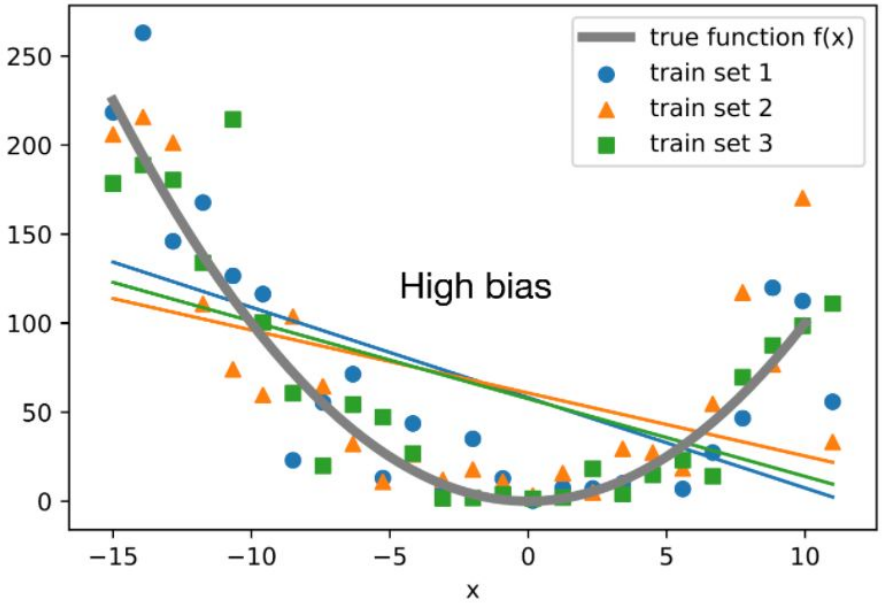
Outline

1. Bias-variance decomposition intuition
2. Validation strategies
3. Feature preprocessing

Bias-variance decomposition intuitions

