https://ttpoll.com/p/747972

COSC 325: Introduction to Machine Learning

Dr. Hector Santos-Villalobos

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE

Lecture 15: Hyperparameter Tuning and Model Evaluation

THE UNIVERSITY OF TENNESSEE KNOXVILLE

https://ttpoll.com/p/747972

Class Announcements

Homework

Homework #5 due 11/06
Homework #6 due 11/13

Course Project:

Midterm Report due 10/27

Lectures:

Extra videos on Notebooks

Quizzes:

Weekly quiz as usual.

Exams:

Next exam 11/21. Same format.

Review

- Basic Data Wrangling Steps
  - Missing values
  - Scaling
  - Encoding of categorical values
- Pipelines

https://ttpoll.com/p/747972

Today's Topics

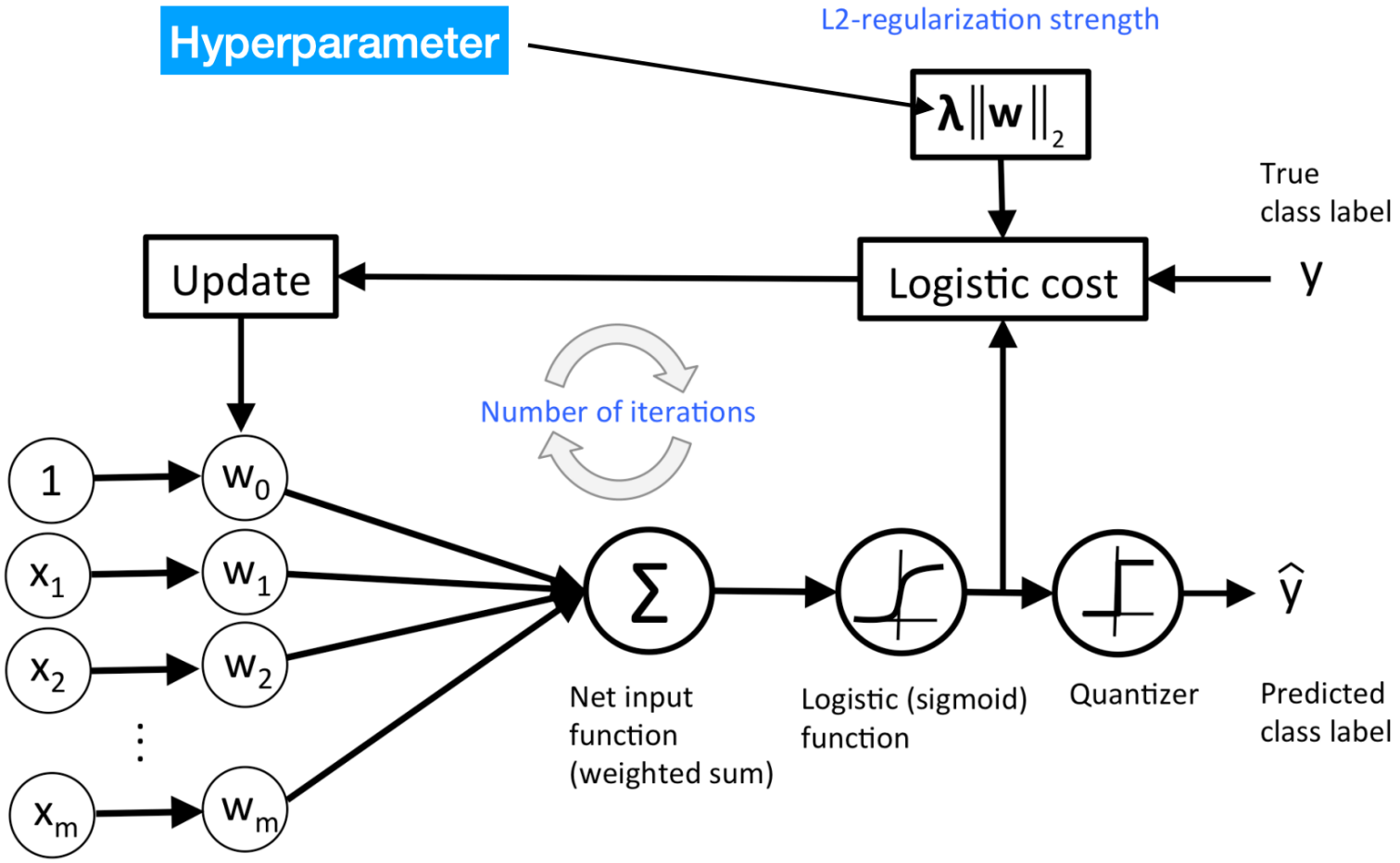
Hyperparameter Tuning

Model Evaluation

Hyperparameter Tuning

THE UNIVERSITY OF TENNESSEE KNOXVILLE

parametric model: logistic regression



Source: Dr. Raschka, ML Course

https://ttpoll.com/p/747972

Hyperparameters

- Learning rate α
- Mini-batch size
- Decision Trees
  - Tree depth
  - Bagging Yes/No
  - Size of Forest
- Polynomial Regression Degree
- Regularization λ
- Learning rate decay

https://ttpoll.com/p/747972

Hyperparameters

- Learning rate α
- Mini-batch size
- Decision Trees
  - Tree depth
  - Bagging Yes/No
  - Size of Forest
- Polynomial Regression Degree
- Regularization λ
- Learning rate decay

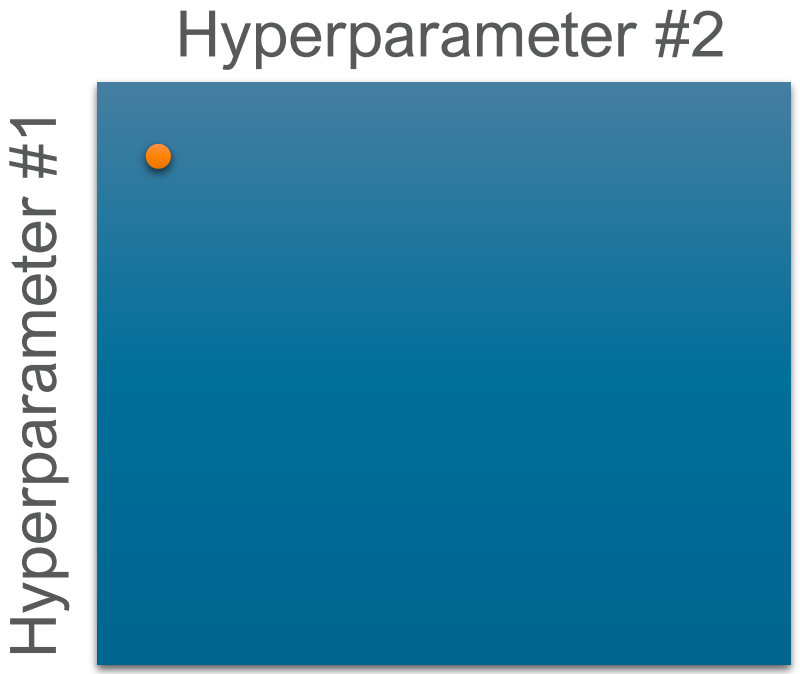
class `sklearn.tree.`**`DecisionTreeClassifier`**(*, *criterion='gini'*, *splitter='best'*, *max_depth=None*, *min_samples_split=2*, *min_samples_leaf=1*, *min_weight_fraction_leaf=0.0*, *max_features=None*, *random_state=None*, *max_leaf_nodes=None*, *min_impurity_decrease=0.0*, *min_impurity_split=None*, *class_weight=None*, *presort='deprecated'*, *ccp_alpha=0.0*)

https://ttpoll.com/p/747972

Hyperparameters

- Learning rate α
- Mini-batch size
- Decision Trees
  - Tree depth
  - Bagging Yes/No
  - Size of Forest
- Polynomial Regression Degree
- Regularization λ
- Learning rate decay

What not to do?

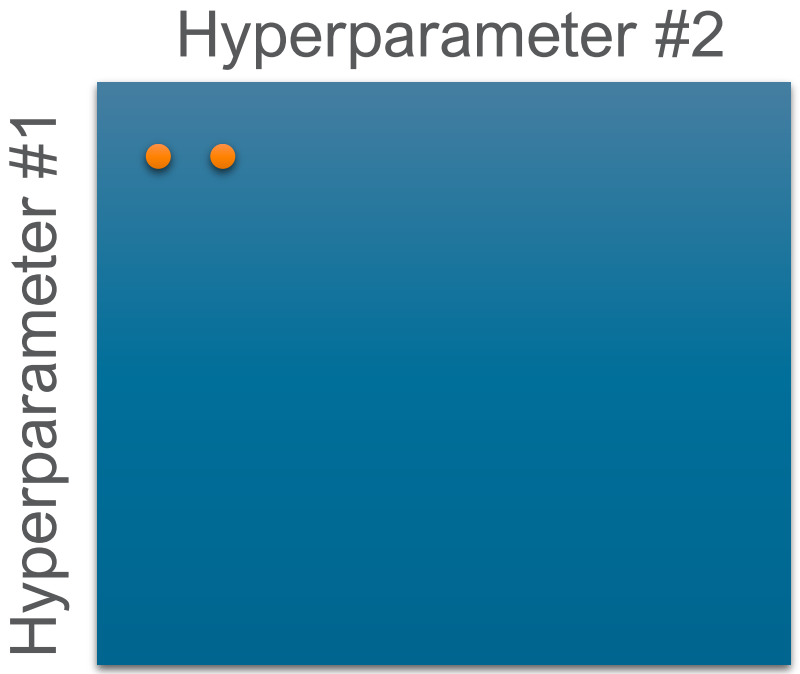


https://ttpoll.com/p/747972

Hyperparameters

- Learning rate α
- Mini-batch size
- Decision Trees
  - Tree depth
  - Bagging Yes/No
  - Size of Forest
- Polynomial Regression Degree
- Regularization λ
- Learning rate decay

What not to do?

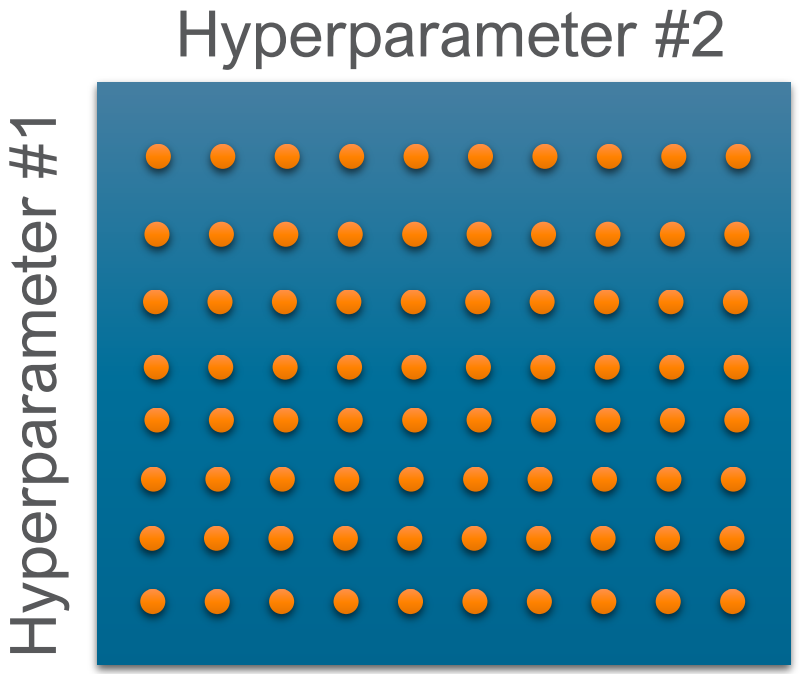


https://ttpoll.com/p/747972

Hyperparameters

- Learning rate α
- Mini-batch size
- Decision Trees
  - Tree depth
  - Bagging Yes/No
  - Size of Forest
- Polynomial Regression Degree
- Regularization λ
- Learning rate decay

What not to do?



