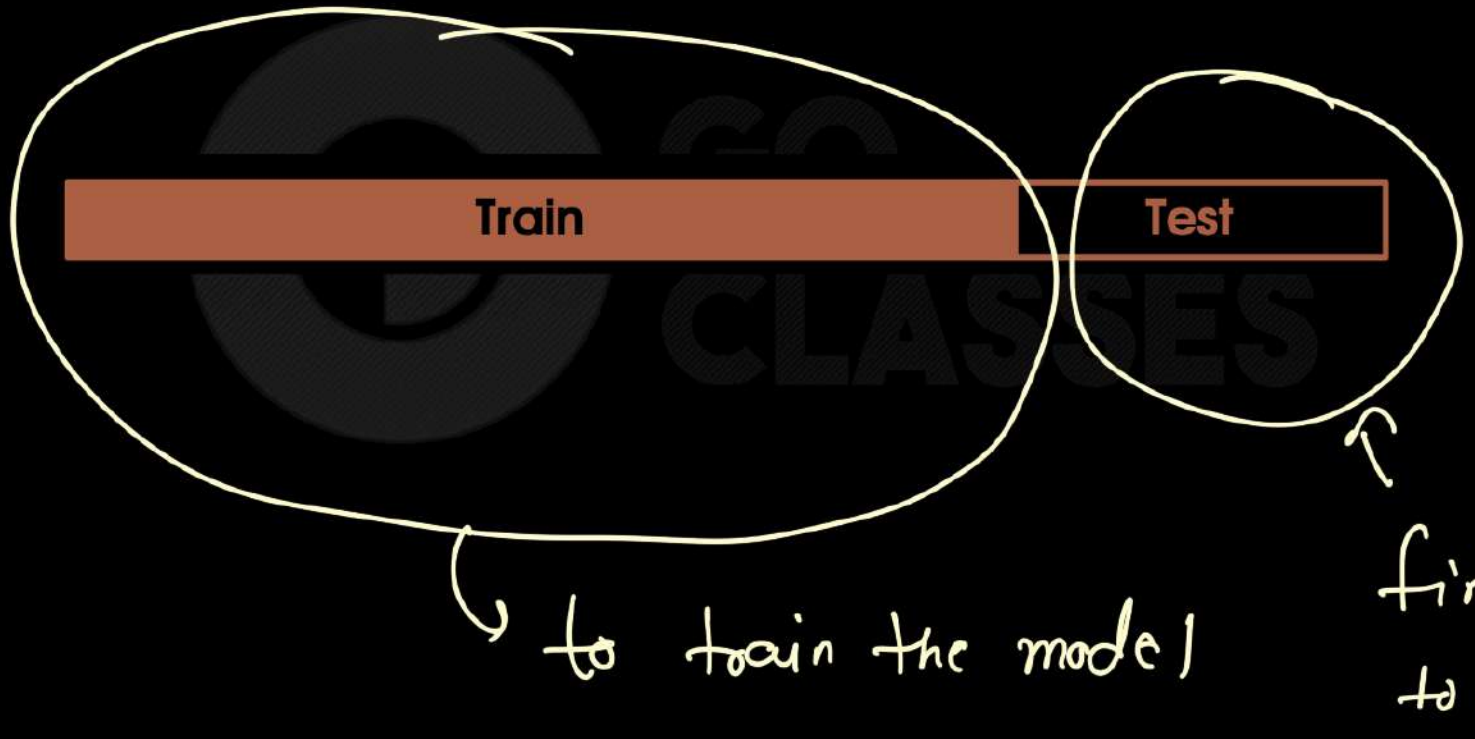
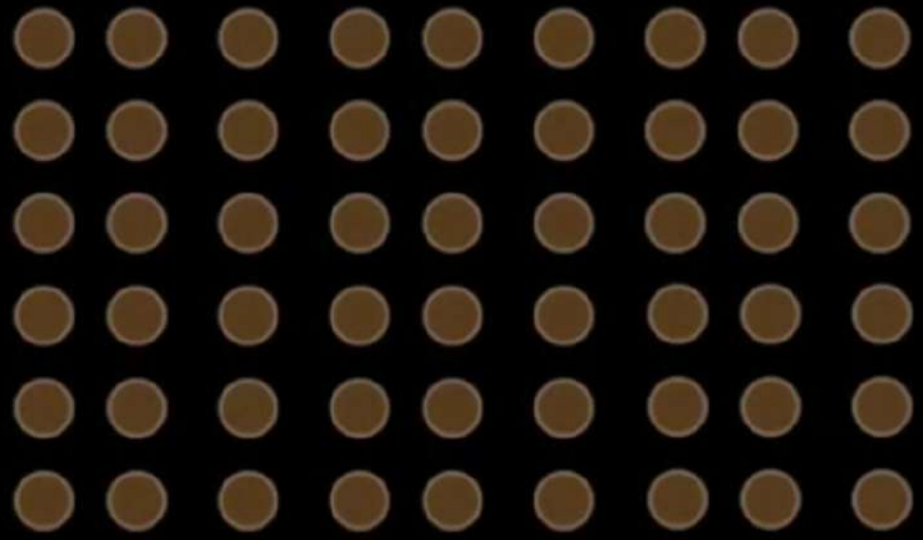
Cross Validation
and
50 Questions on CV

GATE DA Machine Learning Syllabus

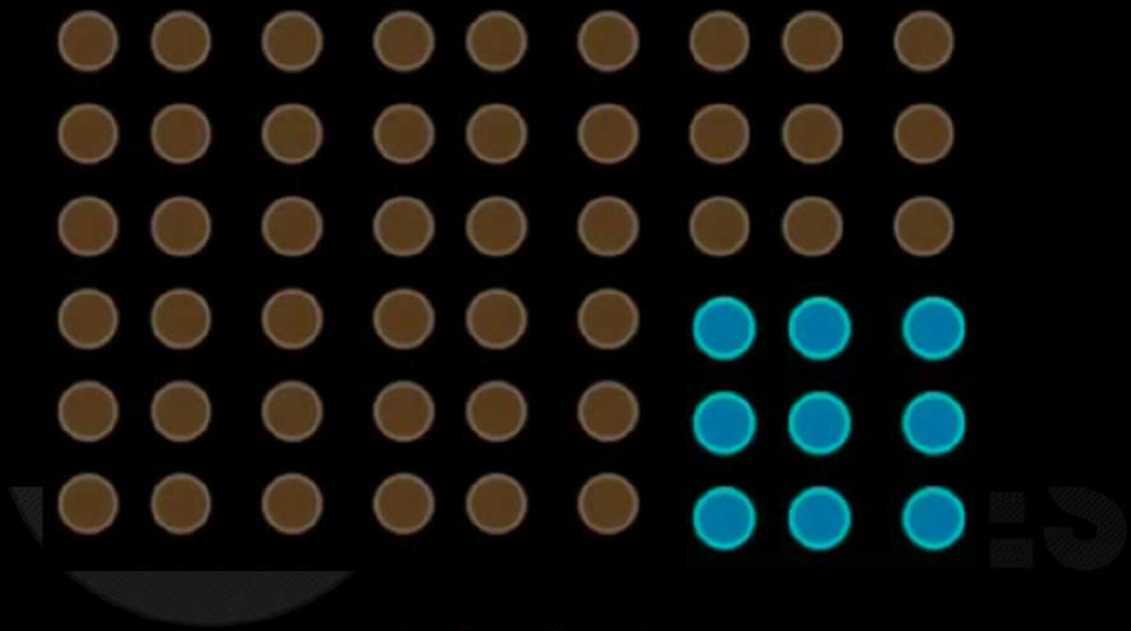
Machine Learning: (i) Supervised Learning: regression and classification problems, simple linear regression, multiple linear regression, ridge regression, logistic regression, k-nearest neighbour, naive Bayes classifier, linear discriminant analysis, support vector machine, decision trees, bias-variance trade-off, cross-validation methods such as leave-one-out (LOO) cross-validation, k-folds cross-validation, multi-layer perceptron, feed-forward neural network; (ii) Unsupervised Learning: clustering algorithms, k-means/k-medoid, hierarchical clustering, top-down, bottom-up: single-linkage, multiple-linkage, dimensionality reduction, principal component analysis.