Bias-variance tradeoff





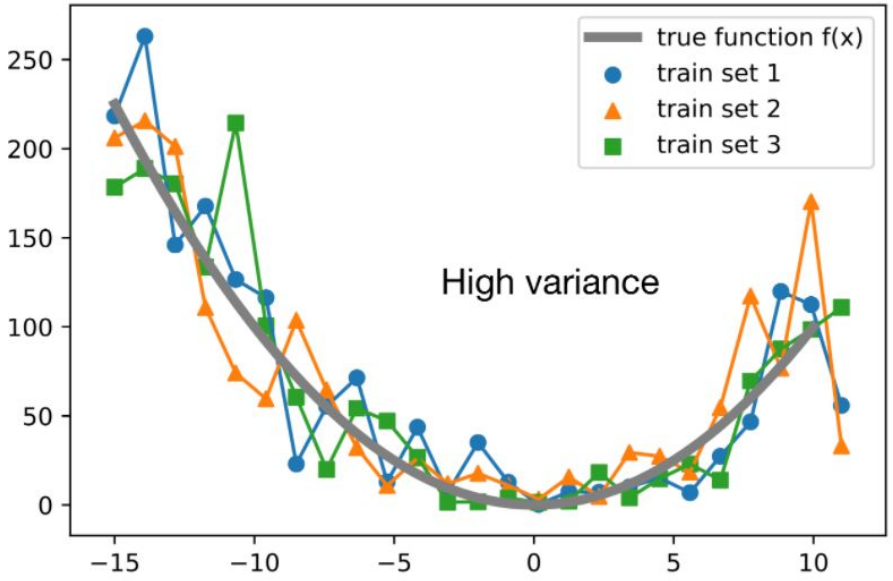
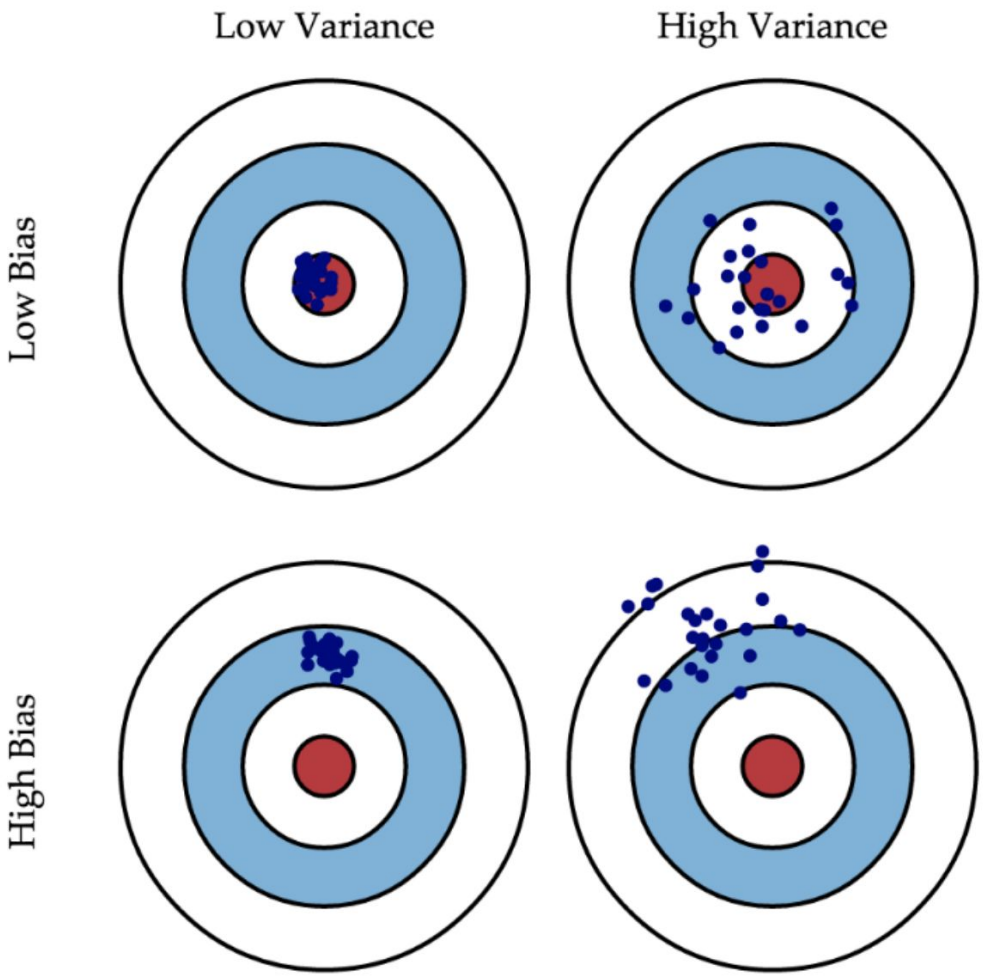
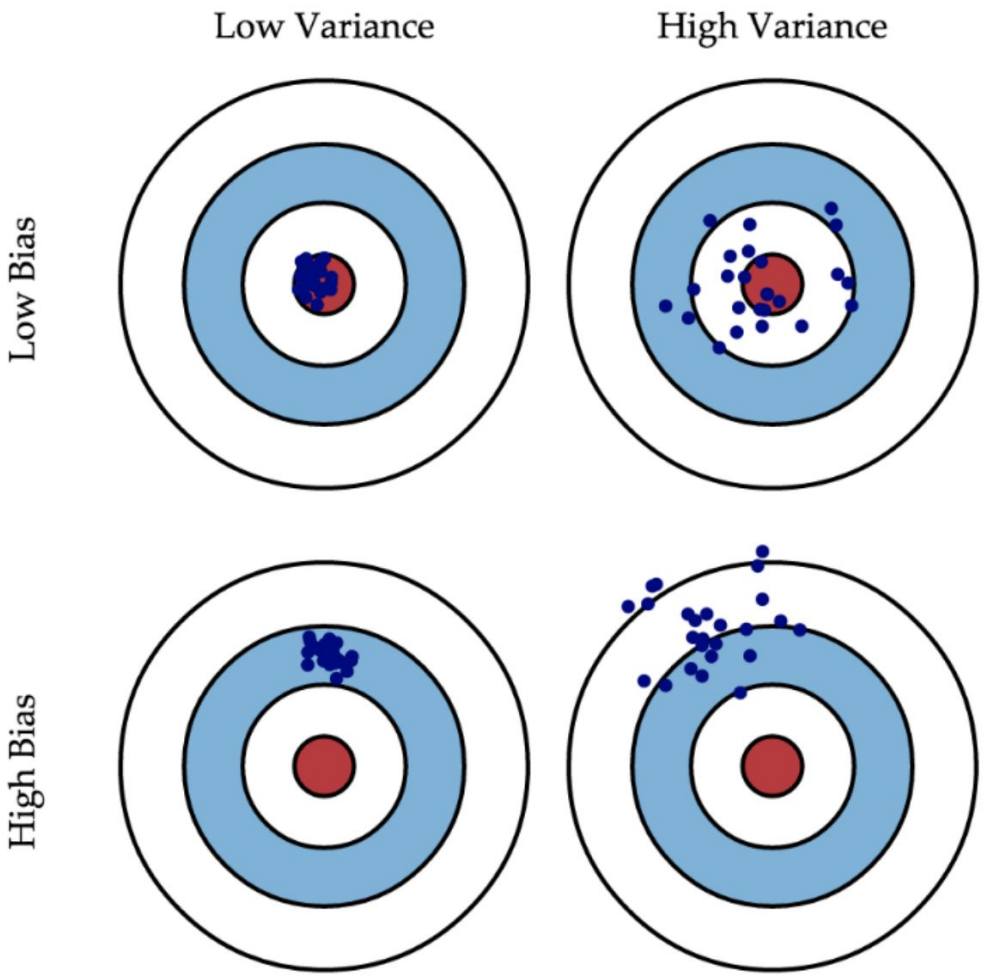