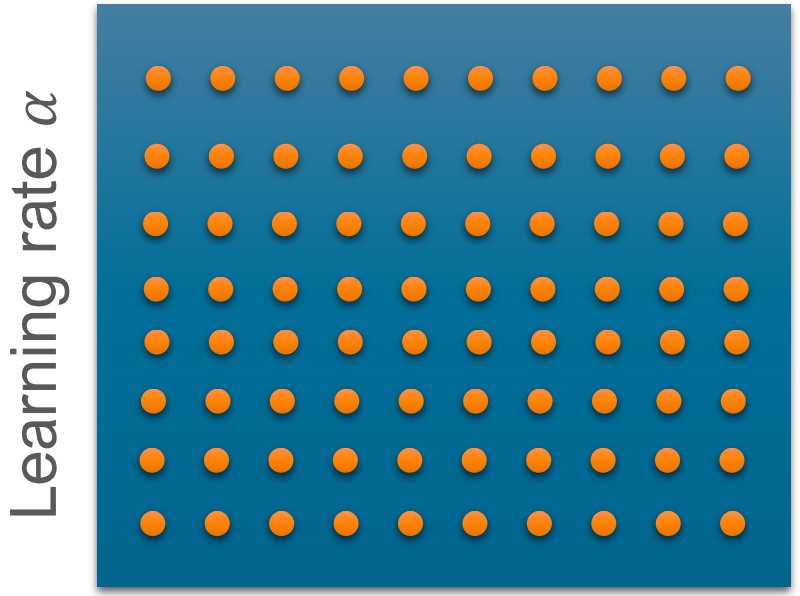
https://ttpoll.com/p/747972

Hyperparameters

- Learning rate α
- Mini-batch size
- Decision Trees
  - Tree depth
  - Bagging Yes/No
  - Size of Forest
- Polynomial Regression Degree
- Regularization λ
- Learning rate decay

What not to do?

Non-division by zero ϵ



https://ttpoll.com/p/747972

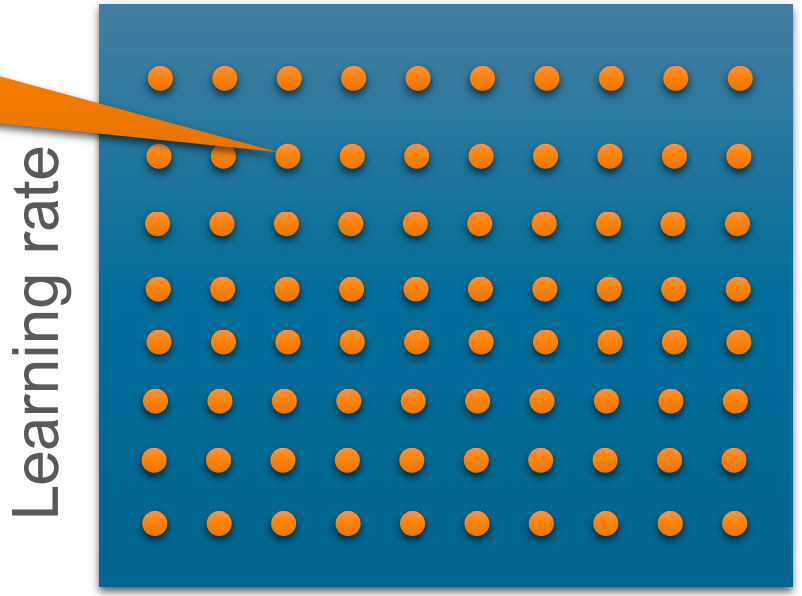
Hyperparameters

- Learning rate α
- Mini-batch
- Decision T
  - Tree dept
  - Bagging Yes/No
  - Size of Forest
- Polynomial Regression Degree
- Regularization λ
- Learning rate decay

We will test 10 values of a low priority parameter without changing a high-priority parameter.

What not to do?

Non-division by zero ϵ



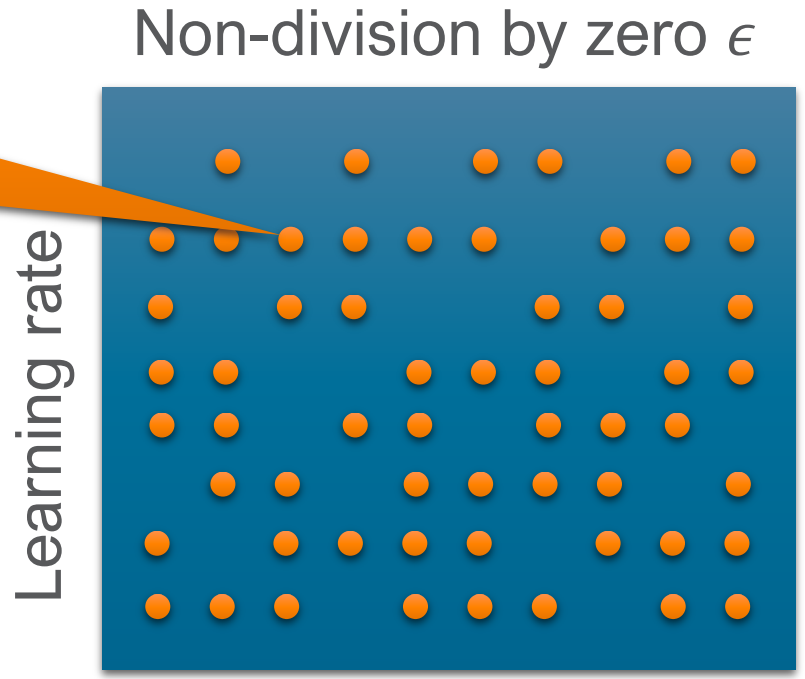
https://ttpoll.com/p/747972

Hyperparameters

- Learning rate α
- Mini-batch
- Decision T
  - Tree dept
  - Bagging Yes/No
  - Size of Forest
- Polynomial Regression Degree
- Regularization λ
- Learning rate decay

Randomly sample the hyperparameter space.

A better approach



https://ttpoll.com/p/747972

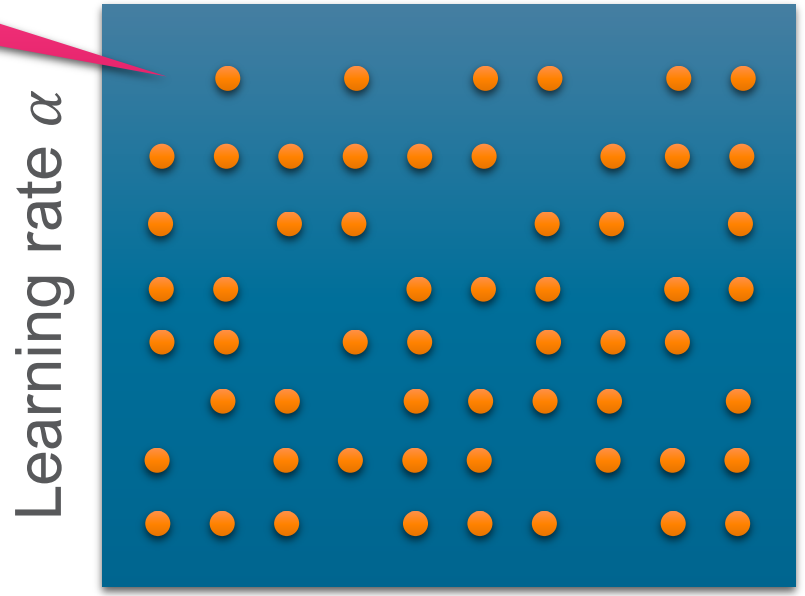
Hyperparameters

- Learning rate α
- Mini-batch size
- Decision Trees
  - Tree depth
  - Bagging Yes/No
  - Size of Forest
- Polynomial Regression Degree
- Regularization λ
- Learning rate decay

An even better approach

Coarse search

Non-division by zero ϵ

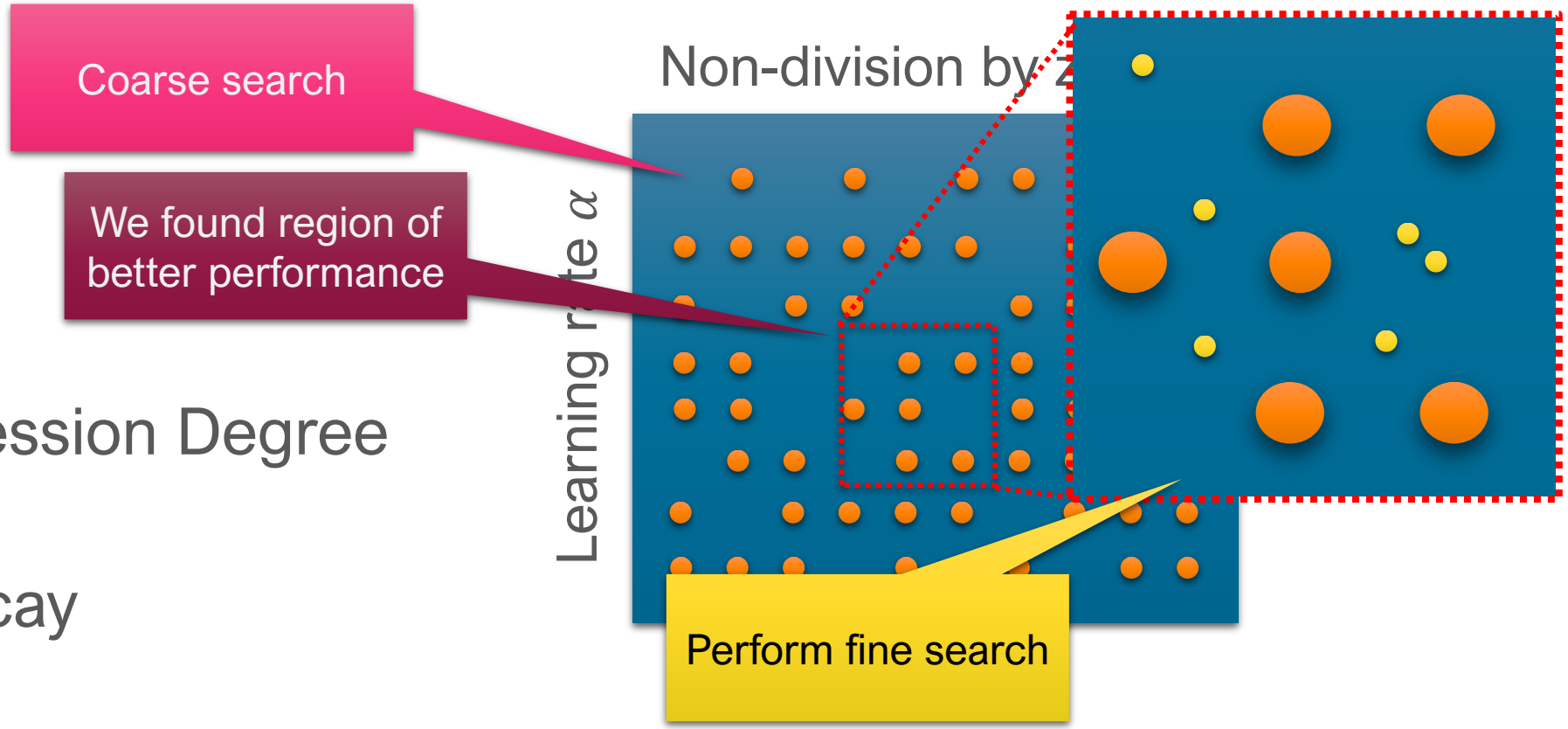


https://ttpoll.com/p/747972

Hyperparameters

- Learning rate α
- Mini-batch size
- Decision Trees
  - Tree depth
  - Bagging Yes/No
  - Size of Forest
- Polynomial Regression Degree
- Regularization λ
- Learning rate decay