Train set and Test Set

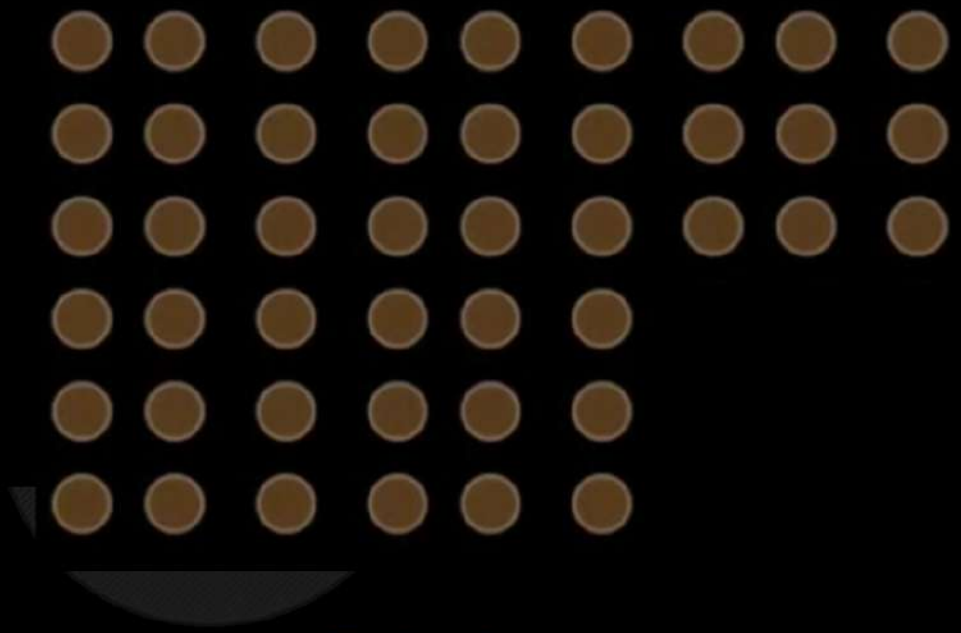




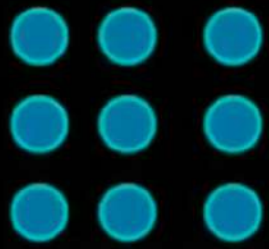
Data



Split the Data

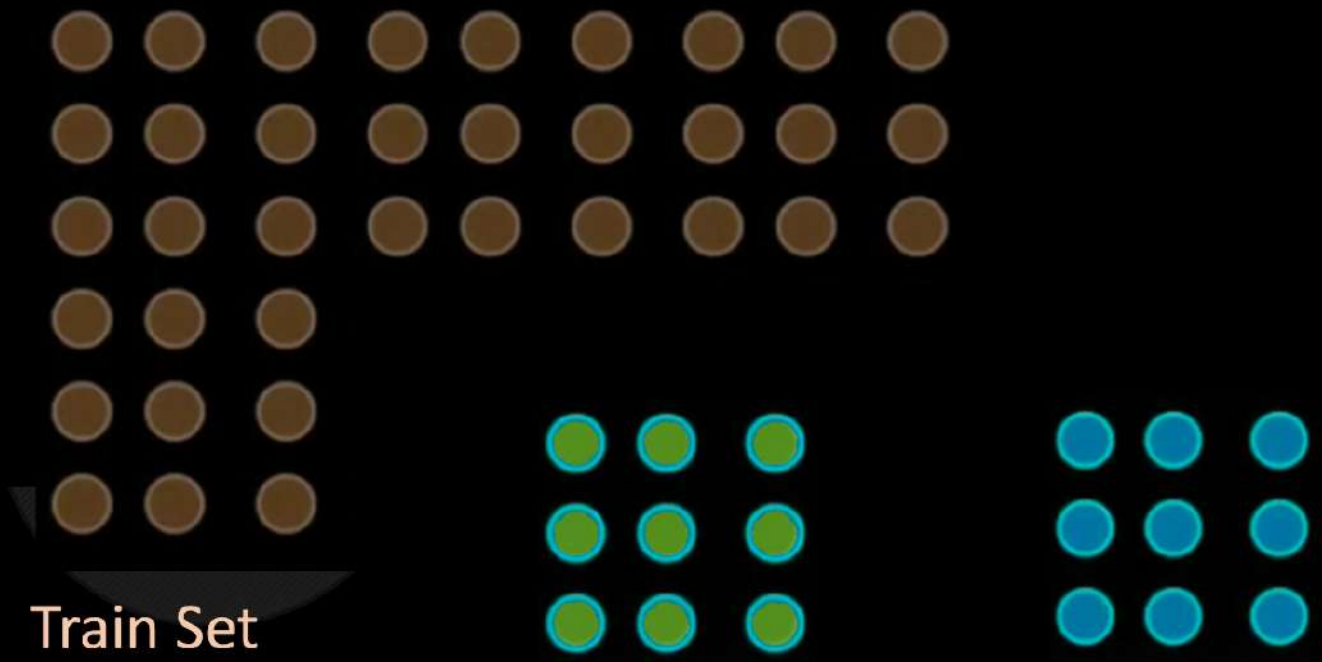


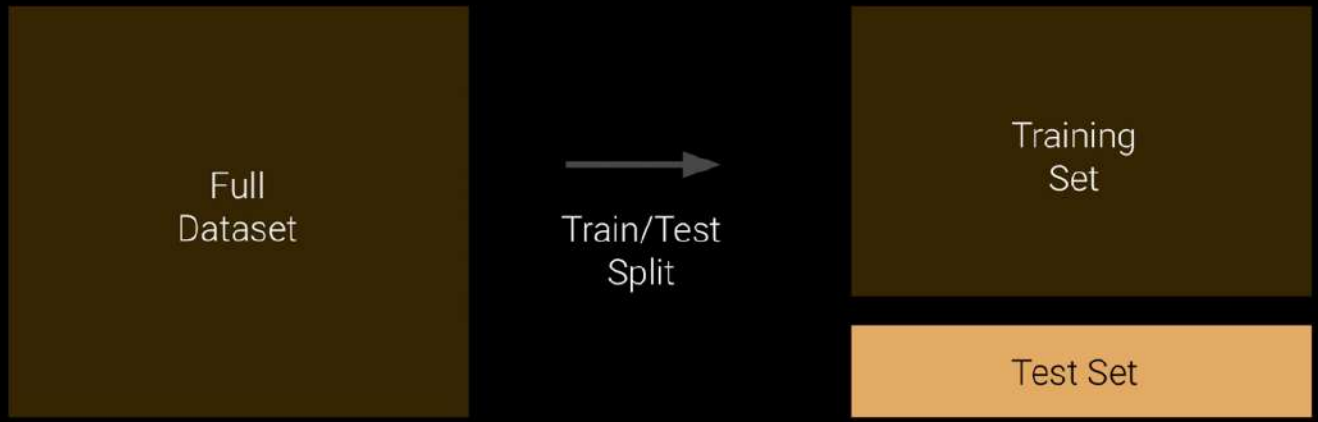
Train Set



Test Set

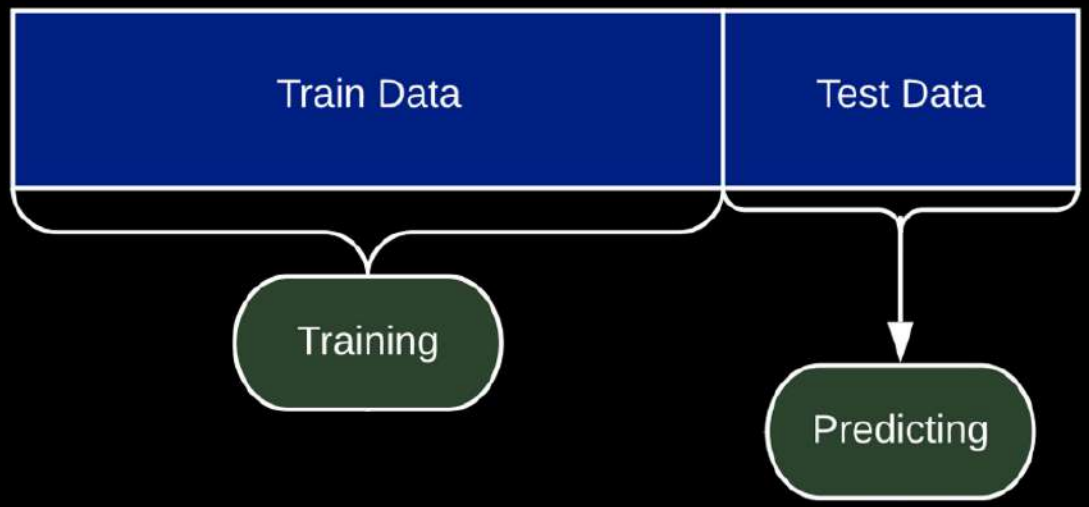

Test Set


Train Set
Validation Set
Test Set

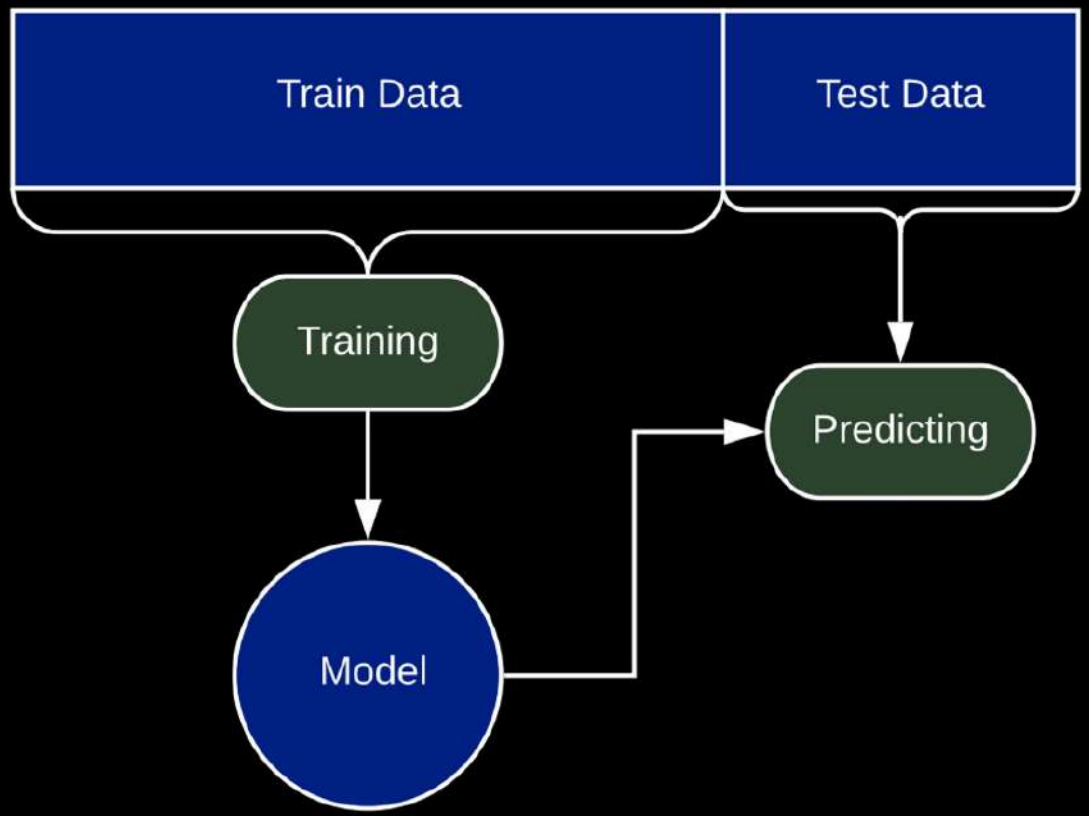

Full
Dataset
Train/Test
Split
Training
Set
Test Set


Full
Dataset
Train/Test
Split
Training
Set
Test Set
Validation
Split
Training
Set
Validation Set