image credit


Low Variance
High Variance
Low Bias
High Bias

Like unicorns, most likely they do not exist

Random Forest (sometimes)



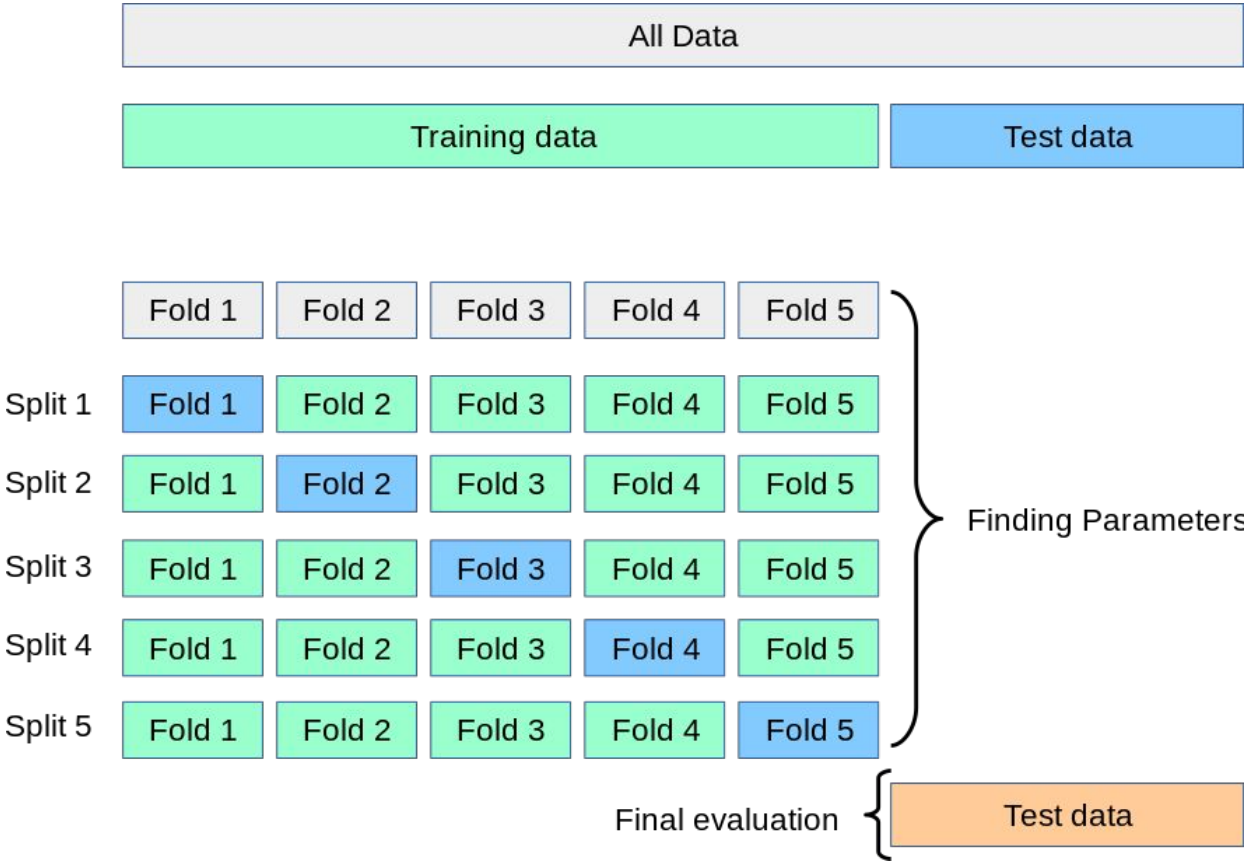
Gradient Boosting
Neural Networks