An even better approach



https://ttpoll.com/p/747972

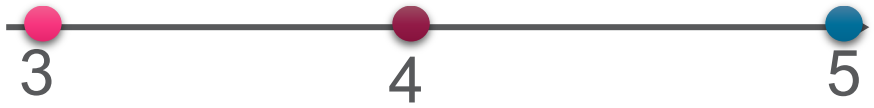
Scaling Hyperparameter Search Space

- Number of trees in a RandomForest T

$$T \in \{100, 200, \ldots, 1000\}$$

- Tree Depth L

$$L \in \{3 \ \ldots, 5\}$$

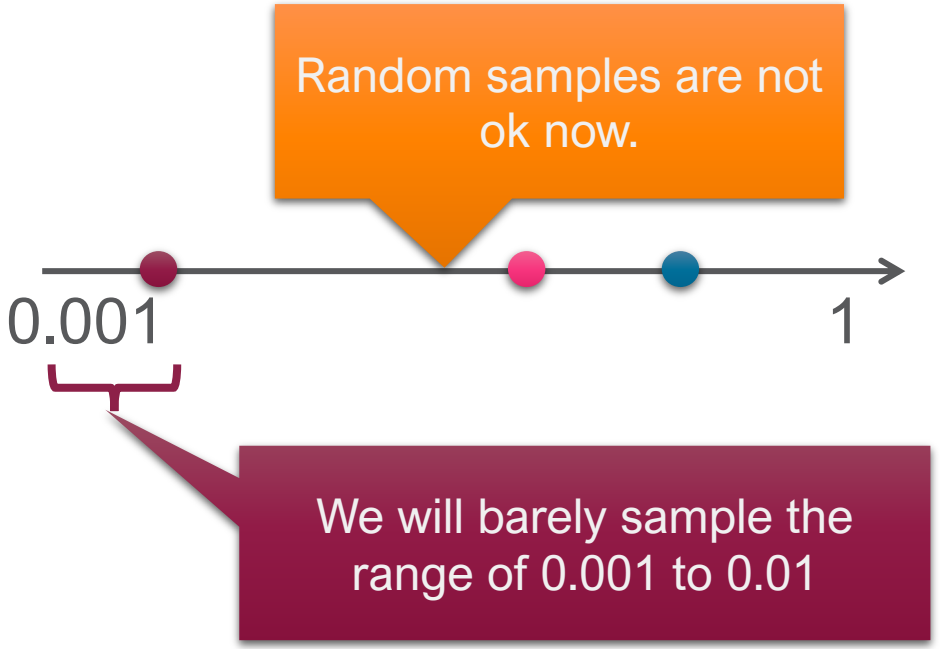


Scaling Hyperparameter Search Space

- Learning rate α

$$\alpha \in \{0.001, \ldots, 1\}$$

- $\alpha = uniform\_sample\ (0.001,\ \ 1.0)$



Scaling Hyperparameter Search Space

- Learning rate α

$$\alpha \in \{0.001, \ldots, 1\}$$

- $\alpha = uniform\_sample\ (0.001,\ \ 1.0)$

Log of starting point

Log of ending point

$\ln(0.001)$

$\ln(1)$

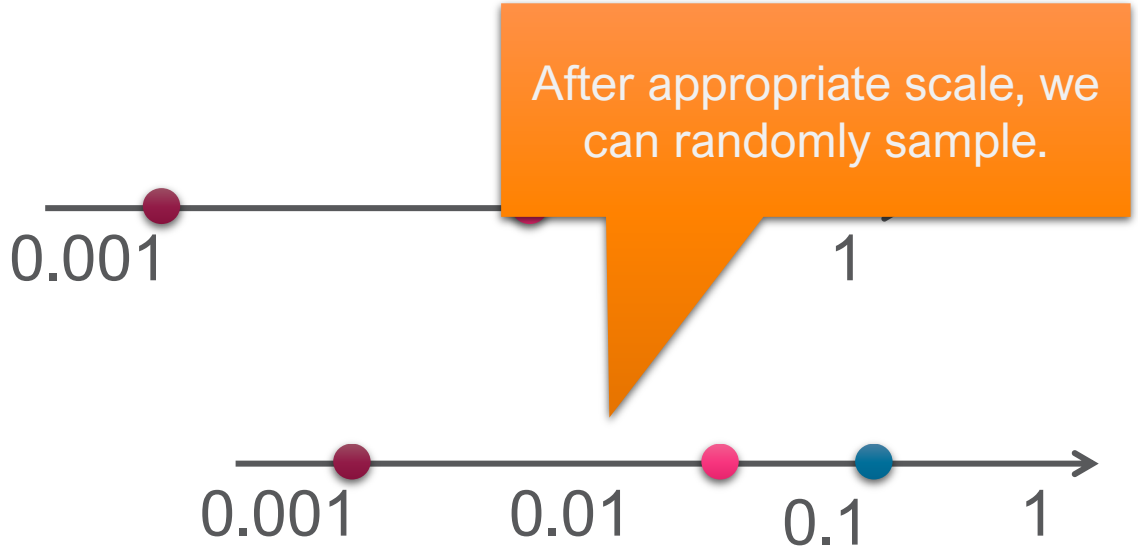
- Instead do $\alpha = 10^{uniform\_sample(-3,0)}$

Scaling Hyperparameter Search Space

- Learning rate α

$$\alpha \in \{0.001, \ldots, 1\}$$

- $\alpha = uniform\_sample\ (0.001,\ \ 1.0)$



- Instead do $\alpha = 10^{uniform\_sample(-3,0)}$

Code Example

```
# Define the parameter range to search over
alpha = np.power(10, rng1.randint(-6,-1,5).astype('float'))  # Numpy approach
learning_rate = np.power(10, rng1.randint(-5,0,5).astype('float'))  # Numpy approach
param_distributions = {
    'MLClassifier__penalty': ['l1', 'l2', 'elasticnet'],    # Regularization
    'MLClassifier__loss': ['hinge', 'log_loss'],            # Loss function
    'MLClassifier__alpha': stats.loguniform(1e-6,1e-2),     # Regularization coefficient
    'MLClassifier__eta0': stats.loguniform(1e-5,1e-1),       # Learning rate
}

# Initialize RandomizedSearchCV
hyperparam_search = RandomizedSearchCV(
    estimator=ml_pipe,
    param_distributions=param_distributions,
```

https://ttpoll.com/p/747972

Pop Quiz

Which of the following is a model parameter? (Select all that apply)

A. Decision Tree Split Features and Thresholds

B. Tree Depth

C. Logistic regression hypothesis coefficients

D. Learning rate

https://ttpoll.com/p/747972

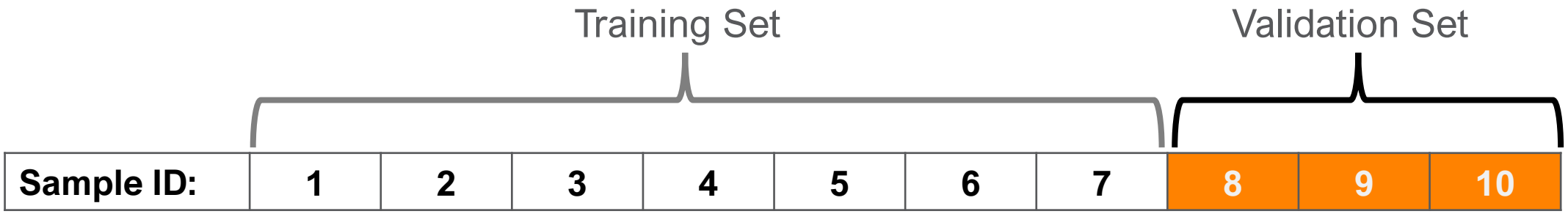
Notebook Time

https://ttpoll.com/p/747972

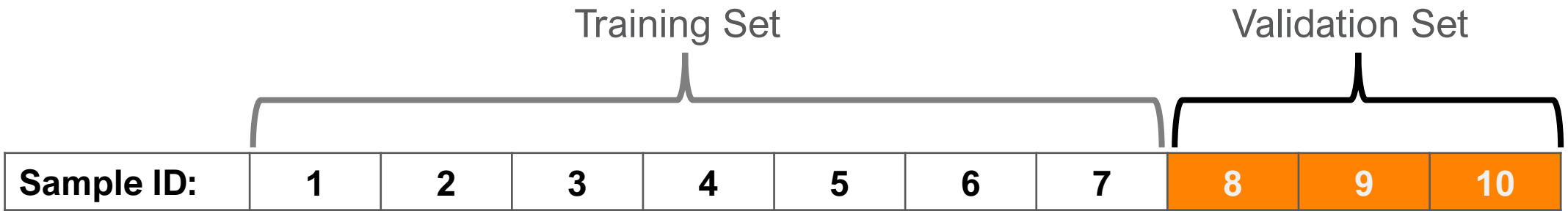
Why do we assess the predictive performance of machine learning models?

- Estimate generalization from performance on unseen data
- We want to tweak hyperparameters to squeeze as much performance as possible from the selected hypothesis space.
- We want to select the best-performing ML algorithm/hypothesis space

Holdout Method

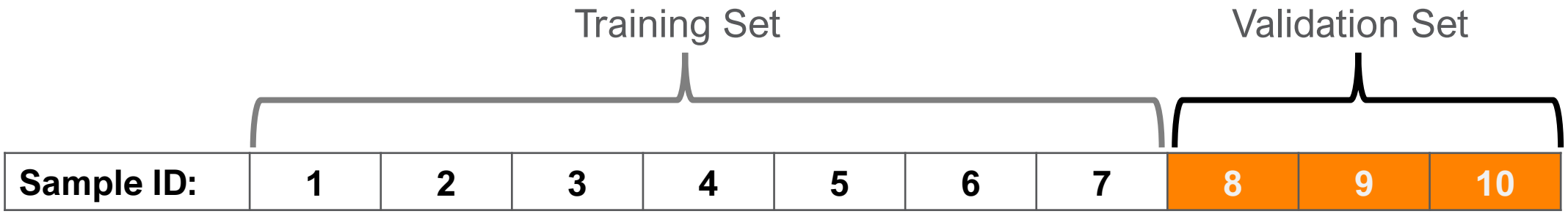


Holdout Method



Sometimes, the holdout method can be misleading.

Holdout Method



Sometimes, the holdout method can be misleading.

Training error is biased and over-optimistic.