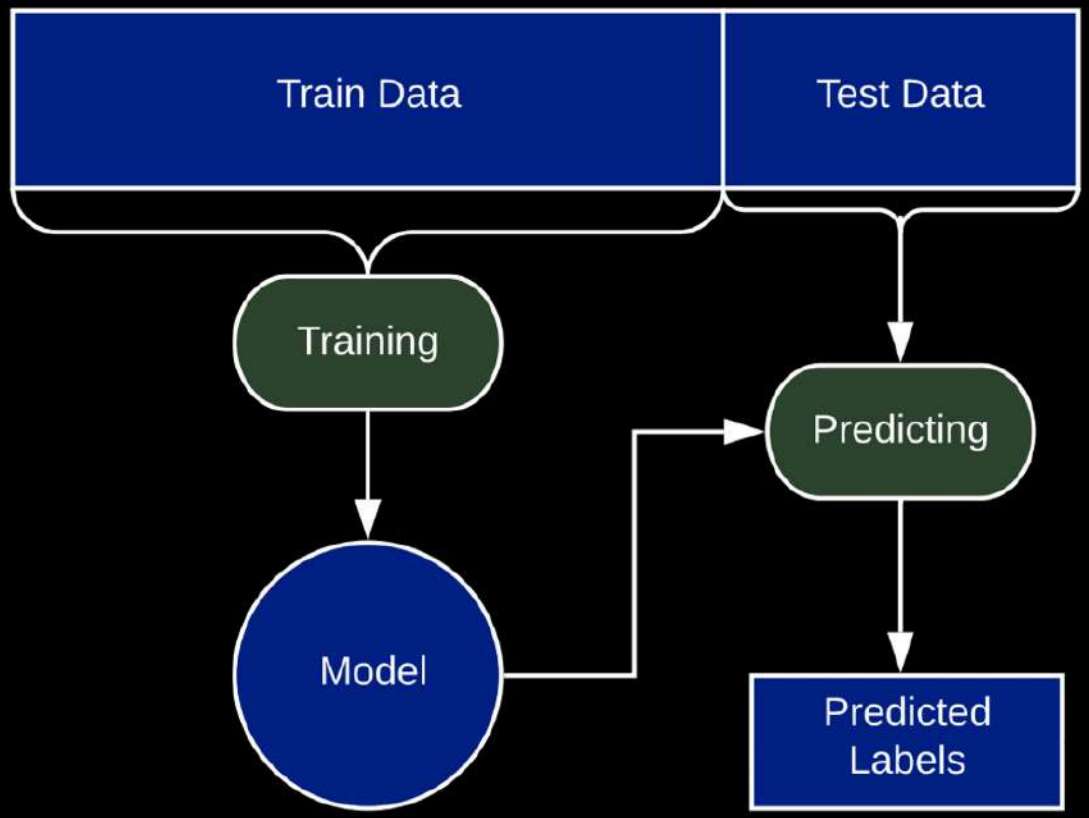
Our General Training Flow



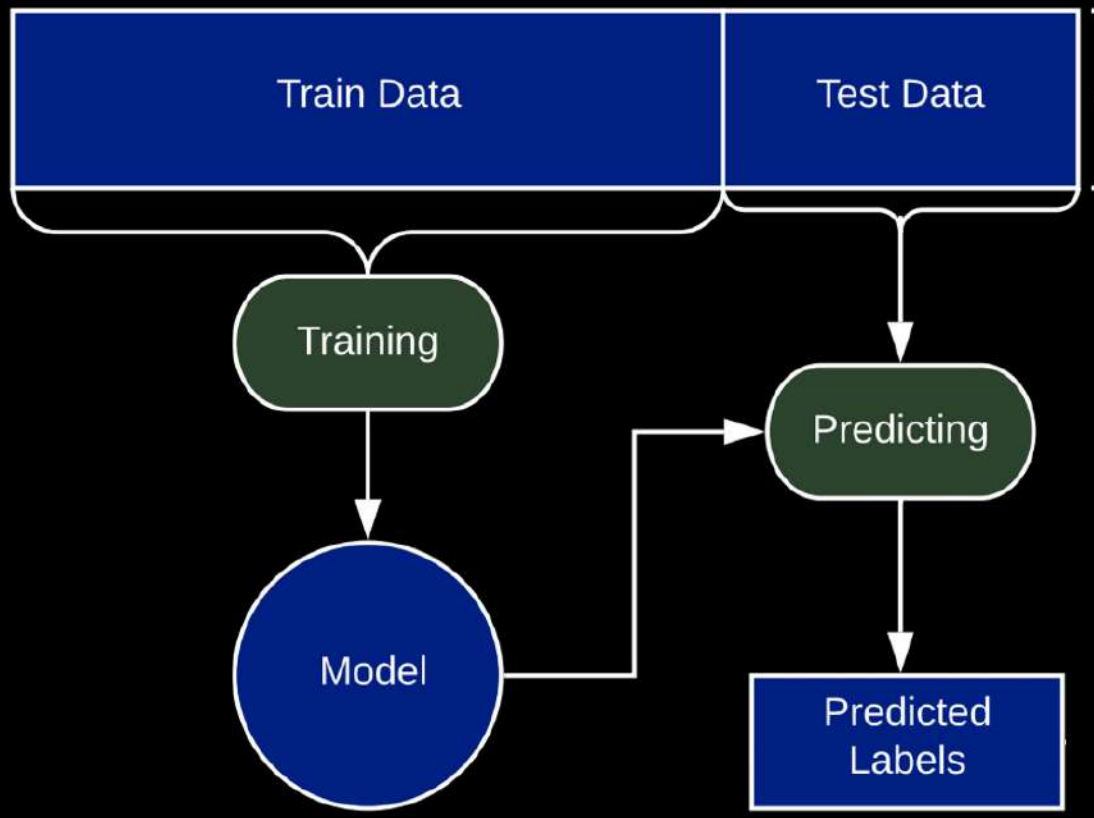
Our General Training Flow



Our General Training Flow



Our General Training Flow



$$\frac{1}{n_{test}} \sum (y_i - \hat{y}_i)^2$$

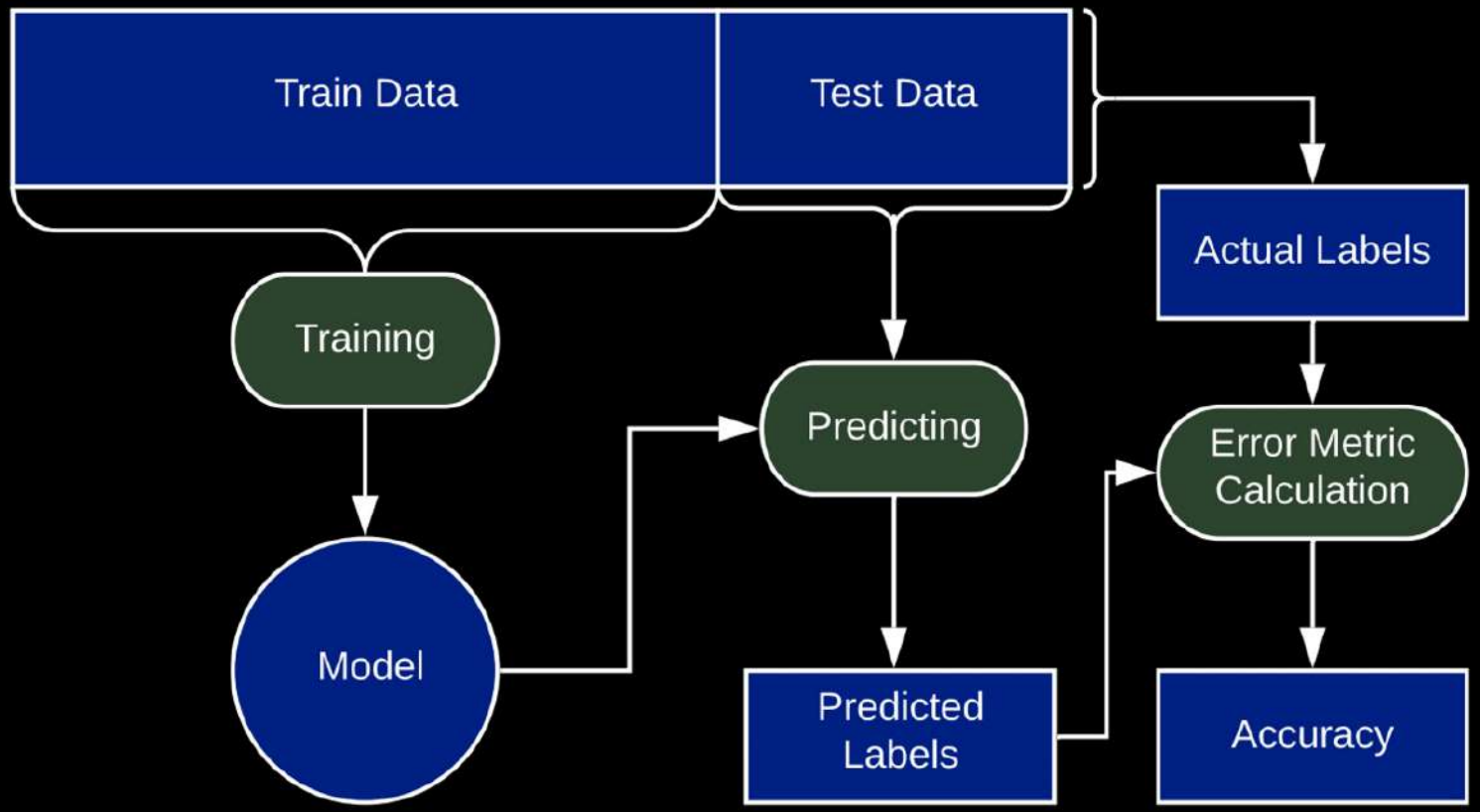
MSE (Regression)

or

Accuracy (classification)

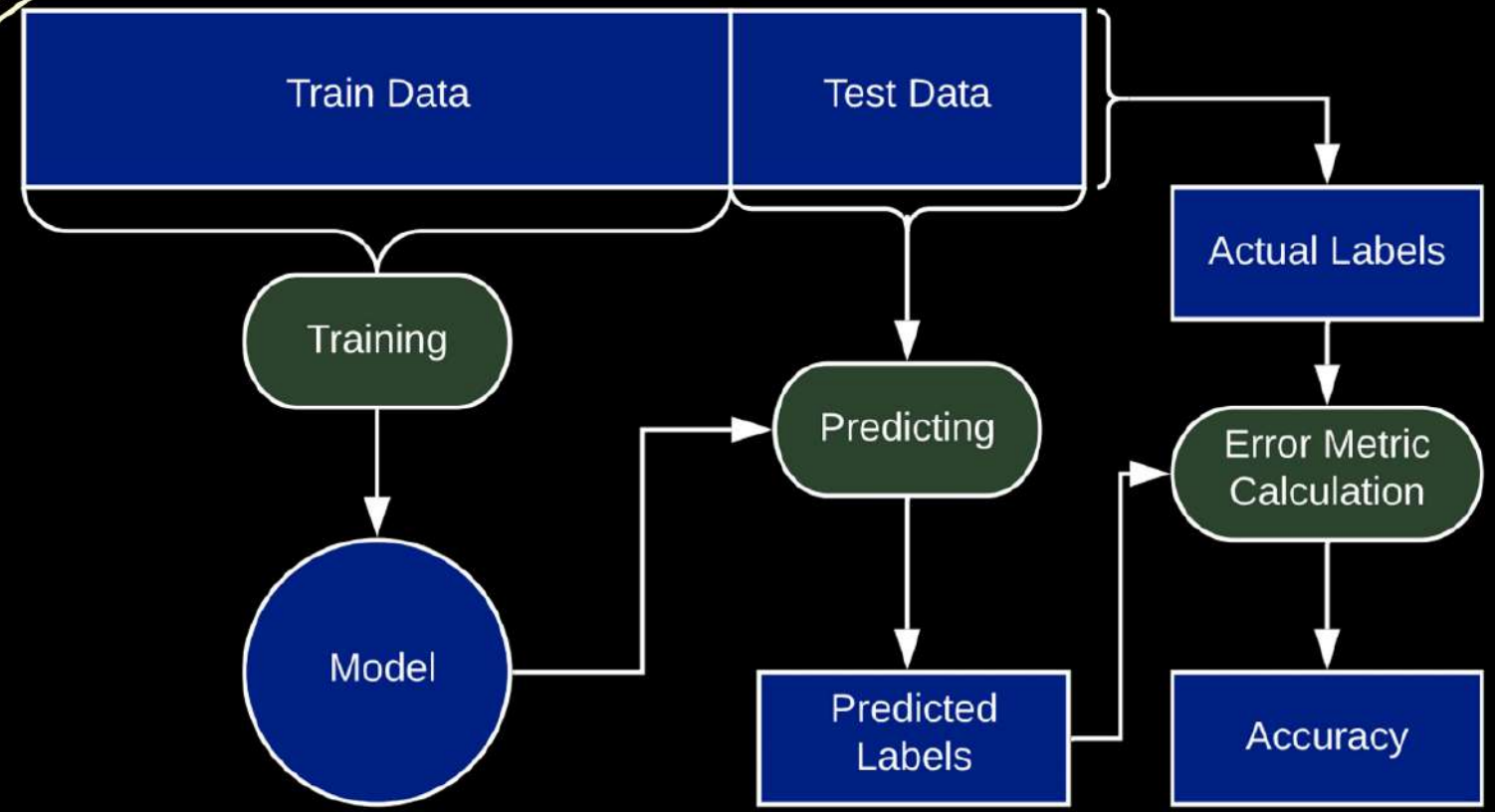
$$\frac{\#\ \text{correctly classified}}{\#\ \text{examples}}$$

Our General Training Flow



Our General Training Flow

- No way to optimise hyperparameters



No Scope to optimise hyper paramater


Full Dataset
Train/Test Split
Training Set
Test Set
Validation Split
Training Set
Validation Set

Validation Set

- To solve the problem we use a validation set
  - Examples that training algorithm does not observe
- Test examples should not be used to make choices about the model hyperparameters
- Training data is split into two disjoint parts
  - First to learn the parameters
  - Other is the validation set to estimate generalization error during or after training
    - allowing for the hyperparameters to be updated
  - Typically 80% of training data for training and 20% for validation

Hyperparameter tuning

- A machine learning model has two types of parameters.
  - The first type of parameters are the parameters that are learned through a machine learning model. → weights
  - The second type of parameters are the hyper parameter that we pass to the machine learning model.
- Normally we randomly set the value for these hyper parameters and see what parameters result in best performance.

C in SVM C ↑↑ ⇒ hard margin

- **Linear Regression:**
  - No major hyperparameters.
  - In SGD-based: **Learning rate.**
- **Ridge Regression:**
  - λ: Regularization strength.
- **Logistic Regression:**
  - In SGD-based: **Learning rate.**
- **K-Nearest Neighbors (KNN):**
  - **n_neighbors** (K): Number of neighbors.
- **Support Vector Machine (SVM):**
  - C: Trade-off between maximizing the margin and minimizing total slacks.

$$\frac{1}{N}\sum\left(y_n - \theta^T x_n\right)^2$$

actual

- These hyperparameters might address model design questions such as:
  - What degree of polynomial features should I use for my linear model?
  - What should be the maximum depth allowed for my decision tree?
  - What should be the minimum number of samples required at a leaf node in my decision tree?
  - How many trees should I include in my random forest?
  - How many neurons should I have in my neural network layer?
  - How many layers should I have in my neural network?
  - What should I set my learning rate to for gradient descent?

- Terminology:
  - **Parameters** = numeric values or structure selected by the learning algorithm
  - **Hyperparameters** = tunable aspects of the model that need to be specified before learning can happen, set outside of the training procedure

$MSE + \lambda \|\theta\|^2$

we need to know λ even before we start optimising

even before we start optimising

https://www.cs.cmu.edu/~mgormley/courses/10601-s24//slides/lecture5-models.pdf

- Terminology:
  - **Parameters** = numeric values or structure selected by the learning algorithm
  - **Hyperparameters** = tunable aspects of the model that need to be specified before learning can happen, set outside of the training procedure
- Example – kNN:
  - Model = the set of all possible nearest neighbor classifiers
  - Parameters = none! kNN is a non-parametric model
  - Hyperparameters = k