KNN

Linear models
Decision stomps
Over-regularized models

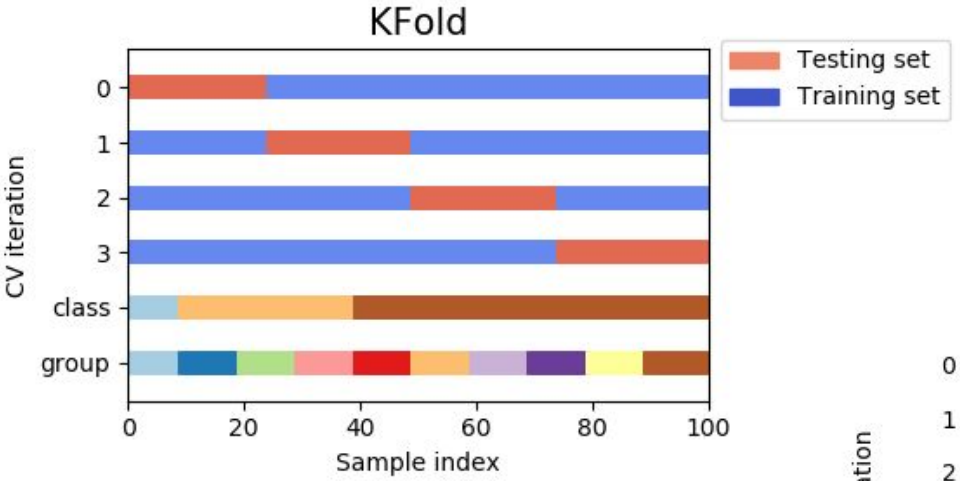
Underfit models

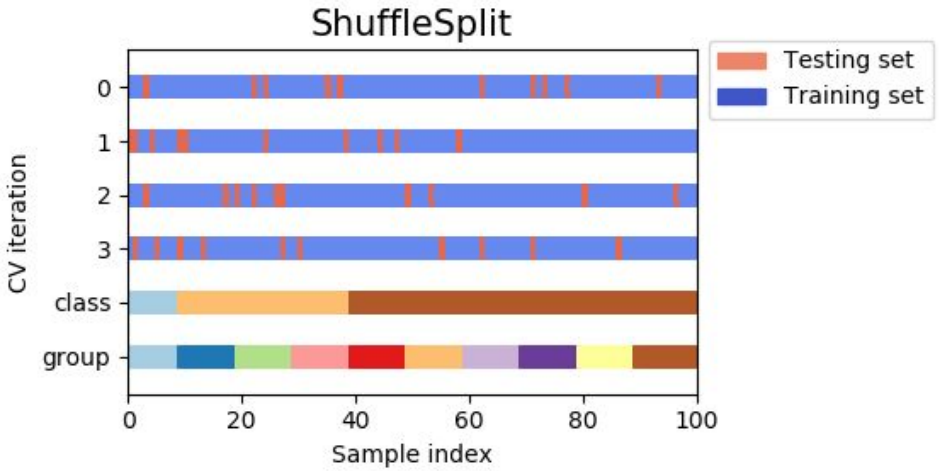
Validation strategies

Validation strategies