Validation/Test error is unbiased (unseen data). But is it optimistic or pessimistic?

https://ttpoll.com/p/747972

Test/Validation Set Optimism

1. Underutilization of data
   - Typical 20-30% validation set reduces samples available for training
2. Does not account of variance in training data
   - E.g., change seed in train_test_split()
3. High variance due to selection split
   - Unbalanced classes or Skewed target distributions
   - Too hard or too easy training/validation/sets
   - E.g., credit card fraud detection example with random split

https://ttpoll.com/p/747972

Test/Validation Set Optimism

1. Underutilization of data
   - Typical 20-30% validation set reduces samples available for training
2. Does not account of variance in training data
   - E.g., change seed in train_test_split()
3. High variance due to selection split
   - Unbalanced classes or Skewed target distributions
   - Too hard or too easy training/validation/sets
   - E.g., credit card fraud detection example with random split

Test/validation either over-optimistic or pessimistic

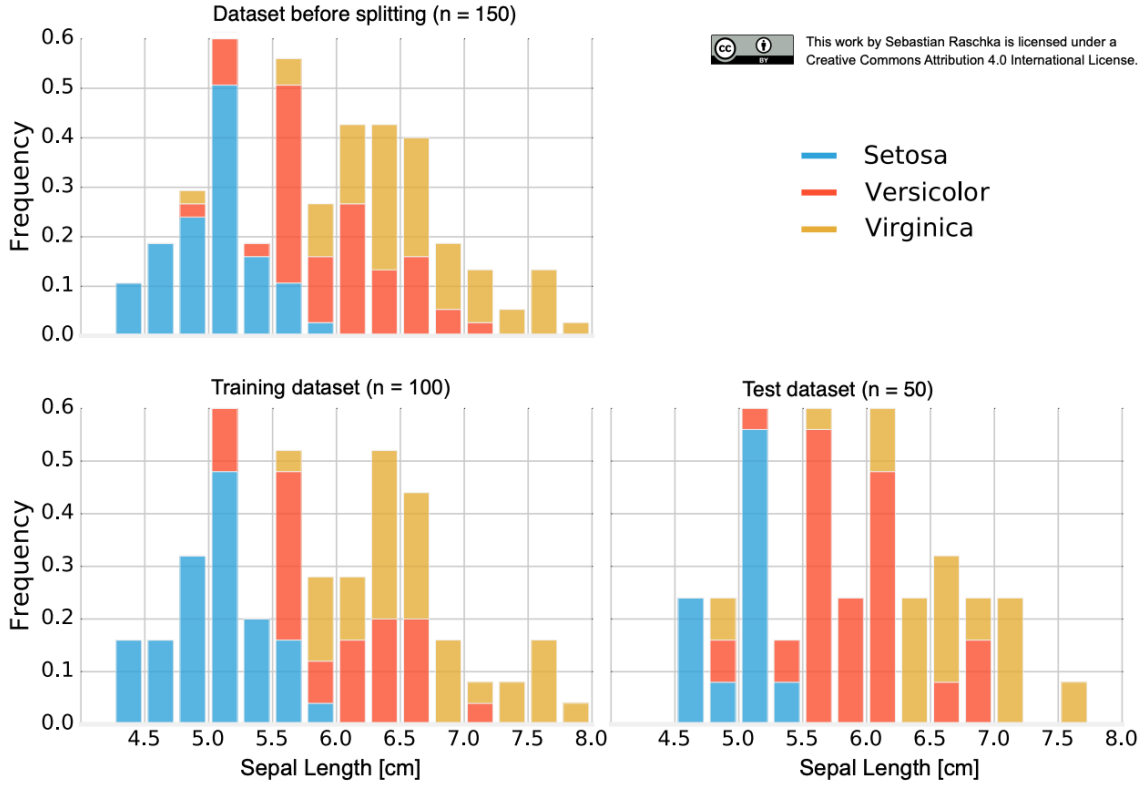
Pessimistic

- Pessimistic test error not necessary a bad thing

- For model evaluation:
  - Correct test error for three models: $h_1 = 25\% < h_2 = 30\% < h_3 = 35\%$
  - Pessimistically biased test error of 5%: $h_1 = 30\% < h_2 = 35\% < h_3 = 40\%$
  - The ranking of the models does not change

Training Data Variance

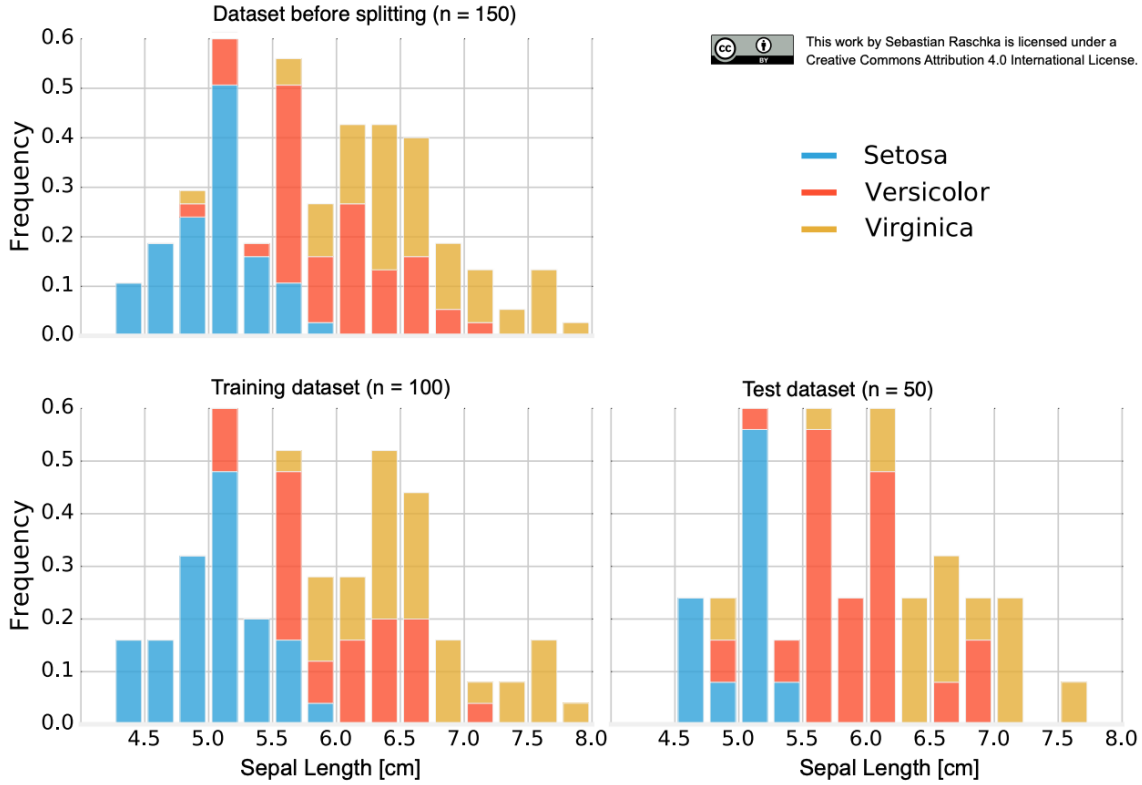
- Original dataset (n=150):
  - 50 (33.3%) Setosa (S)
  - 50 Versicolor (Ve)
  - 50 Virginica (Vi)
- Potential split distribution
  - Training set (100): S-38, Ve-28, Vi-34
  - Test set (50): S-12 Ve-22, Vi-16



Source: Dr. Raschka ML Course

Training Data Variance

- Original dataset (n=150):
  - 50 (33.3%) Setosa (S)
  - 50 Versicolor (Ve)
  - 50 Virginica (Vi)
- Potential split distribution
  - Training set (100): S-38, Ve-28, Vi-34
  - Test set (50): S-12 Ve-22, Vi-16
- One solution: Stratify



Source: Dr. Raschka ML Course

Pessimistic

- Pessimistic test error not necessary a bad thing

- For model evaluation:
  - Correct test error for three models: $h_1 = 25\% < h_2 = 30\% < h_3 = 35\%$
  - Pessimistically biased test error of 5%: $h_1 = 30\% < h_2 = 35\% < h_3 = 40\%$
  - The ranking of the models does not change

What about over-optimism?

https://ttpoll.com/p/747972

How confident are we about the model performance?

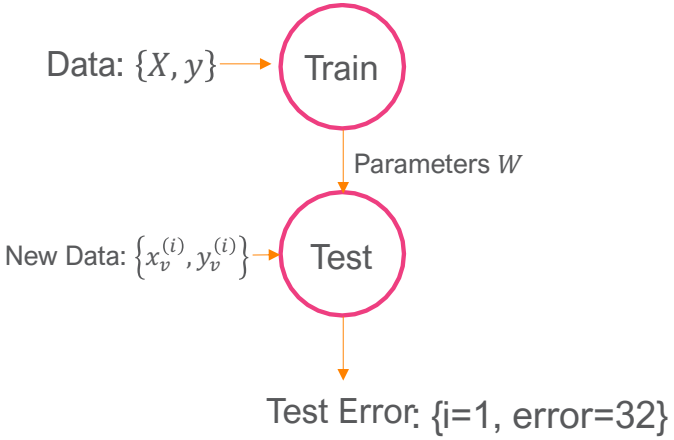
Model Evaluation with Confidence Interval

Confidence Intervals for Model Evaluation

- The confidence interval is a range of values for a variable v.
  - E.g., $v_{low} < v < v_{high}$
- The $X\%$ confidence interval is the probability $X\%$ of v be within the range $[v_{low}, v_{high}]$
  - E.g., 95% confidence interval of $30 < v < 40$

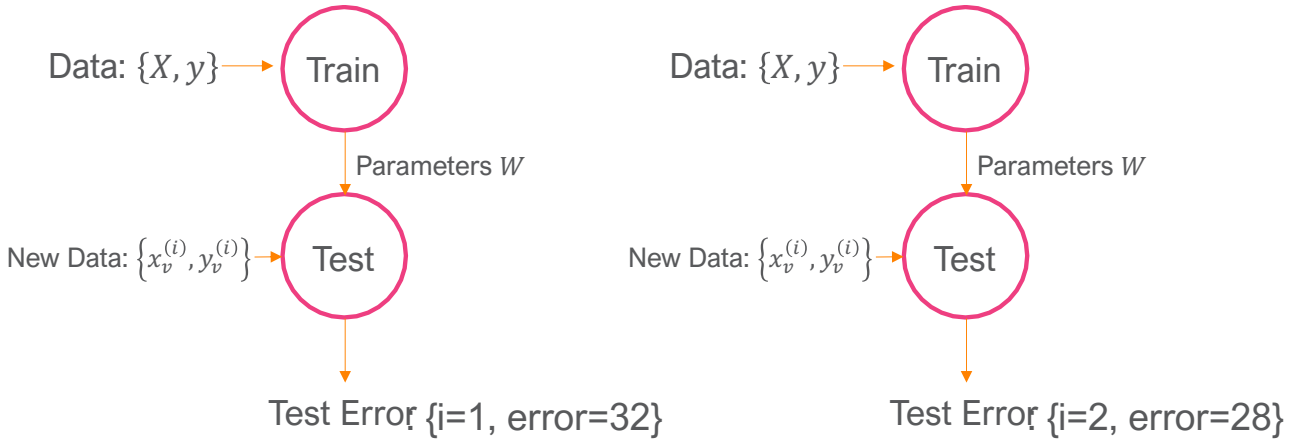
Confidence Intervals for Model Evaluation

- The confidence interval is a range of values for a variable v.
  - E.g., $v_{low} < v < v_{high}$
- The $X\%$ confidence interval is the probability $X\%$ of v be within the range $[v_{low}, v_{high}]$
  - E.g., 95% confidence interval of $30 < v < 40$



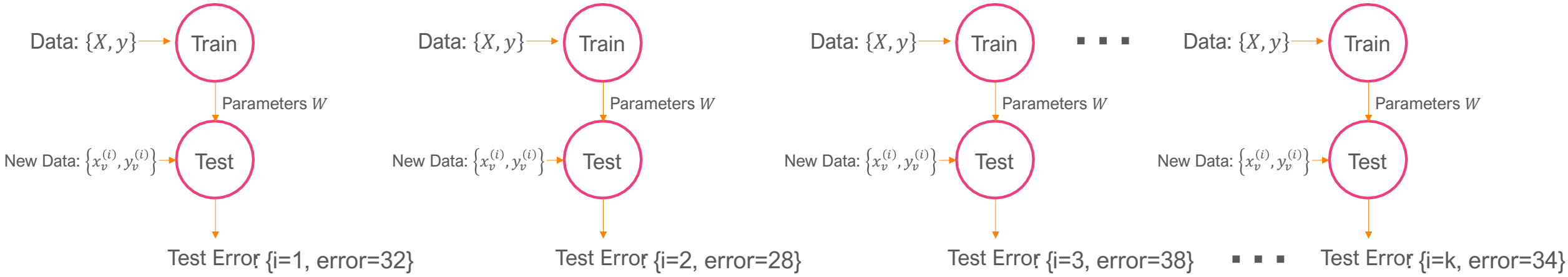
Confidence Intervals for Model Evaluation

- The confidence interval is a range of values for a variable v.
  - E.g., $v_{low} < v < v_{high}$
- The $X\%$ confidence interval is the probability $X\%$ of v be within the range $[v_{low}, v_{high}]$
  - E.g., 95% confidence interval of $30 < v < 40$



Confidence Intervals for Model Evaluation

- The confidence interval is a range of values for a variable v.
  - E.g., $v_{low} < v < v_{high}$
- The $X\%$ confidence interval is the probability $X\%$ of v be within the range $[v_{low}, v_{high}]$
  - E.g., 95% confidence interval of $30 < v < 40$



https://ttpoll.com/p/747972

Pop Quiz

A researcher wants to estimate the average height of a city's adult male population. She collected a random sample of 100 adult males and calculated the sample mean height to be 175 cm (5 ft and 9 inches) with a sample standard deviation of 10 cm (4 inches). She then constructs a 95% confidence interval for the height of the population [173 cm, 177 cm].