Parametric vs. Nonparametric Models

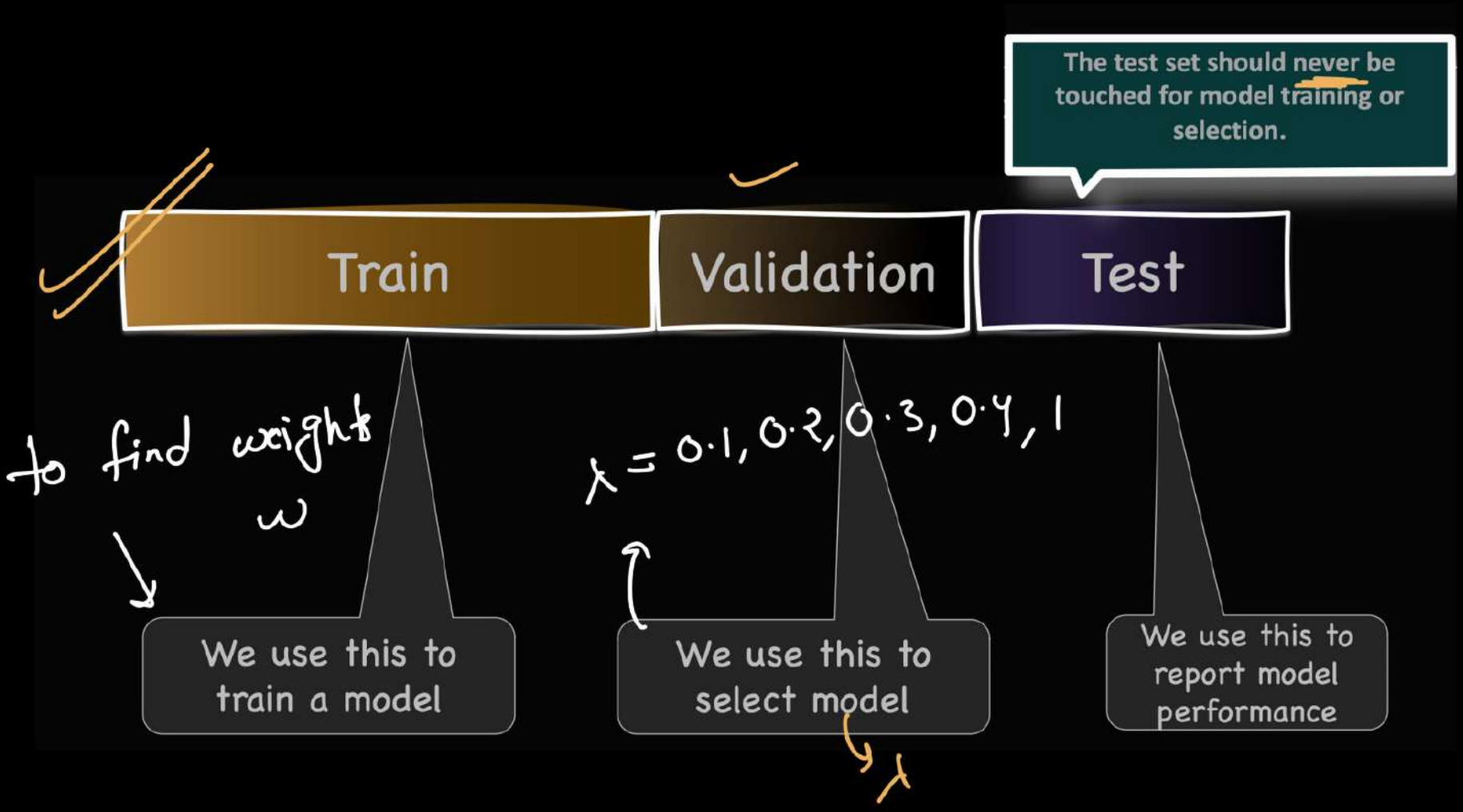
have weights

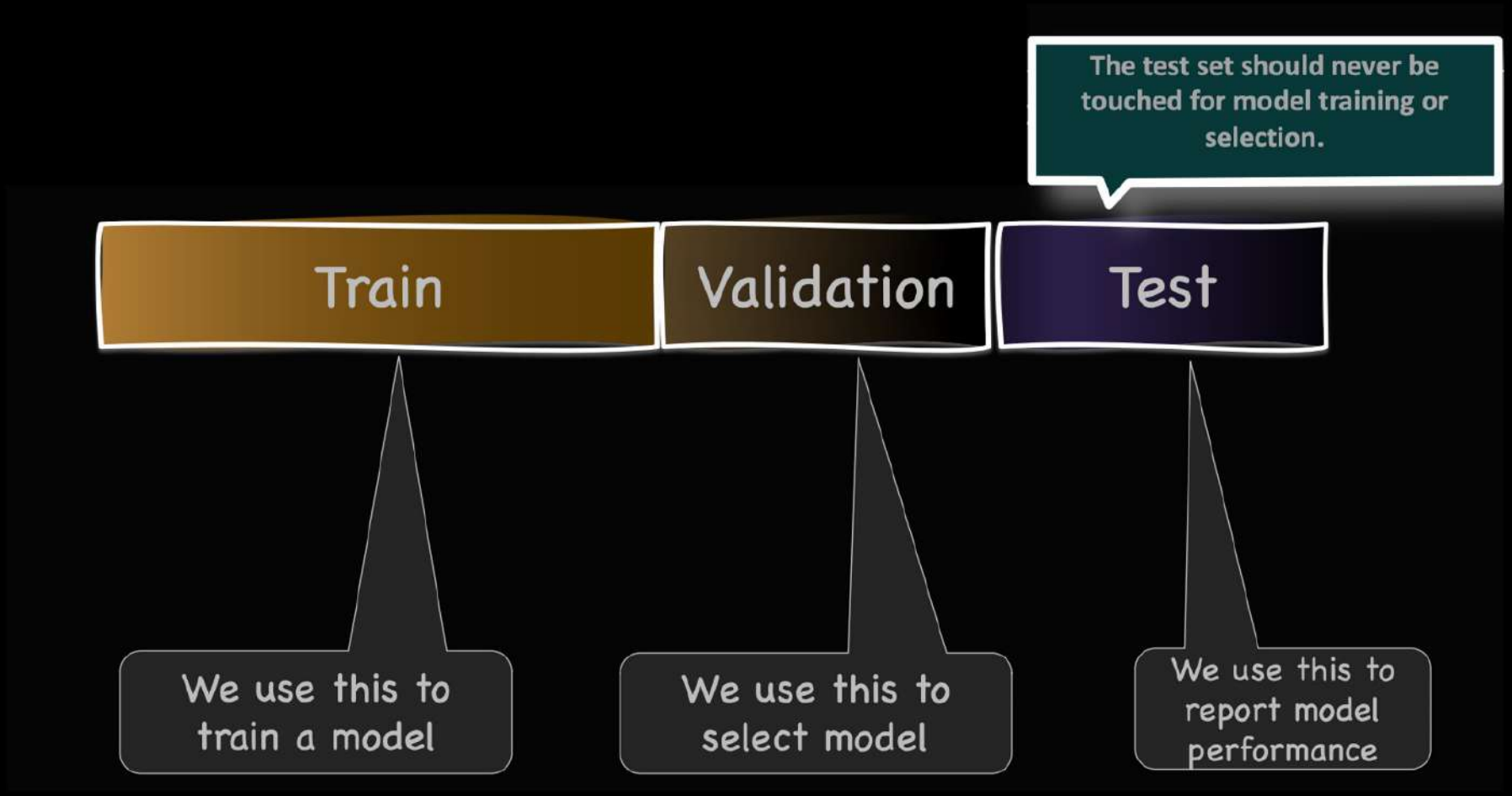
- Parametric models (e.g., Linear regression)
  - Have a parametrized form with parameters learned from training data
  - Can discard training data after parameters have been learned.

$- w^T x_{test}$

- Nonparametric models (e.g., kNN)
  - Have no parameters that are learned from training data; can still have *hyperparameters*
  - Training data generally needs to be stored in order to make predictions

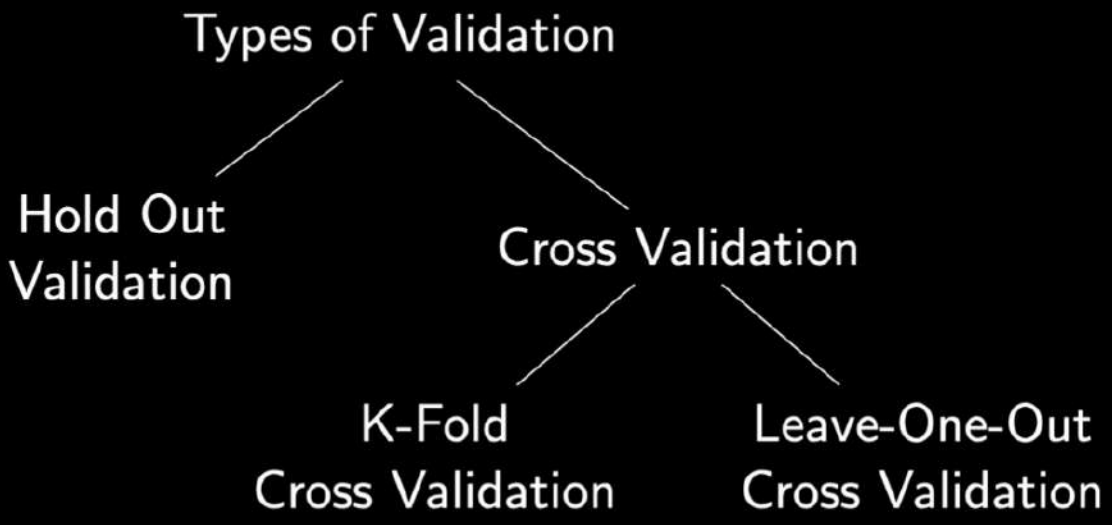
you need training data all the time


The test set should never be touched for model training or selection.
Train
Validation
Test
to find weights w
λ = 0.1, 0.2, 0.3, 0.4, 1
We use this to train a model
We use this to select model
We use this to report model performance
λ


The test set should never be touched for model training or selection.
Train
Validation
Test
We use this to train a model
We use this to select model
We use this to report model performance

Type of Validation

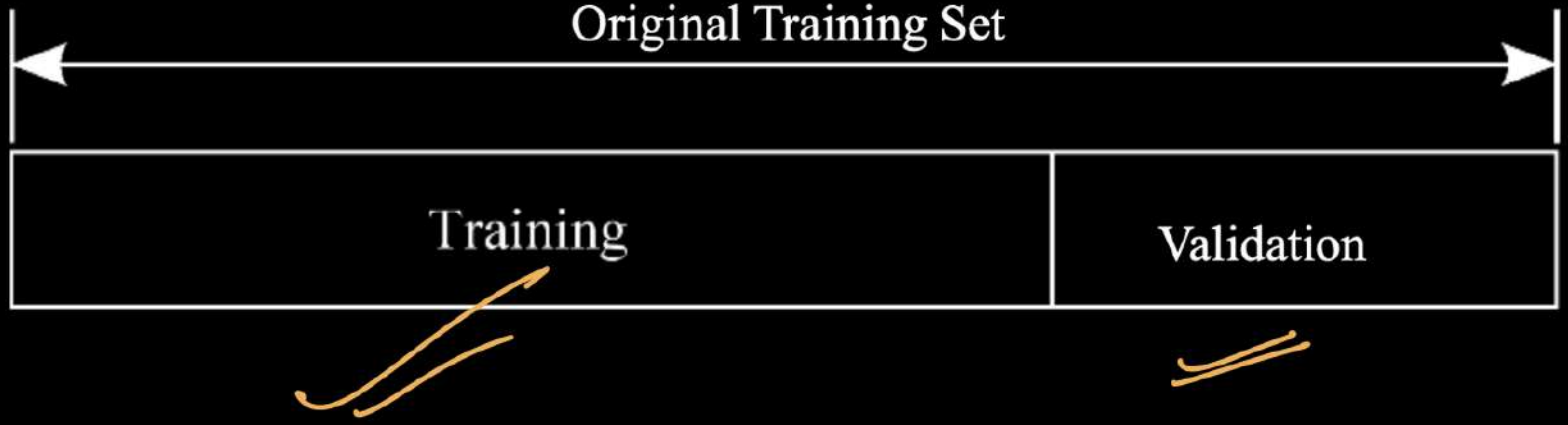
- **Hold Out Validation**
- **Cross Validation**
  - K-Fold Cross Validation
  - Leave-One-Out Cross Validation


Types of Validation
Hold Out Validation
Cross Validation
K-Fold Cross Validation
Leave-One-Out Cross Validation

Hold-Out Validation:

Hold-Out Validation: Split data into training and validation sets.

- Usually 30% as hold-out set.



```
Set lambda_choices = [10, 1, 0.1, 0.01, 0.001]
```

training on train set

λ	Acc
10	60%
1	53%
0.1	65%
0.01	69.8%
0.001	62.5%

Val. Set

1. Train on train set
2. find accuracy on the validation Set

$\lambda =$ → 60%

Best λ ? $\Rightarrow$ 0.01

```
Set lambda_choices = [10, 1, 0.1, 0.01, 0.001]
Initialize all_accuracies as an empty list

 Define training_set as all but last 'validation_size' samples of entire_dataset
 Define validation_set as last 'validation_size' samples of entire_dataset

function train_validation_split(lambda_choices):

    For each lambda in lambda_choices:

        Train model using training_set and current lambda
        Evaluate model on validation_set to get accuracy

        Append accuracy to all_accuracies

    // Find the best lambda
       find the best_lambda that gives maximum accuracy

    Return best_lambda


// Train the final model on the entire dataset using the best hyperparameter
Set final_model to train(entire_dataset, best_lambda )
```

all_accuracies = [ ]

expensive

all_acc.append( )

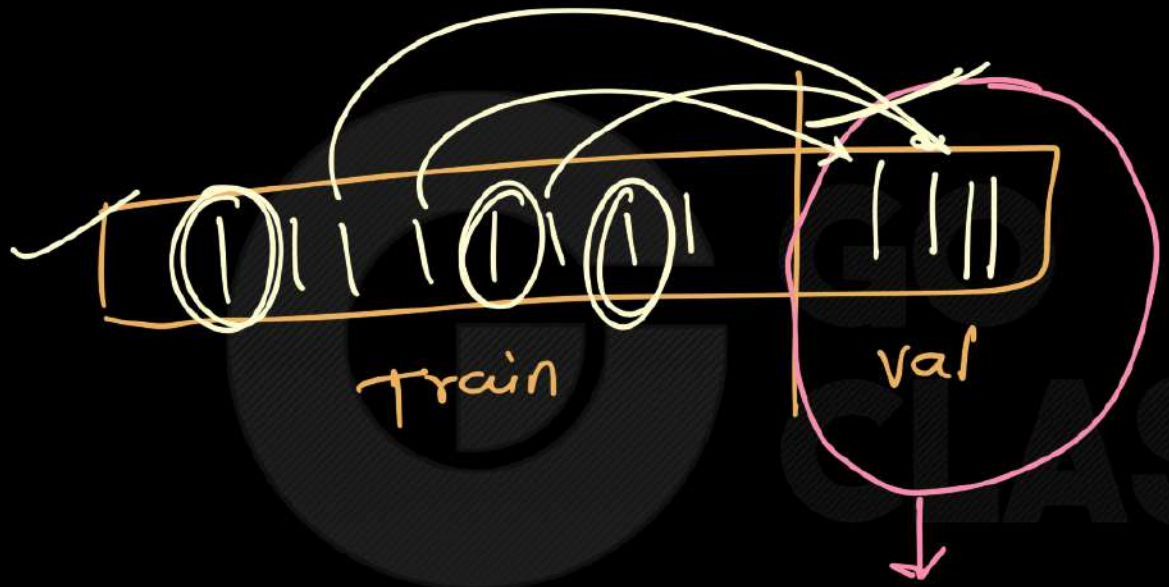
$MSE + \lambda \|\theta\|^2$

How many times i will be training?