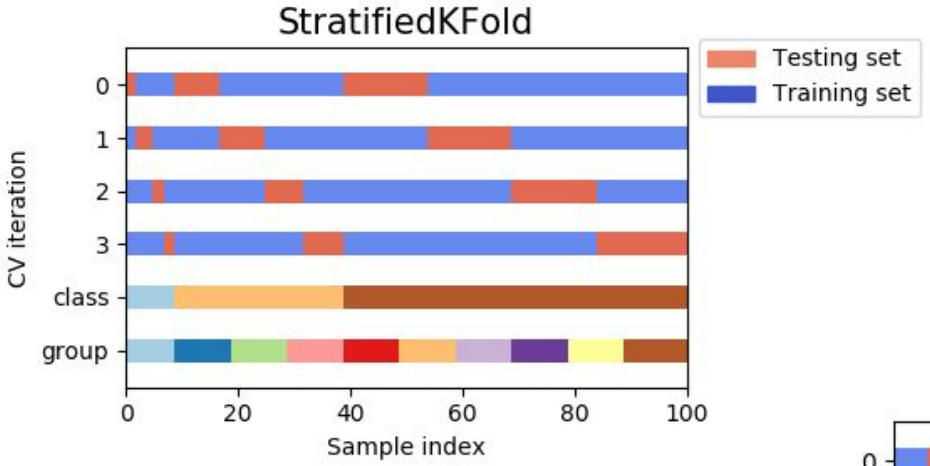
Validation strategies

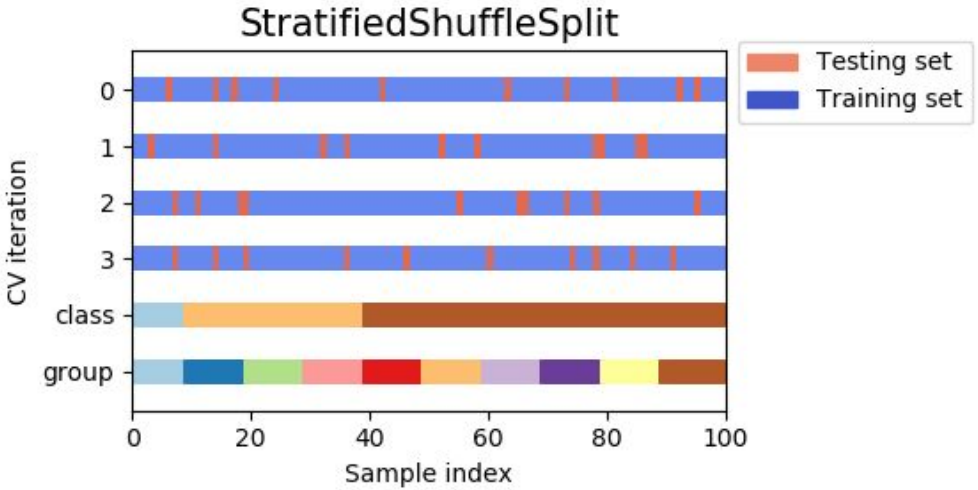




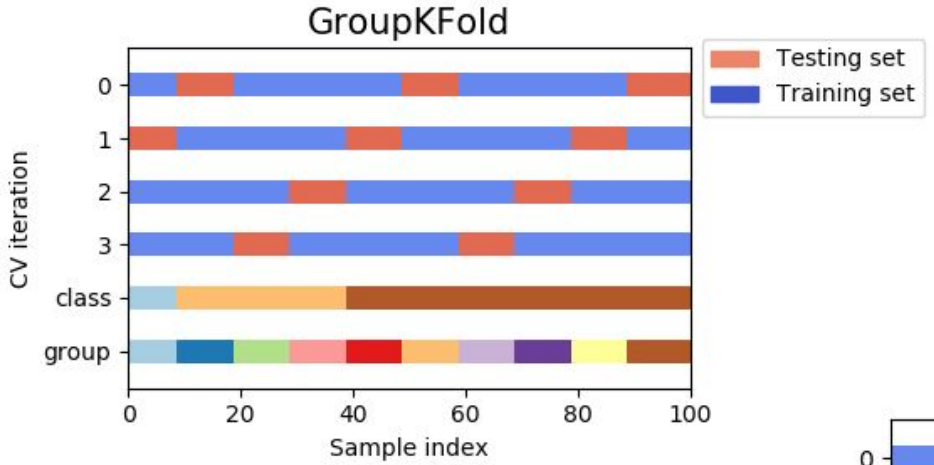
Special case: Leave One Out (LOO) - good for tiny datasets