What is the correct interpretation of this 95% confidence interval?

A. 95% of the time, the sample mean is between 173 cm and 177 cm.

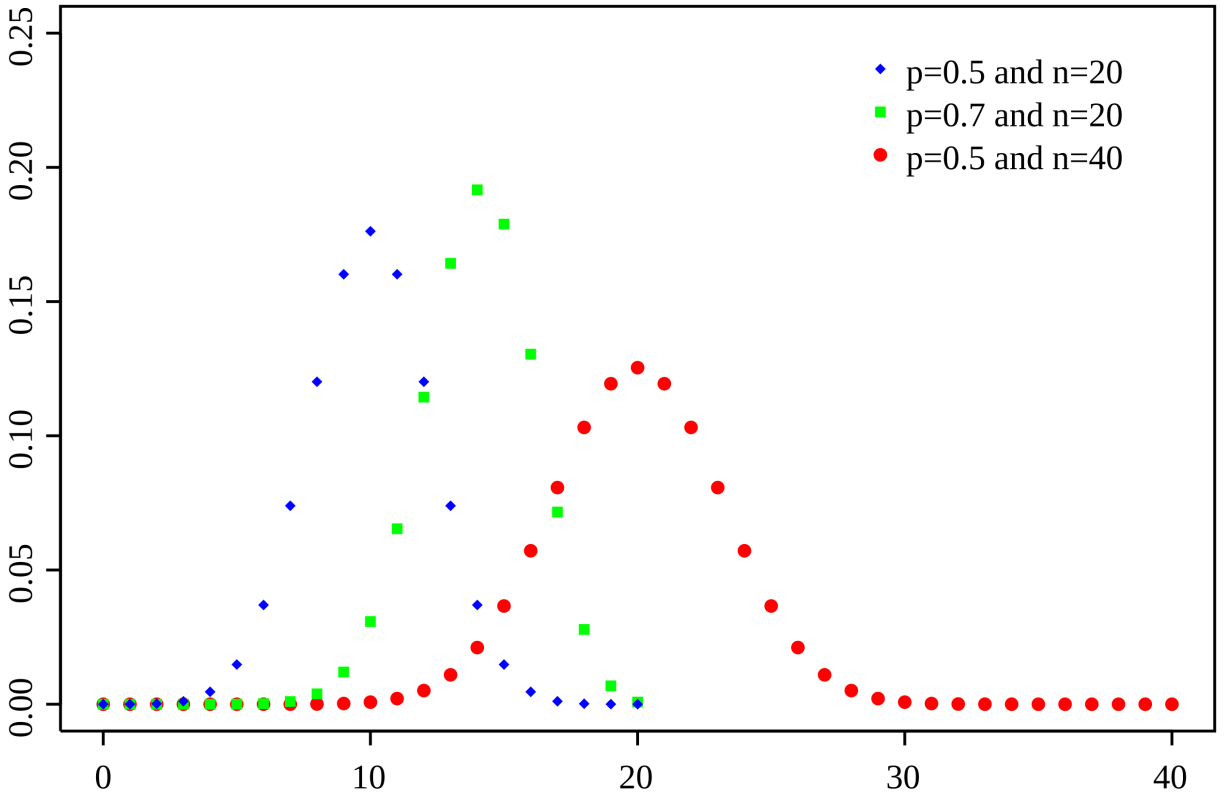
B. 95% of the time, a random adult male height sample is between 173 cm and 177 cm.

C. 95% of the time, the true adult male height mean is between 173 cm and 177 cm.

https://ttpoll.com/p/747972

Binomial Distribution

$$\Pr(X = k) = \left(\frac{n!}{k!\,(n-k)!}\right) p^k (1-p)^{n-k}$$



Notation	$B(n, p)$
Parameters	$n \in \{0, 1, 2, \ldots\}$ – number of trials $p \in [0, 1]$ – success probability for each trial $q = 1 - p$
Support	$k \in \{0, 1, \ldots, n\}$ – number of successes
PMF	$\binom{n}{k} p^k q^{n-k}$
CDF	$I_q(n - \lfloor k \rfloor, 1 + \lfloor k \rfloor)$ (the regularized incomplete beta function)

Mean	np
Median	$\lfloor np \rfloor$ or $\lceil np \rceil$
Mode	$\lfloor (n+1)p \rfloor$ or $\lceil (n+1)p \rceil - 1$
Variance	$npq = np(1-p)$
Skewness	$\frac{q-p}{\sqrt{npq}}$
Excess kurtosis	$\frac{1-6pq}{npq}$
Entropy	$\frac{1}{2}\log_2(2\pi enpq) + O\left(\frac{1}{n}\right)$ in shannons. For nats, use the natural log in the log.

Coin Flip (Bernoulli Trial)

- Mean μ_k is the average number of Heads in n trials

$$\mu_k = np = n\left(\frac{k}{n}\right) = \mathrm{k}$$

- Mean Variance $\sigma_k^2 = np(1-p) = k - \frac{k^2}{n}$

- Standard deviation $= sqrt(\sigma_k^2) = \sigma_k = \sqrt{(np(1-p))}$

Our ML results with a 0-1 Loss

- $ERR_s = \frac{1}{n}\sum_{x \in S} \delta\big(f(x), h(x)\big) = \frac{1}{n}\sum_{x \in S} I(y, \hat{y}) = p$
- Mean Variance $\sigma^2 = p(1-p)$
- Standard Error $SE = \sqrt{\frac{\sigma^2}{n}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$

Our ML results with a 0-1 Loss

- $ERR_s = \frac{1}{n}\sum_{x \in S} \delta\big(f(x), h(x)\big) = \frac{1}{n}\sum_{x \in S} I(y, \hat{y}) = \boldsymbol{p}$
- Mean Variance $\sigma^2 = p(1-p)$
- Standard Error $SE = \sqrt{\frac{\sigma^2}{n}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$

Normal Approximation Interval

- $n > 40$
- $np > 5$
- $n(1-p) > 5$

$$CI = p \pm z\sqrt{\frac{p(1-p)}{n}}$$

Our ML results with a 0-1 Loss

- $ERR_s = \frac{1}{n}\sum_{x \in S} \delta\big(f(x), h(x)\big) = \frac{1}{n}\sum_{x \in S} I(y, \hat{y}) = \boldsymbol{p}$
- Mean Variance $\sigma^2 = p(1-p)$
- Standard Error $SE = \sqrt{\frac{\sigma^2}{n}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$

Normal Approximation Interval

- $n > 40$
- $np > 5$
- $n(1-p) > 5$

$$CI = p \pm \boldsymbol{z}\sqrt{\frac{p(1-p)}{n}}$$

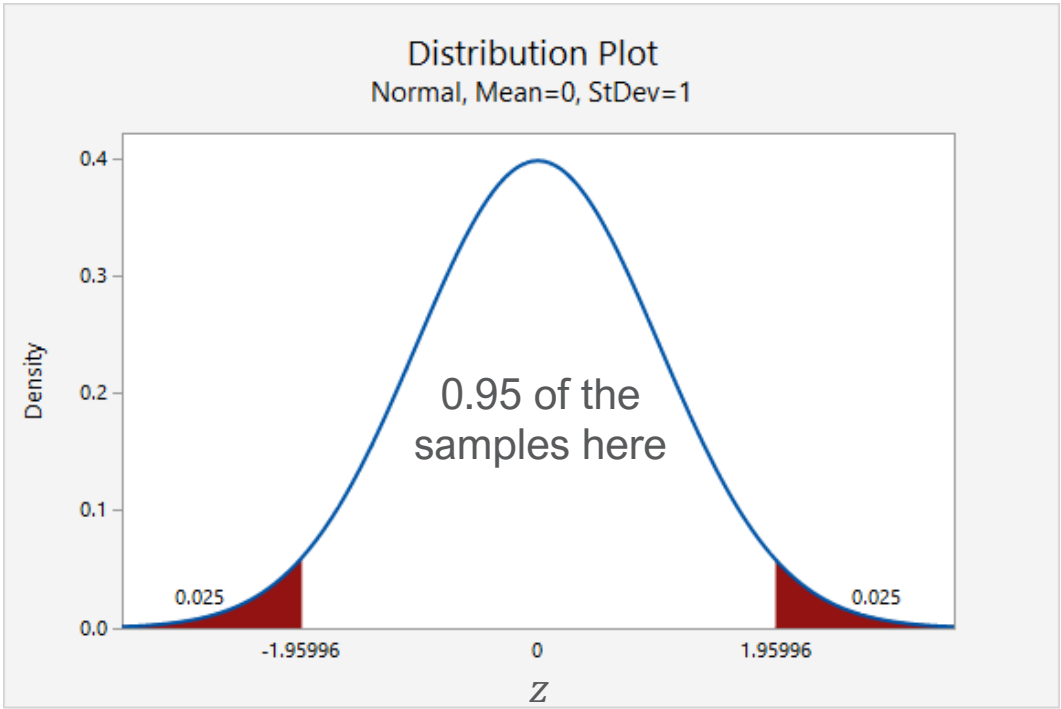
$X\%$ Confidence Interval

Normal Approximation Interval

Normal Approximation Interval

- $n > 40$
- $np > 5$
- $n(1-p) > 5$

$$CI = p \pm \boldsymbol{z}\sqrt{\frac{p(1-p)}{n}}$$



CI for other z values: $99\%, z = 2.58$, $90\%, z = 1.64$

https://ttpoll.com/p/747972

Code example

```python
# Number of samples
n = len(y_val)

# Z-value for Confidence interval of 95%
# Recall 99% -> z=2.58; 95% -> z=1.96; 90% -> z=1.64
Z = 1.96

# Compute error or accuracy
p = accuracy_score(y_val, y_hat)

# Compute confidence interval
CI = Z * np.sqrt(p*(1-p)/n)
lower_bound = p - CI
upper_bound = p + CI

# Display results
print(f"The 95% Confidence Interval for our model accuracy is between {lower_bound*100:.2f} and {u
```

✓ 0.0s Python

```
The 95% Confidence Interval for our model accuracy is between 74.75 and 87.94
```

https://ttpoll.com/p/747972

Estimate on a single trained model

Our error could still be over-optimistic due to sampling random effects.

Repeated Holdout

- Also known as Monte Carlo Cross Validation

- Average performance over k repetitions

$$ACC_{avg} = \frac{1}{k}\sum_{j=1}^{k} ACC_j$$

- ACC_j is the accuracy estimate for the jth test set of size m

$$ACC_j = 1 - \frac{1}{m}\sum_{i=1}^{m} I\left(y^{(i)}, \hat{y}^{(i)}\right)$$

https://ttpoll.com/p/747972

Bootstrapping and Empirical Confidence Interval

- Bootstrapping: resampling technique to estimate a sampling distribution
  - Introduced by Bradley Efron 1979
  - We are interested in estimating the distribution of our model error
  - Sampling with replacement
  - In contrast, holdout method samples without replacement

Bootstrap Algorithm

1. Inputs: Data size $\{X, y\}$ of size n
2. For B bootstraps trials:
   1. Draw n samples with replacement from $\{X, y\}$ to form $\{X, y\}_b$
   2. Fit model with $\{X, y\}_b$
   3. Compute model accuracy

$$ACC_b = \frac{1}{n}\sum_{i=1}^{n} 1 - L\left(f\left(x_b^{(i)}\right), h\left(x_b^{(i)}\right)\right)$$

3. Model accuracy is the average bootstrap accuracy

$$ACC_{boot} = \frac{1}{B}\sum_{b=1}^{B} ACC_b$$

Bootstrap Algorithm

1. Inputs: Data size $\{X, y\}$ of size n
2. For B bootstraps trials:
   1. Draw n samples with replacement from $\{X, y\}$ to form $\{X, y\}_b$
   2. Fit model with $\{X, y\}_b$
   3. Compute model accuracy

$$ACC_b = \frac{1}{n}\sum_{i=1}^{n} 1 - L\left(f\left(x_b^{(i)}\right), h\left(x_b^{(i)}\right)\right)$$

3. Model accuracy is the average bootstrap accuracy

$$ACC_{boot} = \frac{1}{B}\sum_{b=1}^{B} ACC_b$$

Too Optimistic!