= 5

- Objective: Optimize model performance using hyperparameters.
- Steps:
  - Split data into training and validation sets.
  - Train the model with different hyperparameter values.
  - Evaluate model accuracy on validation data.
  - Identify the best hyperparameter for final training.
- Final Training: Train the model on the complete dataset using the best hyperparameter found.

- The holdout method has basic drawback



Goal

find "λ" that perform best on unseen data (test data)

val set may not be good enough (in other words, true reflection of the test set)

- **The holdout method has basic drawback**
  - Since it is a single train-and-test experiment, the holdout estimate of error rate will be misleading if we happen to get an "unfortunate" split
  - This may happen especially when dataset is smaller.
  - This method does not provide better estimates of generalization error due to its dependence on the data split.

val

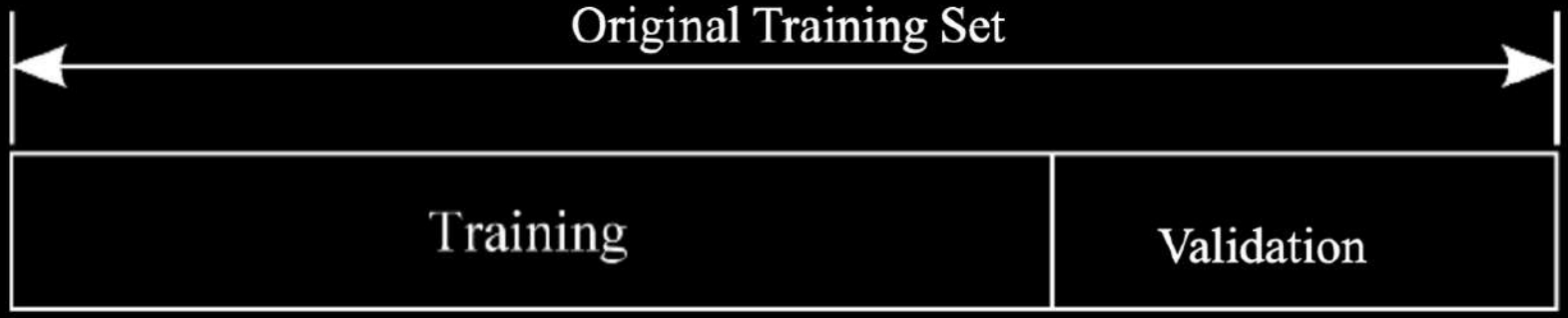
if data split is bad (val. set is very similar to train) then we do not generalize well.

main problem only when we have small data


large
large
Train
validation

Hold-Out Validation: Split data into training and validation sets.

- Usually 30% as hold-out set.



Problem:

- Estimation of error rate might be misleading

Motivation

Validation techniques are motivated by two fundamental problems in pattern recognition: model selection and performance estimation.

- **Model selection**
  - Almost invariably, all pattern recognition techniques have one or more free parameters
    - The number of neighbors in a kNN classification rule
    - The network size, learning parameters and weights in MLPs
  - How do we select the "optimal" parameter(s) or model for a given classification problem?
- **Performance estimation**

Cross Validation

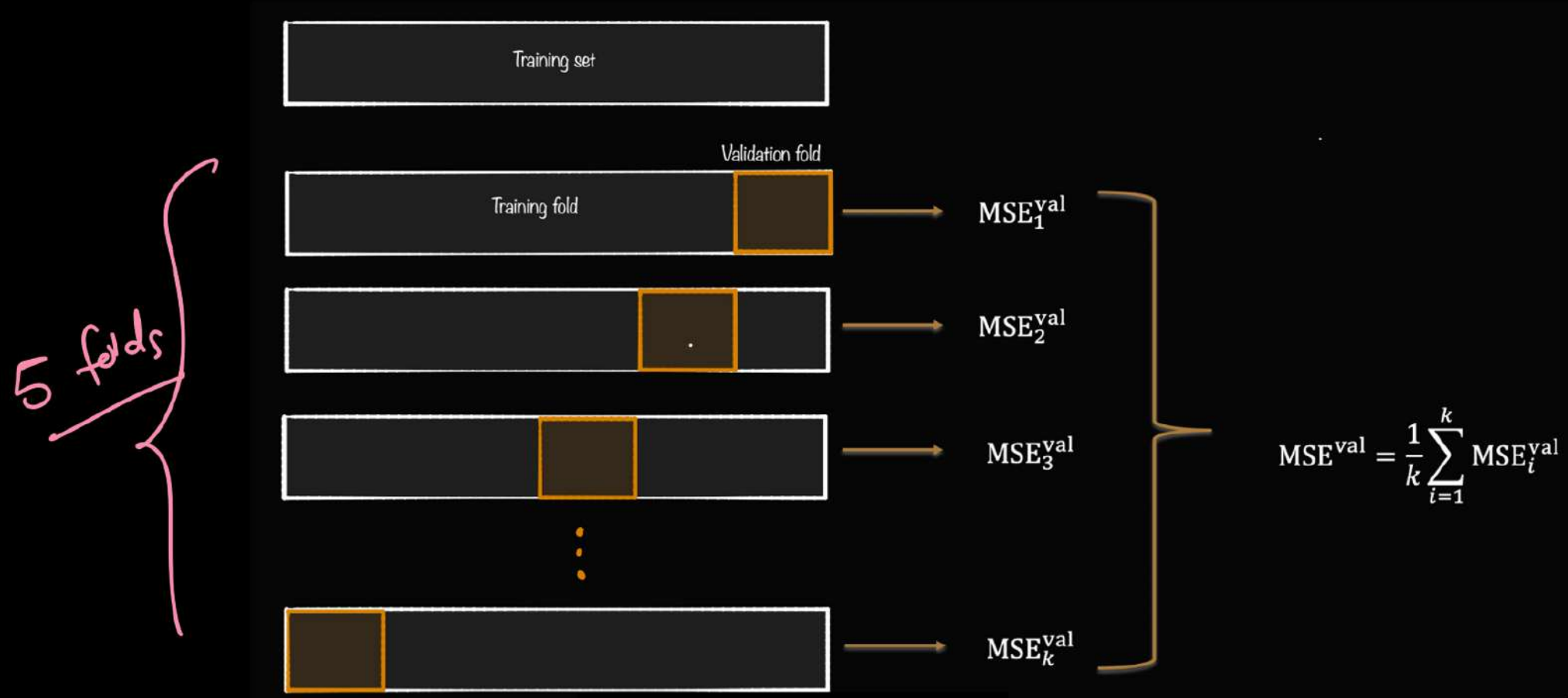
Cross Validation: Motivation

Using a single validation set to select amongst multiple models can be problematic.

One solution to the problems raised by using a single validation set is to evaluate each model on **multiple** validation sets and average the validation performance.

https://harvard-iacs.github.io/2021-CS109A/lectures/lecture05/presentation/ModelSelection_withCV.pdf

For each λ do this: $\lambda = 0.1$



For each λ do this: $\lambda = 10$

60
MSE_1^{val}
MSE_2^{val} 65
MSE_3^{val} 75
72
MSE_k^{val} 63

$$MSE^{val} = \frac{1}{k}\sum_{i=1}^{k} MSE_i^{val}$$

λ	Acc
3 → 10	
5 → 1	
5 → 0.1	
0.01	
0.001	

Val. Set

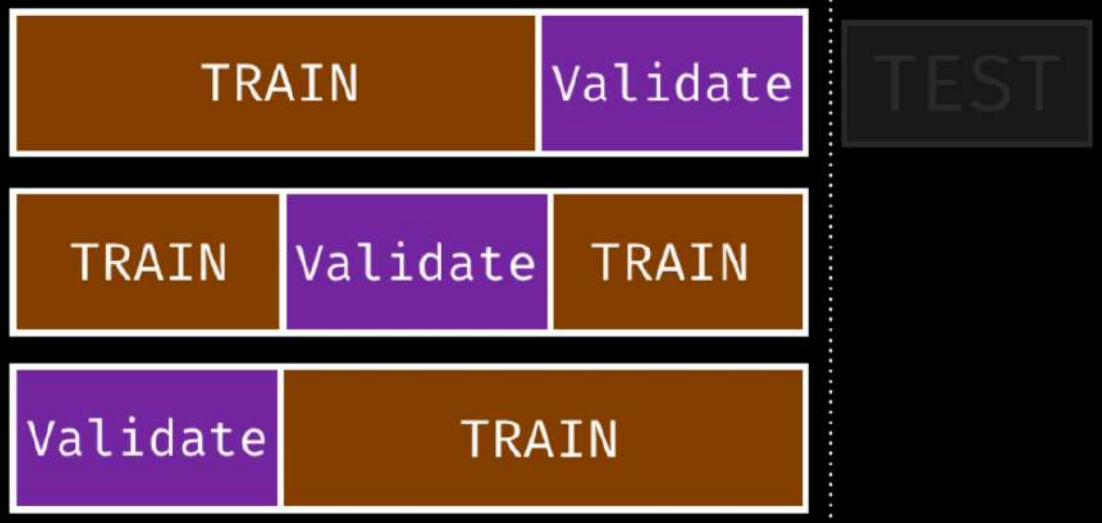
trainings

$5 \times 5 = 25$

Train-Validate Split

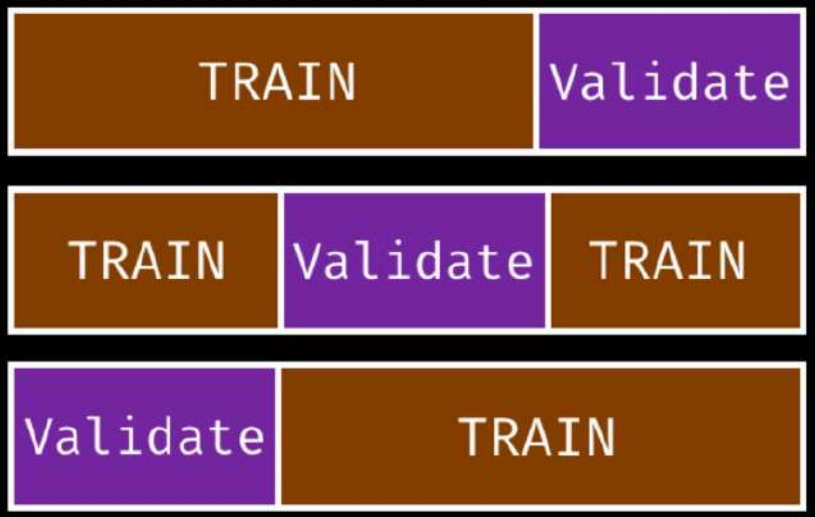
Train-Validate Split

Train-Validate Split



Question: 1

How many times we need to train the model for $\lambda = 0.1$?



Answer : 3 times

Hyperparameters

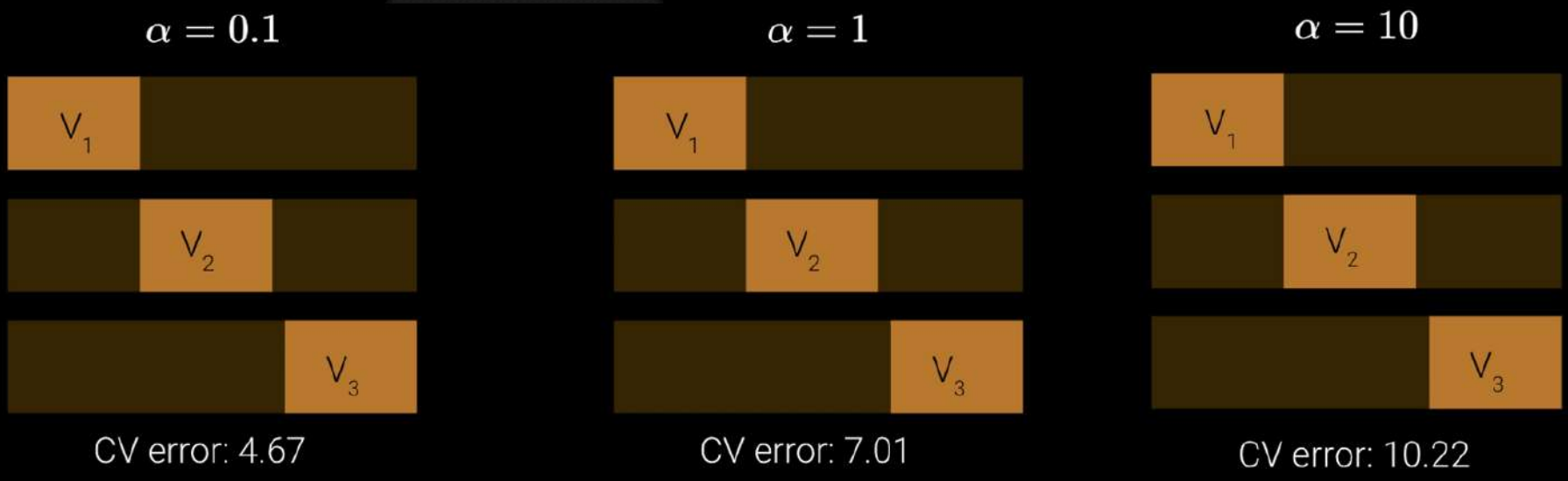
Cross-validation is often used for **hyperparameter** selection.