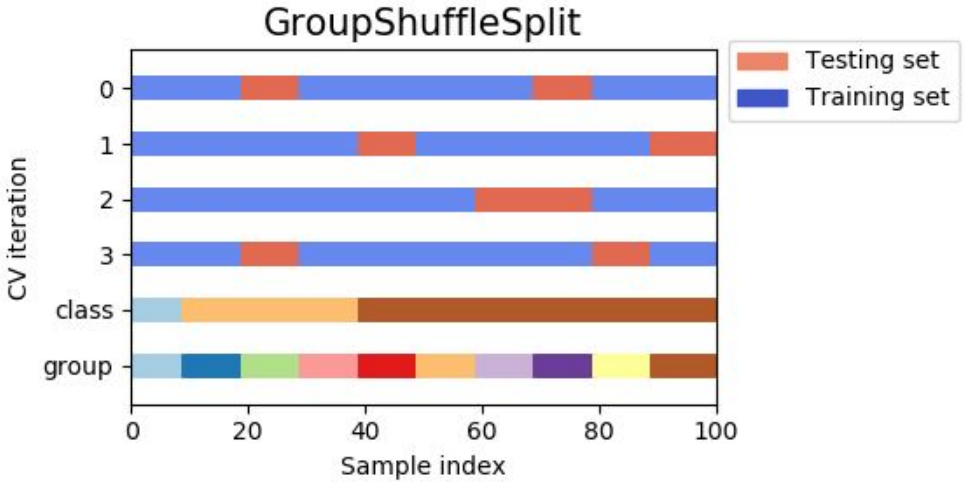
Validation strategies



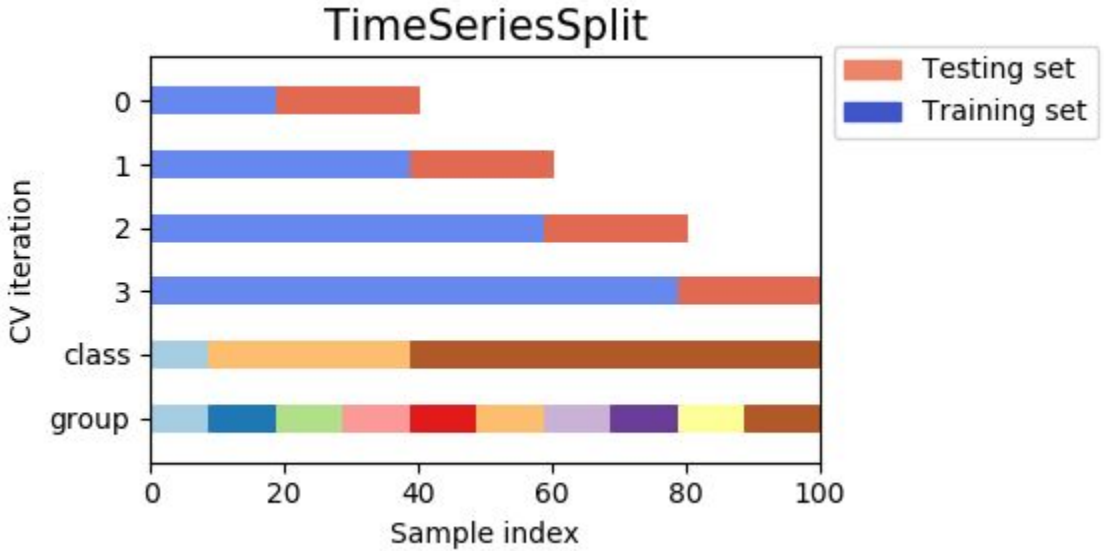


Validation strategies





Special case: time series



Never use `train_test_split` in this case!

Feature preprocessing

Feature Types

- Numeric (e.g. height) → use as is, maybe normalize
- Binary (e.g. physiological gender) → use as is
- Categorical (e.g. eye color) → one-hot encoding
- Ordinal (e.g. birth month) → one-hot encoding / as-is

Data Normalization

- Standardization

$$\{f_i\}_{i \in I} \rightarrow \left\{ \frac{f_i - mean\{f_t\}_{t \in I}}{std\{f_t\}_{t \in I}} \right\}$$

- Min-Max Normalization

$$\{f_i\}_{i \in I} \rightarrow \left\{ \frac{f_i - min\{f_t\}_{t \in I}}{max\{f_t\}_{t \in I} - min\{f_t\}_{t \in I}} \right\}$$

- Max Scaling

$$\{f_i\}_{i \in I} \rightarrow \left\{ \frac{f_i}{max\{|f_t|\}_{t \in I}} \right\}$$

Data Normalization

```
from sklearn import preprocessing as pp
```

- Standardization

```
scaler = pp.StandardScaler()
```

- Min-Max Normalization

```
scaler = pp.MinMaxScaler()
```

- Max Scaling

```
scaler = pp.MaxAbsScaler()
```

```
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

Thanks for your attention!