Better Bootstrapping Approaches

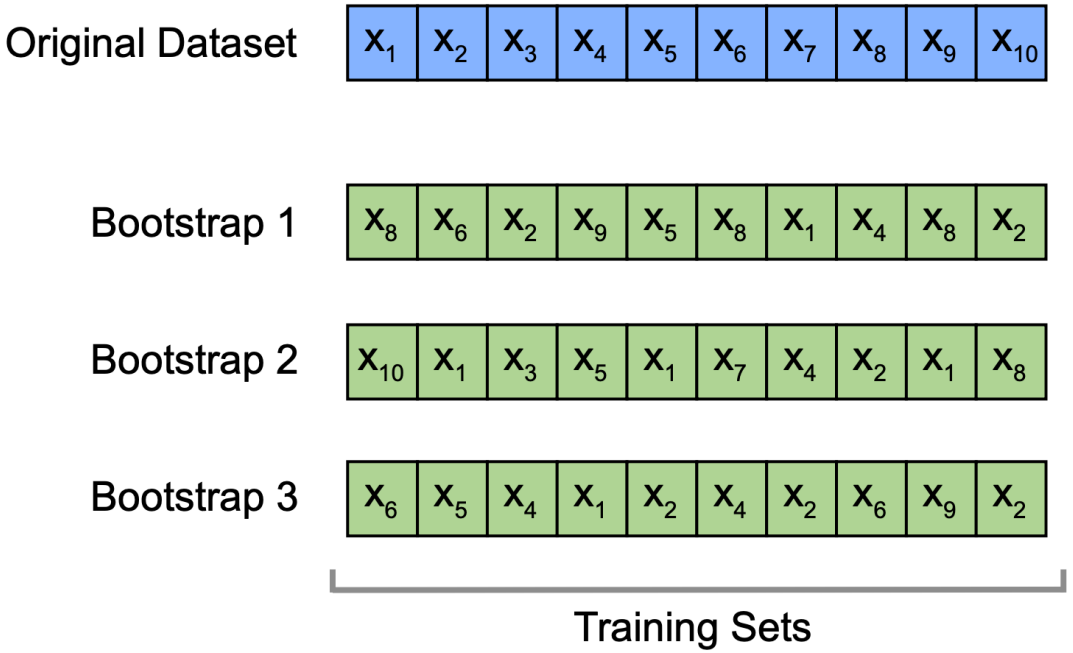
- Leave-One-Out Bootstrap (LOOB)
  - Same as above
  - Ensures one of the samples at each bootstrap trial is left out for validation
  - Accuracy measured on validation sample
- Out-of-bag Bootstrapping
  - Samples left out during bootstrapping sampling are used for testing
  - Accuracy measured on test samples
  - Recall a little over 36% of samples are left out at each bootstrap trial

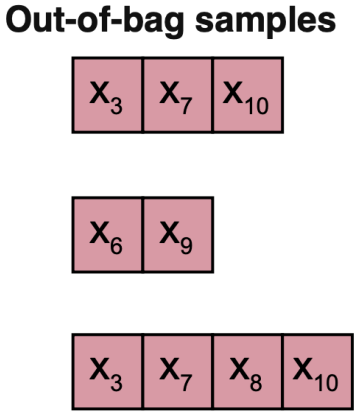
https://ttpoll.com/p/747972

Out-of-Bag Bootstrapping (OOB)

Source: Dr. Raschka, Machine Learning Course

Out-of-Bag Bootstrapping (OOB)





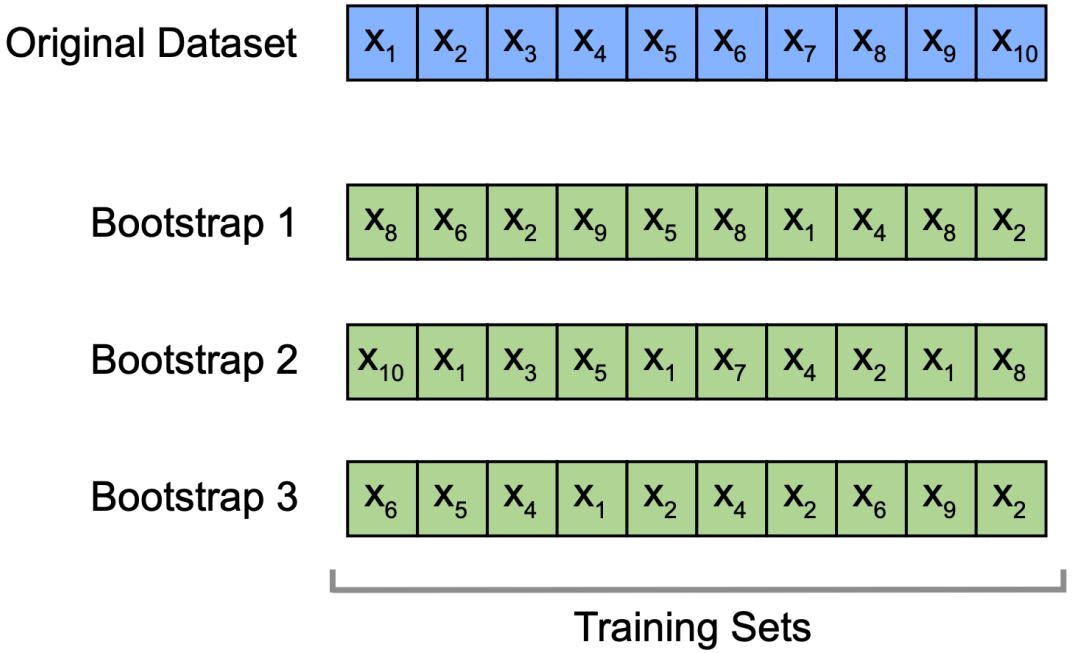
$$ACC_{boot} = \frac{1}{B}\sum_{b=1}^{B} ACC_b$$

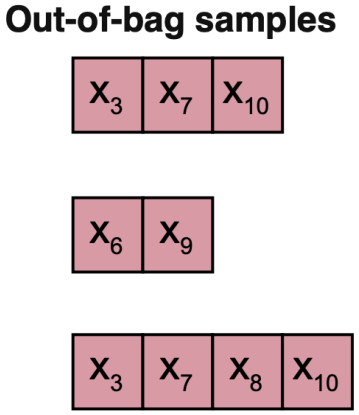
$$SE_{boot} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(ACC_b - ACC_{boot})^2}$$

$$CI = ACC_{boot} \pm t \times SE_{boot}$$

Source: Dr. Raschka, Machine Learning Course

Out-of-Bag Bootstrapping (OOB)





$$ACC_{boot} = \frac{1}{B}\sum_{b=1}^{B} ACC_b$$

$$SE_{boot} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(ACC_b - ACC_{boot})^2}$$

$$CI = ACC_{boot} \pm \boldsymbol{t} \times SE_{boot}$$

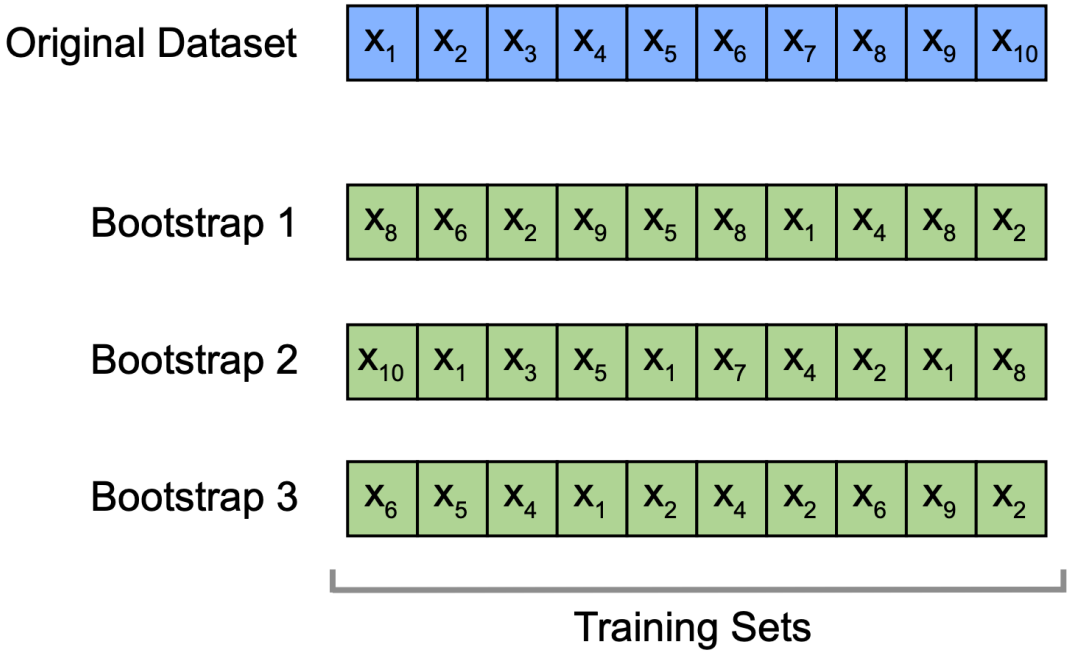
T Distribution

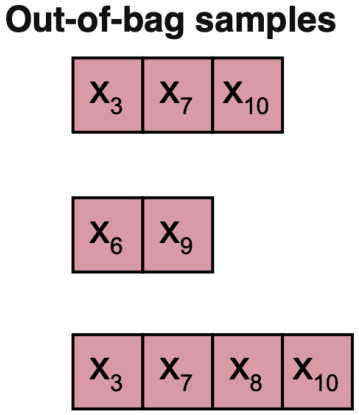
For $B = 200$ and 95% CI

$t_{\alpha=0.05,99} = 1.984$

Source: Dr. Raschka, Machine Learning Course

Out-of-Bag Bootstrapping (OOB)





$$ACC_{boot} = \frac{1}{B}\sum_{b=1}^{B} ACC_b$$

$$SE_{boot} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(ACC_b - ACC_{boot})^2}$$

$$CI = ACC_{boot} \pm \boldsymbol{t} \times SE_{boot}$$

At least 200 bootstrap trials are recommended.