Hyperparameter: value in a model chosen *before* the model is fit to data

- Cannot solve for hyperparameters via calculus, OLS, gradient descent, etc – we must choose it ourselves
- Examples
  - Degree of polynomial model
  - Gradient descent learning rate, α
  - Regularization penalty, λ (to be introduced later this lecture)

Question: 2

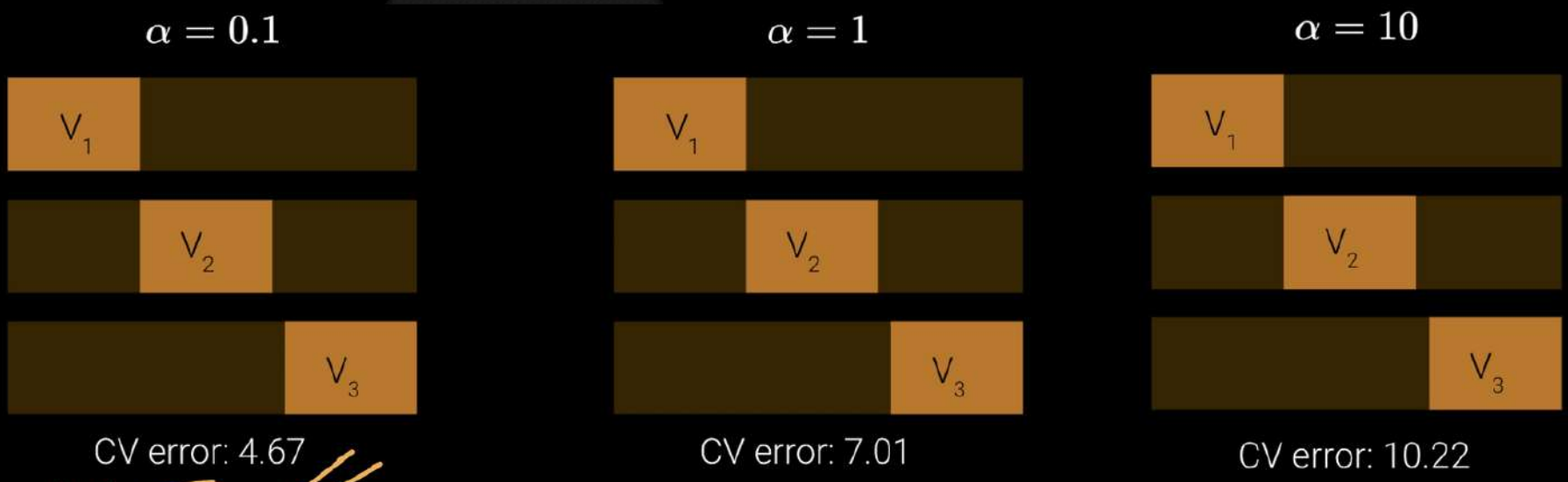
What is the best value of α ?



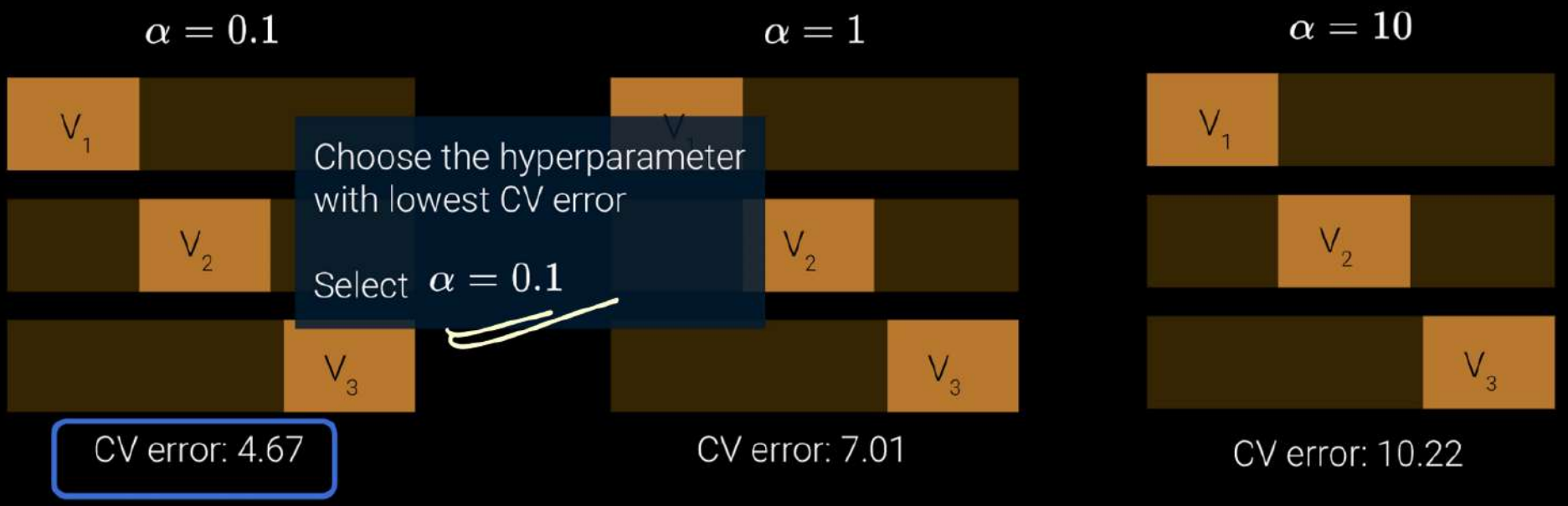
Question: 2

What is the best value of α ?

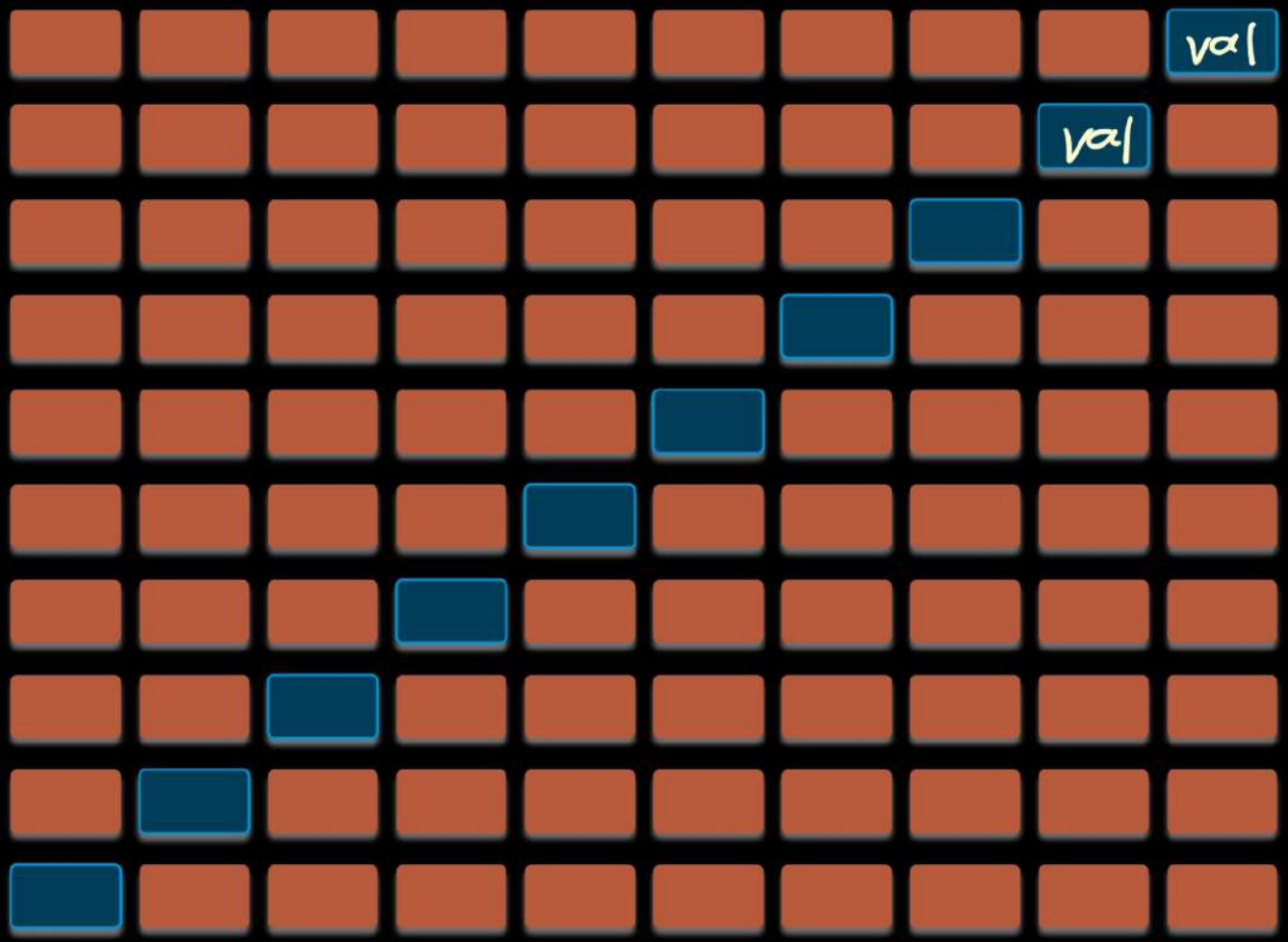
total 9 times training



Nvidia ⟹ $\alpha = 0.1$ gives you least error


$\alpha = 0.1$
$\alpha = 1$
$\alpha = 10$
V_1
V_2
V_3
Choose the hyperparameter with lowest CV error
Select $\alpha = 0.1$
CV error: 4.67
CV error: 7.01
CV error: 10.22

10-fold Cross Validation



$\lambda = 0.1$

↓

train

10 times

4-Fold Cross Validation

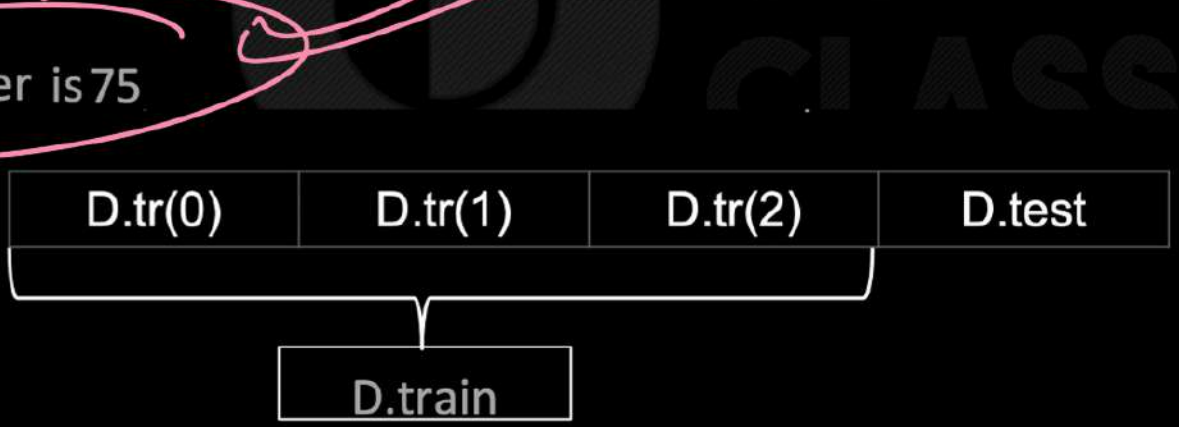
Quiz Question

Assuming 4-fold cross-validation and data D of size 100, what is the size of D.train?

Quiz Question

Assuming 4-fold cross-validation and data D of size 100, what is the size of D.train?

answer is 75



If $k = 5$, then you have 5 parts $T_1, \ldots, T_5$ you would run 5 training runs

- Train on T_1, T_2, T_3, T_4 evaluate on T_5.
- Train on T_1, T_2, T_3, T_5 evaluate on T_4.
- Train on T_1, T_2, T_4, T_5 evaluate on T_3.
- Train on T_1, T_3, T_4, T_5 evaluate on T_2.
- Train on T_2, T_3, T_4, T_5 evaluate on T_1.