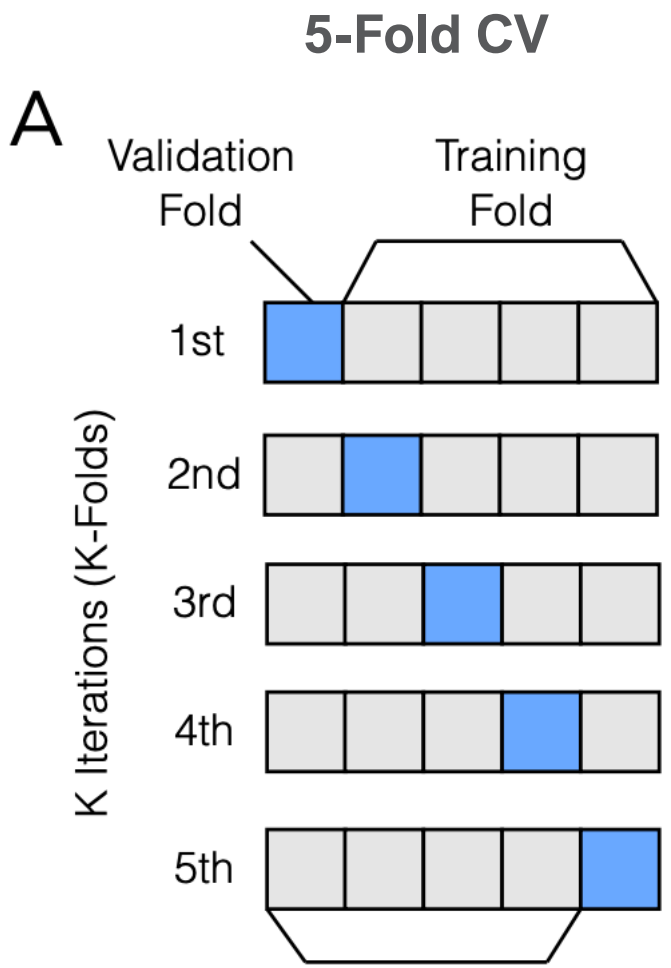
Source: Dr. Raschka, Machine Learning Course

K-Fold Cross Validation

THE UNIVERSITY OF TENNESSEE KNOXVILLE

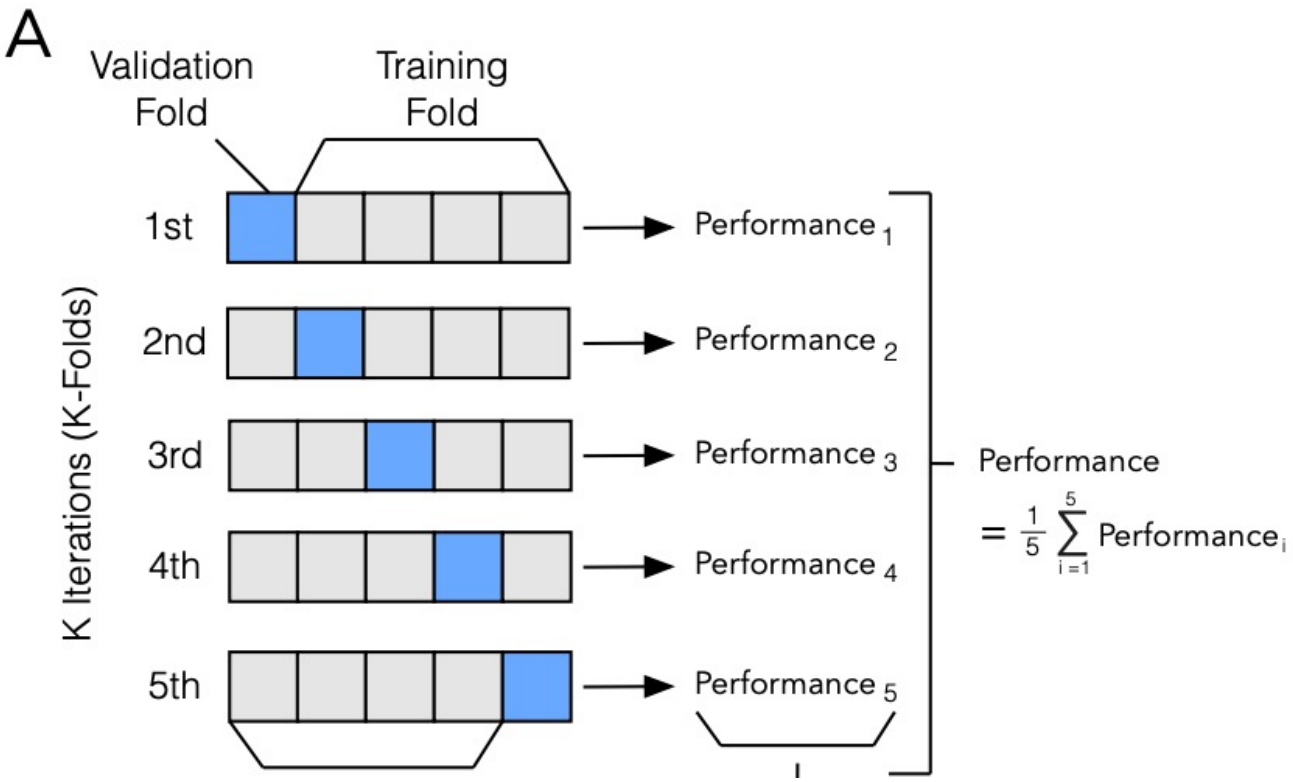
K-Fold Cross Validation

- Non-overlapping validation folds
  - Utilizes all data for validation (aggregated folds)
- Overlapping training folds
  - Sampling without replacement results in poor measurement of variance
- Pessimistic for small k
  - Smaller training sets



Source: Dr. Raschka, Machine Learning Course

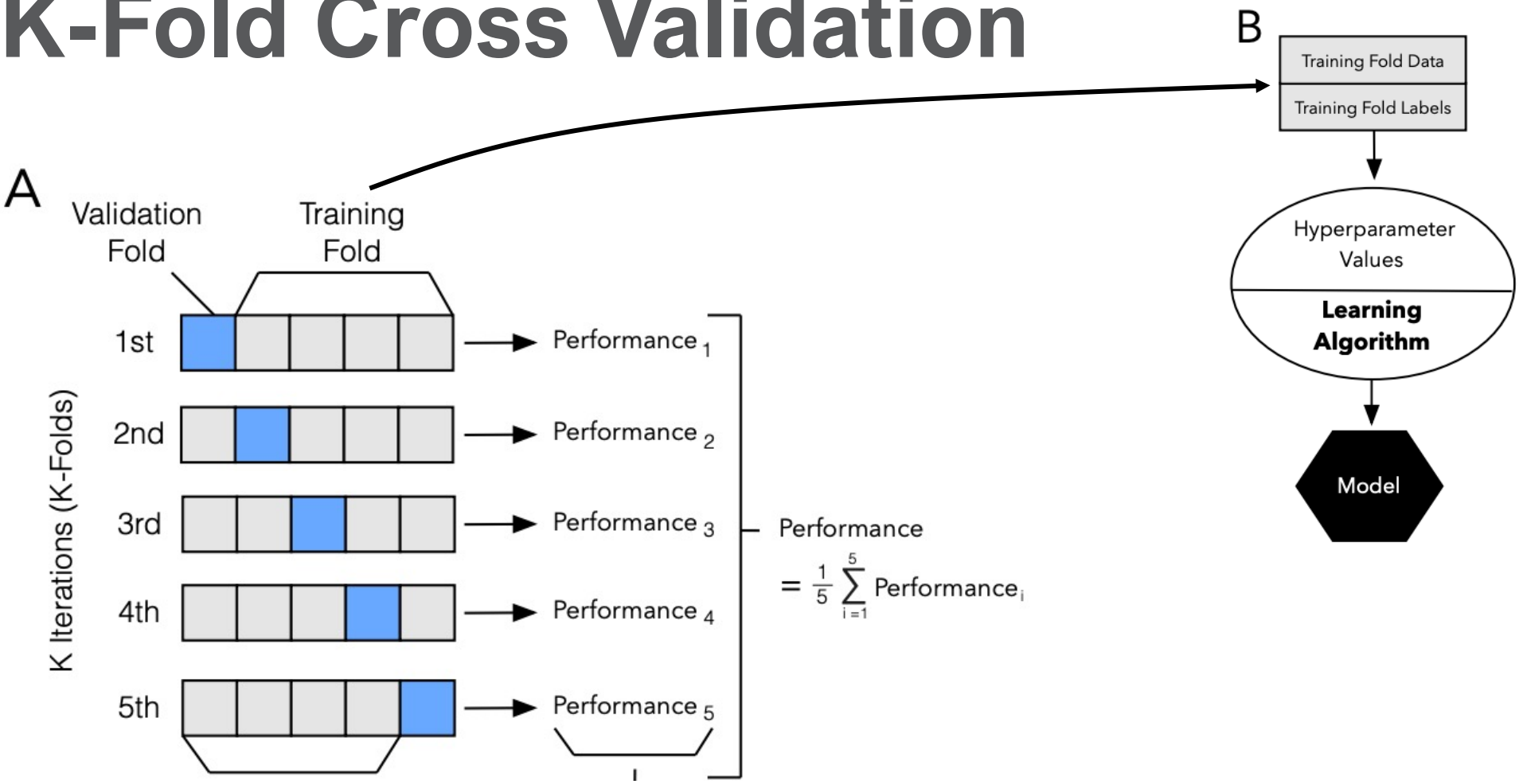
K-Fold Cross Validation



Source: Dr. Raschka, Machine Learning Course

https://ttpoll.com/p/747972

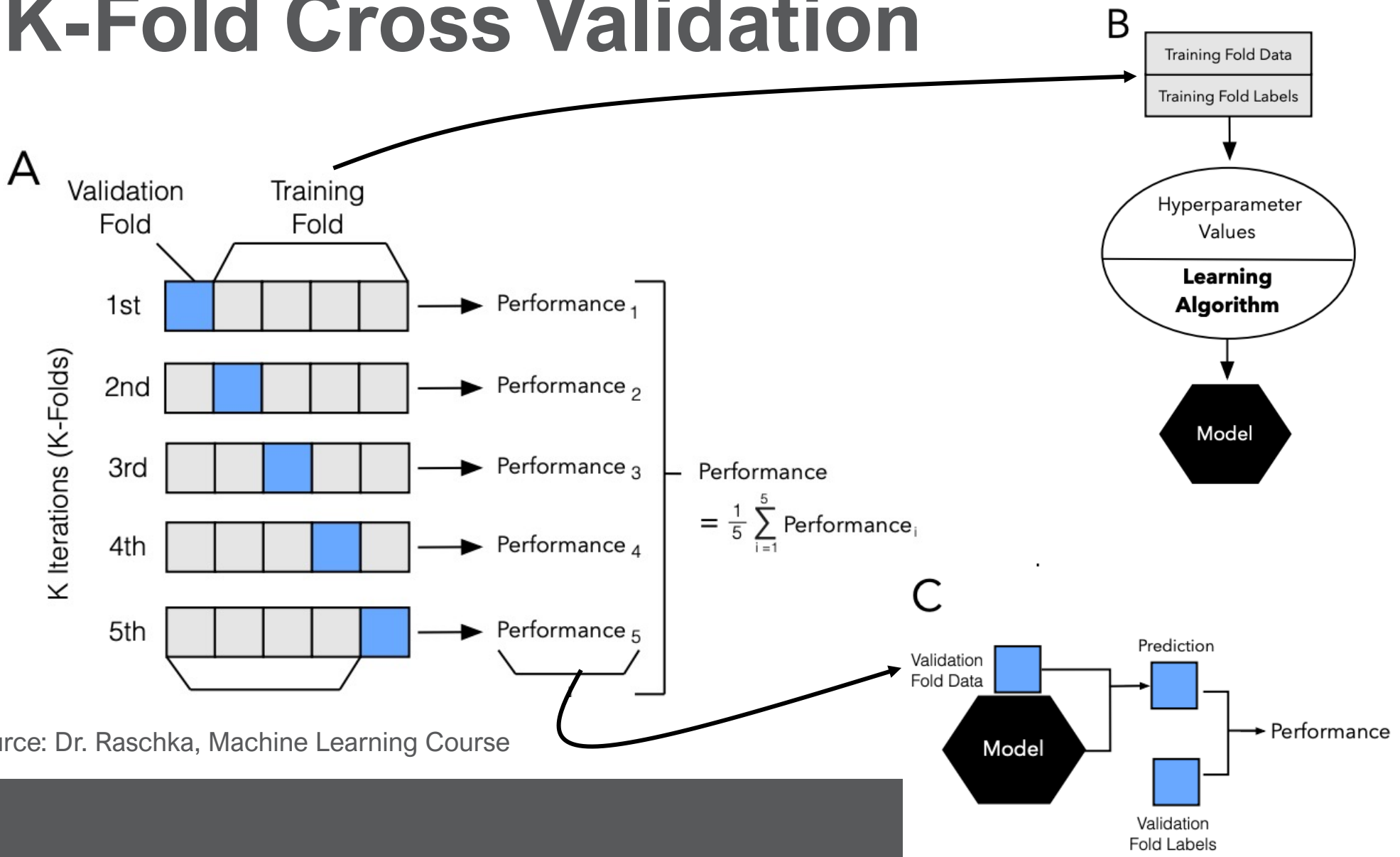
K-Fold Cross Validation



Source: Dr. Raschka, Machine Learning Course

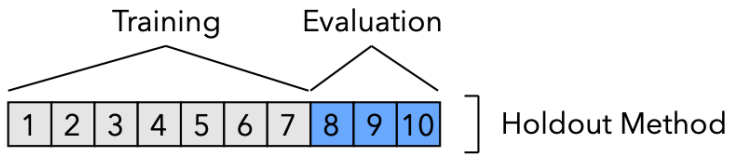
https://ttpoll.com/p/747972

K-Fold Cross Validation

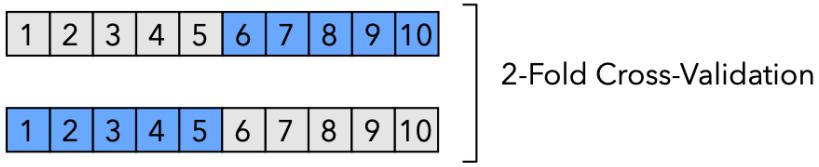


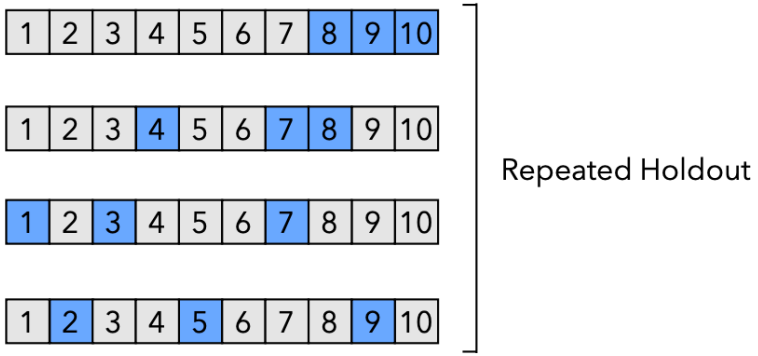
Source: Dr. Raschka, Machine Learning Course

K-Fold Cross Validation: Special Cases



k=2

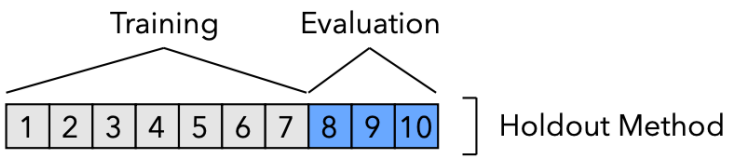




Source: Dr. Raschka, Machine Learning Course

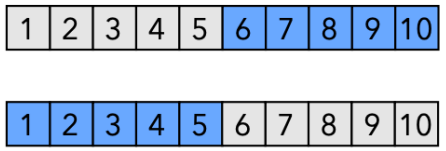
https://ttpoll.com/p/747972

K-Fold Cross Validation: Special Cases

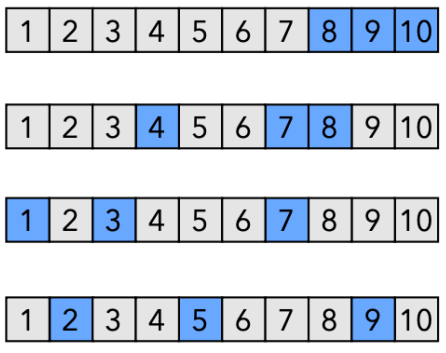


Holdout Method

k=2

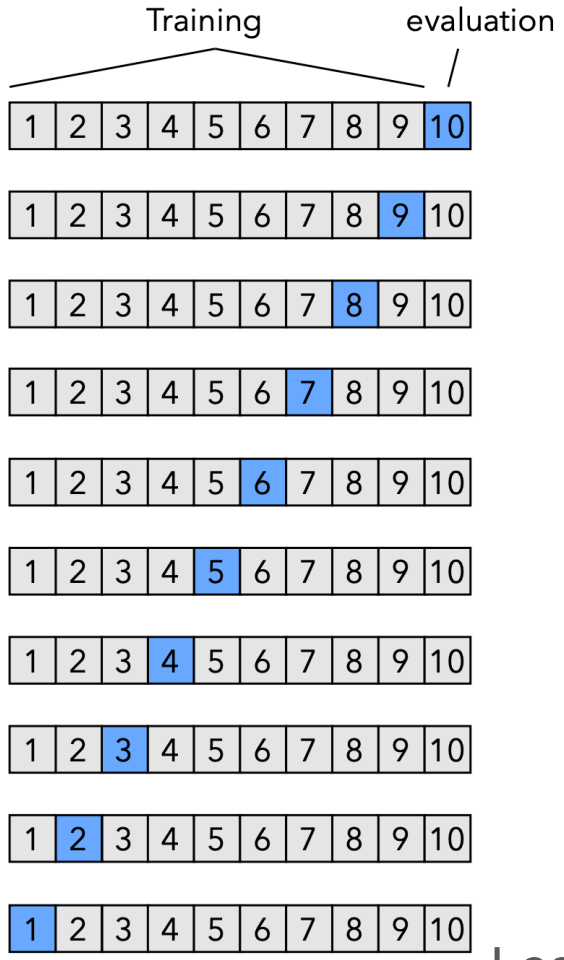


2-Fold Cross-Validation



Repeated Holdout

k=n



LOOCV
Leave-One-Out CV

Source: Dr. Raschka, Machine Learning Course

LOOCV vs Holdout

Experiment	Mean	Standard deviation
True R^2 — q^2	0.010	0.149
True R^2 — hold 50	0.028	0.184
True R^2 — hold 20	0.055	0.305
True R^2 — hold 10	0.123	0.504

1. Hawkins, D. M., Basak, S. C., & Mills, D. (2003). Assessing model fit by cross-validation. *Journal of Chemical Information and Computer Sciences*, *43*(2), 579-586.

The reported "mean" refers to the averaged difference between the true coefficients of determination (R^2) and the coefficients obtained via LOOCV (here called q^2) after repeating this procedure on multiple, different 100-example training sets

In rows 2-4, the researchers used the holdout method for fitting models to the 100-example training sets, and they evaluated the performances on holdout sets of sizes 10, 20, and 50 samples. Each experiment was repeated 75 times, and the mean column shows the average difference between the estimated R^2 and the true R^2 values.

https://ttpoll.com/p/747972

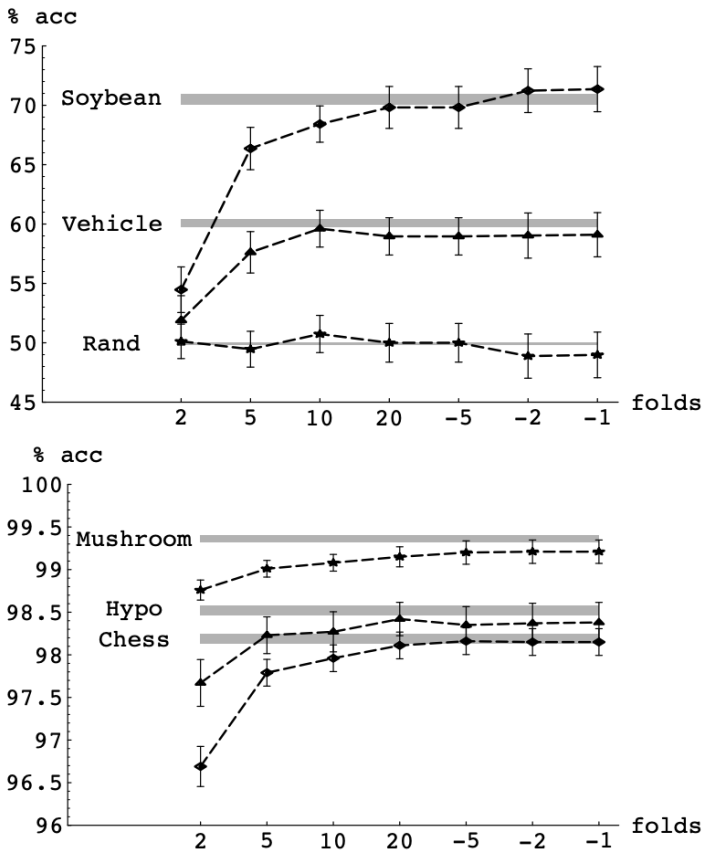


Figure 1: C4.5: The bias of cross-validation with varying folds. A negative k folds stands for leave-k-out. Error bars are 95% confidence intervals for the mean. The gray regions indicate 95% confidence intervals for the true accuracies. Note the different ranges for the accuracy axis.

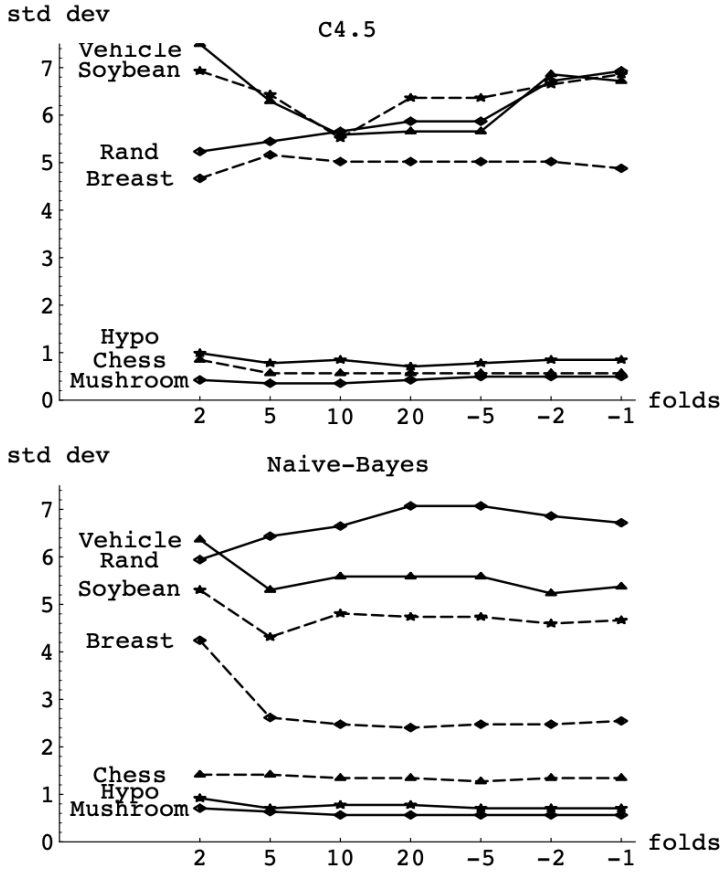


Figure 3: Cross-validation: standard deviation of accuracy (population). Different line styles are used to help differentiate between curves.

Recommendations:

- For the average dataset
  - 10-Fold CV
- For small datasets
  - LOOCV
- For Large datasets
  - 5-Fold CV

Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (Vol. 14, No. 2, pp. 1137-1145).

https://ttpoll.com/p/747972

The Law of Parsimony

Occam's Razor: "Among competing hypotheses, the one with the fewest assumptions should be selected."

https://ttpoll.com/p/747972

Pop Quiz

True or False. In K-Fold cross-validation, the dataset is partitioned in K non-overlapping training sets.

A. True

B. False

Review

- Hyperparameter tuning
  - Grid Search
  - Random Search -> 3+ parameters
  - Scale search space
- Model evaluation
  - Holdout Method
    - Confidence intervals
    - Stratified sampling
  - Repeated Holdout
    - LOOB, OOBB
  - K-Fold Cross-Validation
    - 10-Fold, LOOCV (n-Fold)

Next Lecture

- Feature selection
- Model explainability

https://ttpoll.com/p/747972

Helper Slides

https://ttpoll.com/p/747972