5-Fold Cross-Validation Results for $\lambda = 0.1$

Folds	Train folds	Validation fold	Accuracy
12345	2,3,4,5	1	85%
12345	1,3,4,5	2	83%
12345	1,2,4,5	3	91%
12345	1,2,3,5	4	88%
12345	1,2,3,4	5	84%

Table: Cross-validation accuracy results

Average accuracy $= 88.2\%$

5-Fold Cross-Validation Results for $\lambda = 0.5$

Folds	Train folds	Validation fold	Accuracy
12345	2,3,4,5	1	80%
12345	1,3,4,5	2	82%
12345	1,2,4,5	3	89%
12345	1,2,3,5	4	87%
12345	1,2,3,4	5	85%

Table: Cross-validation accuracy results

Average accuracy $= 82.6\%$

5-Fold Cross-Validation Results for $\lambda = 1.0$

Folds	Train folds	Validation fold	Accuracy
12345	2,3,4,5	1	78%
12345	1,3,4,5	2	81%
12345	1,2,4,5	3	87%
12345	1,2,3,5	4	86%
12345	1,2,3,4	5	82%

Table: Cross-validation accuracy results

Average accuracy $= 82.8\%$

5-Fold Cross-Validation Results for $\lambda = 5.0$

Folds	Train folds	Validation fold	Accuracy
12345	2,3,4,5	1	75%
12345	1,3,4,5	2	79%
12345	1,2,4,5	3	84%
12345	1,2,3,5	4	83%
12345	1,2,3,4	5	80%

Table: Cross-validation accuracy results

Average accuracy $= 80.2\%$

Which Lambda is Best?

Lambda (λ)	Average Accuracy	Best Fold Accuracy
0.1	88.2%	91%
0.5	82.6%	89%
1.0	82.8%	87%
5.0	80.2%	84%

Table: Summary of Cross-Validation Results

Which Lambda is Best?

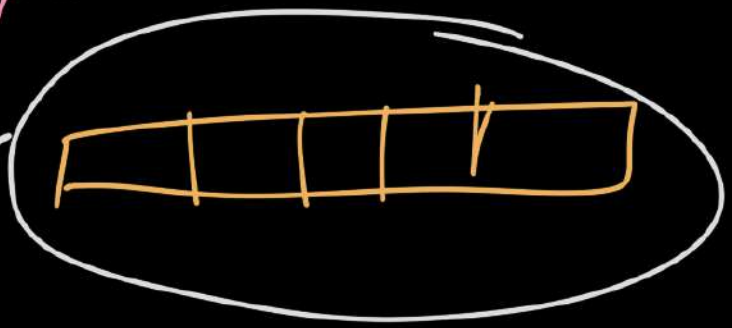
Lambda (λ)	Average Accuracy	Best Fold Accuracy
0.1	88.2%	91%
0.5	82.6%	89%
1.0	82.8%	87%
5.0	80.2%	84%

Table: Summary of Cross-Validation Results

Not useful for anything

0.1 Answer

$\lambda = 0.1$



What do you do after k-fold cross validation.

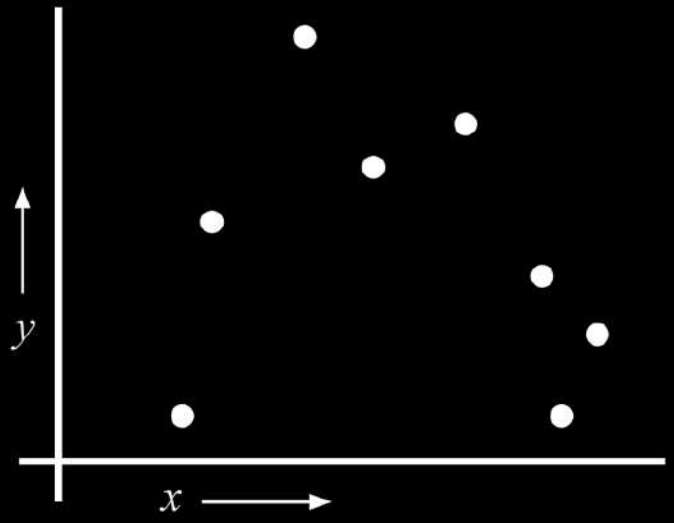
which λ

- Once you have decided which model or set of parameters to use, you then train a new model over the whole data set and use that for prediction.
- For example you could test if SVMs and Logistic regression on the same data-set and use k-fold cross validation to decide which model would perform best. Once you know this, you can then retrain on the whole data-set and use this model in production.

Pseudo Code of k-fold Cross Validation

```
lambda_choices = [10, 1, 0.1, 0.01, 0.001]
all_avg_accuracies = []


function k_fold_CV(k, lambda_choices):


    for each lambda in lambda_choices:
        accuracies = []

        for fold = 1 to K:
            validation_set = fold
            training_set = all other folds

            model = train(training_set, lambda)
            accuracy = evaluate(model, validation_set)
            accuracies.append(accuracy)

        avg_accuracy = average(accuracies)
        all_avg_accuracies.append(avg_accuracy)

    return lambda with the best avg_accuracy


# Train the final model on the entire dataset using the best hyperparameter
final_model = train(entire_dataset, best_lambda)
```

LOOCV (Leave-one-out Cross Validation)



For k=1 to n

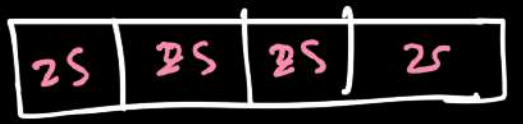
1. Let $(\mathbf{x}^k, \mathbf{y}^k)$ be the kth example
2. Temporarily remove $(\mathbf{x}^k, \mathbf{y}^k)$ from the dataset

D.train = 99 examples

kth example only

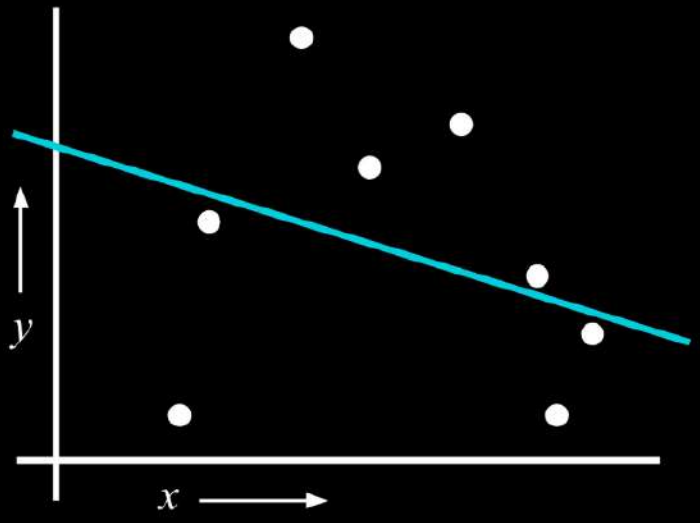
n - samples

100 = training samples



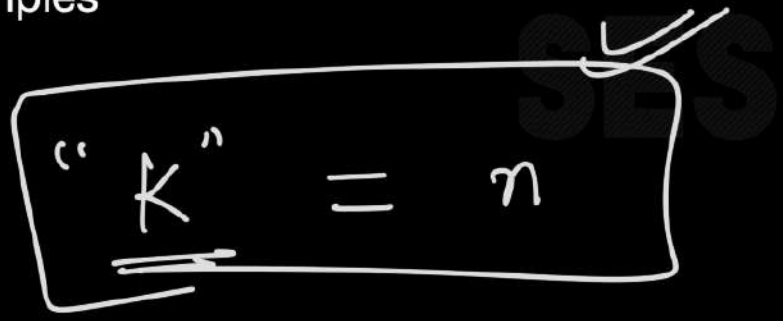
slide by Olga Veksler

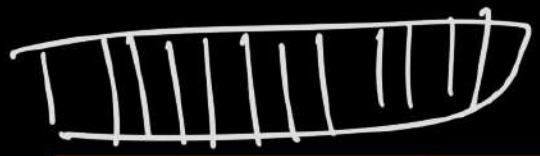
LOOCV (Leave-one-out Cross Validation)



For k=1 to n

1. Let $(\mathbf{x}^k, \mathbf{y}^k)$ be the k^{th} example
2. Temporarily remove $(\mathbf{x}^k, \mathbf{y}^k)$ from the dataset
3. Train on the remaining n-1 examples





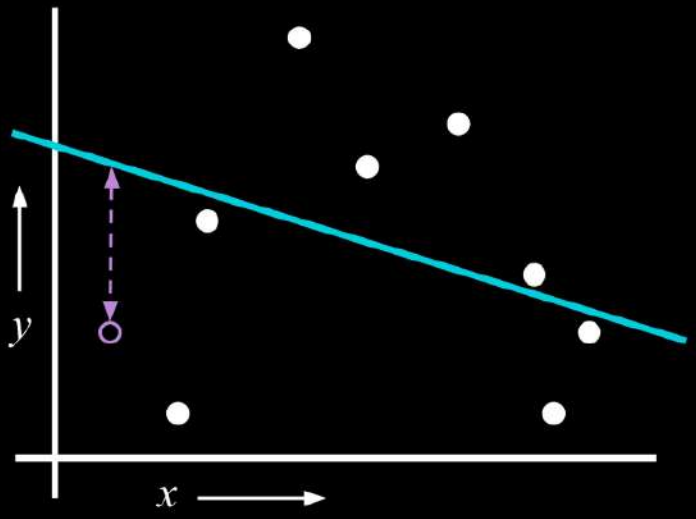
slide by Olga Veksler

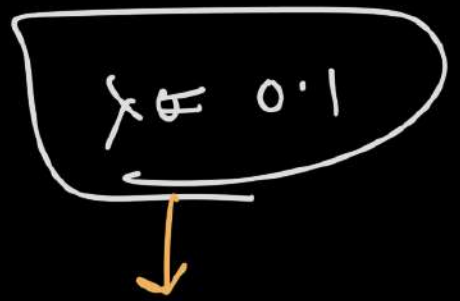
LOOCV (Leave-one-out Cross Validation)

$\lambda = 0.1$

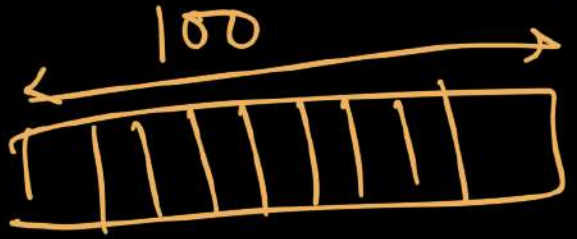
For k=1 to n

1. Let $(\mathbf{x}^k, \mathbf{y}^k)$ be the kth example
2. Temporarily remove $(\mathbf{x}^k, \mathbf{y}^k)$ from the dataset
3. Train on the remaining n-1 examples
4. Note your error on $(\mathbf{x}^k, \mathbf{y}^k)$



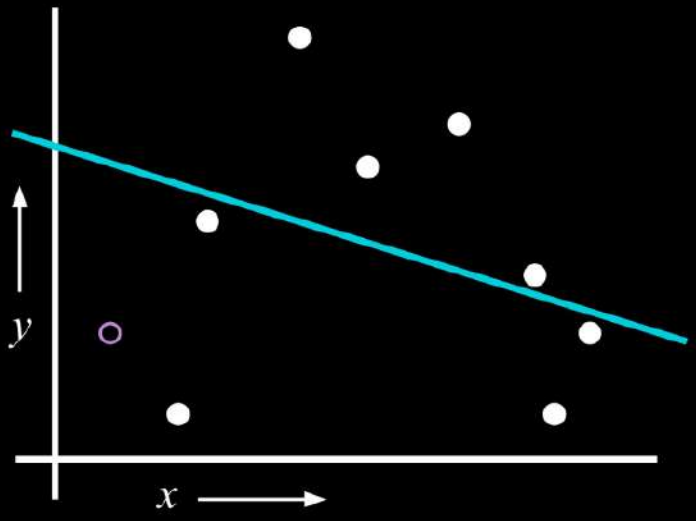


train = 100 times



slide by Olga Veksler

LOOCV (Leave-one-out Cross Validation)



For k=1 to n

1. Let $(\mathbf{x}^k, \mathbf{y}^k)$ be the k^{th} example
2. Temporarily remove $(\mathbf{x}^k, \mathbf{y}^k)$ from the dataset
3. Train on the remaining n-1 examples
4. Note your error on $(\mathbf{x}^k, \mathbf{y}^k)$

When you've done all points, report the mean error

slide by Olga Veksler

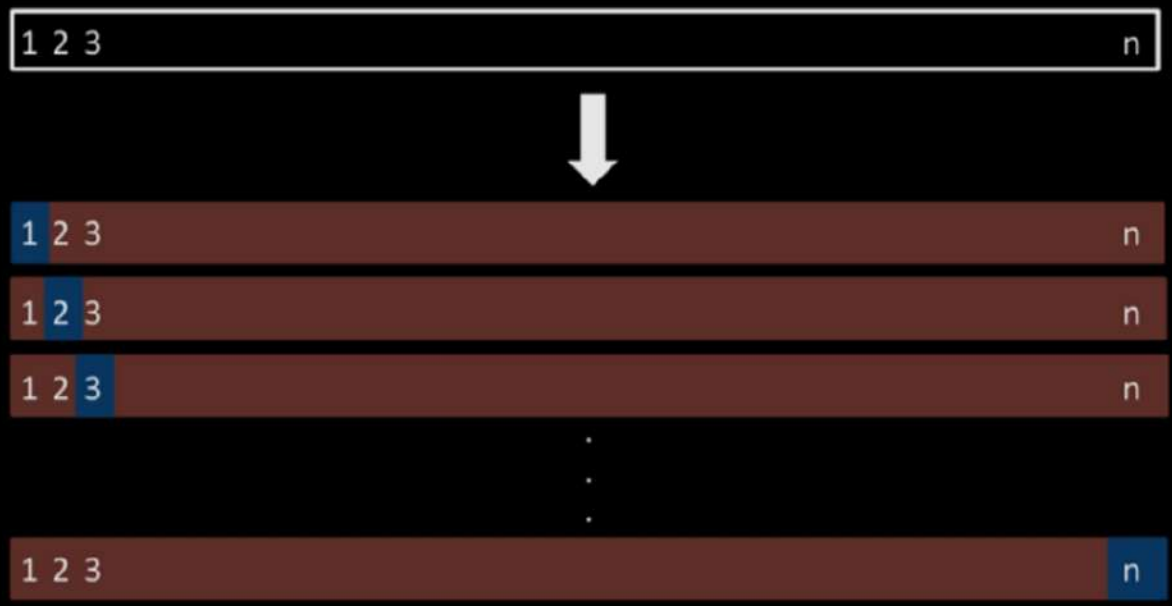
- Consider the **leave-one-out** approach.
- Let the data set be

$$\mathcal{D}_n = (\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_{n-1}, y_{n-1}), \cancel{(\boldsymbol{x}_n, y_n)}, (\boldsymbol{x}_{n+1}, y_{n+1}), \ldots, (\boldsymbol{x}_N, y_N)$$

- Remove the n-th training sample

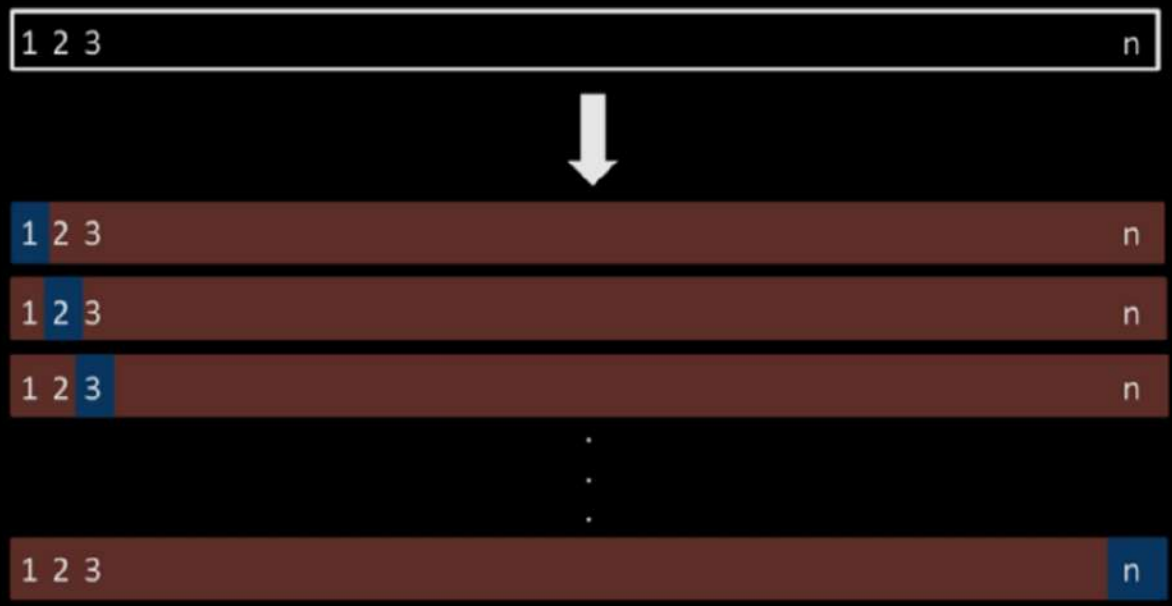
Question: 3

Which cross validation technique is being used here?



Question: 3

Which cross validation technique is being used here?



Leave one out CV

LOOCV

Question: 4 True/False

(1) True/False: Imagine we are creating a model to predict Seattle air quality. In order to improve our model performance, we should create an validation dataset from test dataset for hyperparameter tuning.

(A) True

(B) False

https://courses.cs.washington.edu/courses/cse446/24sp/sections/05/practice_midterm_solutions.pdf

(1) True/False: Imagine we are creating a model to predict Seattle air quality. In order to improve our model performance, we should create an validation dataset from test dataset for hyperparameter tuning.

(A) True

(B) False

Solution:

The solution is (B).

validation should be from training set

Question: 5

(2) To prevent overfitting of our linear model, we would apply regularization on the model. Which of the following datasets should we evaluate using our model in order to decide the right amount of regularization?

(A) Train

(B) Validation

(C) Test

(D) All of the above

https://courses.cs.washington.edu/courses/cse446/24sp/sections/05/practice_midterm_solutions.pdf

(2) To prevent overfitting of our linear model, we would apply regularization on the model. Which of the following datasets should we evaluate using our model in order to decide the right amount of regularization?

(A) Train

(B) Validation

(C) Test

(D) All of the above

The solution is (B).

Question: 6

(2) (2 points) Your friend needs help selecting between two linear regression models, desiring the one that generalizes best. Which model should they use? (Circle your answer and provide a short justification.)

Model 1

Training MSE: 100, Validation MSE: 105

Model 2

Training MSE: 50, Validation MSE: 110

(2) (2 points) Your friend needs help selecting between two linear regression models, desiring the one that generalizes best. Which model should they use? (Circle your answer and provide a short justification.)

Model 1

Training MSE: 100, Validation MSE: 105

Model 2

Training MSE: 50, Validation MSE: 110

Solution: Model 1 since even though it has worse training error, it yields slightly better validation error. Since validation error is a better predictor of future performance, this model is more likely to generalize best.

https://cseweb.ucsd.edu/classes/fa19/cse151-a/pract_quiz1-sol.pdf

Question: 7 True/False

(8) [1 Pt.] It is important to frequently evaluate models on the test data throughout the process of model development.

(8) [1 Pt.] It is important to frequently evaluate models on the test data throughout the process of model development.

Solution: **False.** Noooooooooooo. Once test data is used it is no longer test data. You should create validation datasets or use cross-validation procedures to evaluate models.

Question: 8

(e) [2 Pts] To choose the most optimal λ, Namjoon uses 5-fold cross-validation to train 5 models and stores the appropriate root-mean square error for each fold and λ. Given the collected statistics below about the **cross-validation error for each model**, help Namjoon choose the values of λ (among the 5 listed) that is the **best** to use with exponential regularization.

	λ = 0.001000	λ = 0.010000	λ = 0.100000	λ = 1.000000	λ = 10.000000
mean	105.017370	62.983541	126.653039	2873.545309	inf
min	8.027953	7.550870	56.968206	1380.204219	inf

- ○ A. $\lambda = 0.001$
- ○ B. $\lambda = 0.01$
- ○ C. $\lambda = 0.1$
- ○ D. $\lambda = 1$
- ○ E. $\lambda = 10$

(e) [2 Pts] To choose the most optimal λ, Namjoon uses 5-fold cross-validation to train 5 models and stores the appropriate root-mean square error for each fold and λ. Given the collected statistics below about the **cross-validation error for each model**, help Namjoon choose the values of λ (among the 5 listed) that is the **best** to use with exponential regularization.

	λ = 0.001000	λ = 0.010000	λ = 0.100000	λ = 1.000000	λ = 10.000000
mean	105.017370	62.983541	126.653039	2873.545309	inf
min	8.027953	7.550870	56.968206	1380.204219	inf

irrelevant info

this Never matter to us

- ○ A. $\lambda = 0.001$
- ○ **B.** $\lambda = 0.01$
- ○ C. $\lambda = 0.1$
- ○ D. $\lambda = 1$
- ○ E. $\lambda = 10$

Solution: We wish to use the hyperparameter associated with the smallest mean cross-validated error. In this case, that is $\lambda = 0.01$.

Question: 9

(b) [2 Pts] Jessica decides to use LASSO regression. If Jessica has 5 candidate values of regularizer parameter λ, how many validation errors will she calculate if she runs 4-fold cross-validation?

Answer = [ ]

$\lambda \Rightarrow 5$

$4 \Rightarrow$ folds

20 times training will happen

(b) [2 Pts] Jessica decides to use LASSO regression. If Jessica has 5 candidate values of regularizer parameter λ, how many validation errors will she calculate if she runs 4-fold cross-validation?

Answer = [ ]

Solution: In cross-validation, we must calculate one validation error for each fold for each possible parameter (λ in this case). As a result, our final answer is $5 * 4 = 20$.

The number of columns in our feature matrix does not impact how many validation errors we need to calculate.

Question: 10

14. What is cross-validation **not** used for?
    (a) To evaluate the performance of a machine learning model on unseen data.
    (b) To select a model's hyperparameters.
    (c) To determine the generalization of a machine learning model.
    (d) To train multiple machine learning models on different datasets.

https://courses.cs.washington.edu/courses/cse446/23au/exams/pastexams/23sp-midterm-solutions.pdf

Question: 10

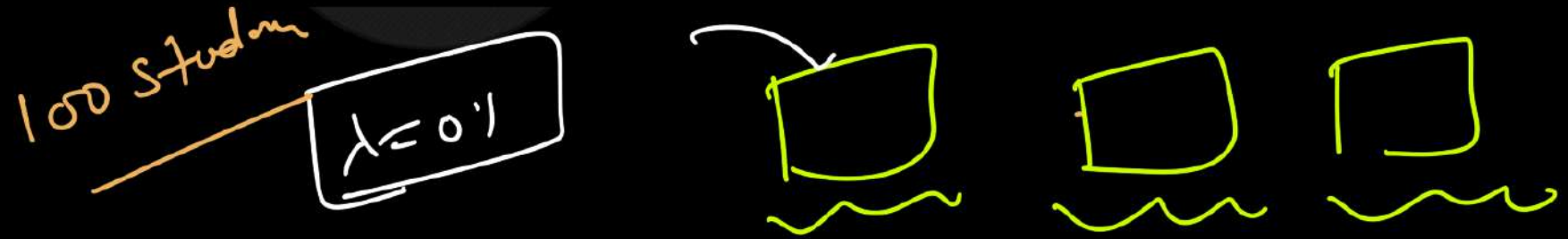
14. What is cross-validation **not** used for?

(a) To evaluate the performance of a machine learning model on unseen data. TRUE

(b) To select a model's hyperparameters. TRUE

(c) To determine the generalization of a machine learning model. (a & c are Same)

(d) To train multiple machine learning models on different datasets.



https://courses.cs.washington.edu/courses/cse446/23au/exams/pastexams/23sp-midterm-solutions.pdf

14. What is cross-validation **not** used for?
    (a) To evaluate the performance of a machine learning model on unseen data.
    (b) To select a model's hyperparameters.
    (c) To determine the generalization of a machine learning model.
    (d) To train multiple machine learning models on different datasets.

Correct answers: (d)

Explanation: The answer "to train multiple ML models on different datasets" is the correct one. We could argue that CV trains the same machine learning model on different partitions of the same dataset, but **not** *multiple* ML models on *different* datasets

Question: 11

Ella wants to use gradient descent to determine the optimal parameter values for her regularized model. To do so, she needs to select a value for α, the gradient descent learning rate.

She performs **4-fold cross-validation** using a dataset of 60 observations to help her choose from three possible values of α. The validation errors computed in each fold of her cross-validation (CV) procedure are shown below. Assume that she is not considering a test set, so all 60 observations are used in the cross-validation procedure.

	Fold #1	Fold #2	Fold #3	Fold #4
$\alpha = 0.1$	2	10	3	5
$\alpha = 1$	6	1	2	3
$\alpha = 10$	4	9	1	2

(e) [2 Pts] How many observations are in each validation fold of Ella's CV procedure?

Number of observations: _____

(f) [3 Pts] Given the validation errors displayed above, which value of α should Ella choose?

- ◯ A. $\alpha = 0.1$
- ◯ B. $\alpha = 1$
- ◯ C. $\alpha = 10$
- ◯ D. α should be determined using ordinary least squares, not cross-validation

https://ds100.org/sp24/resources/assets/exams/su23/su23_final_sol.pdf

Question: 11

Ella wants to use gradient descent to determine the optimal parameter values for her regularized model. To do so, she needs to select a value for α, the gradient descent learning rate.

She performs **4-fold cross-validation** using a dataset of 60 observations to help her choose from three possible values of α. The validation errors computed in each fold of her cross-validation (CV) procedure are shown below. Assume that she is not considering a test set, so all 60 observations are used in the cross-validation procedure.

	Fold #1	Fold #2	Fold #3	Fold #4
$\alpha = 0.1$	2	10	3	5
$\alpha = 1$	6	1	2	3
$\alpha = 10$	4	9	1	2

Avg
20/4
12/4
16/4
5 3 4

(e) [2 Pts] How many observations are in each validation fold of Ella's CV procedure?

Number of observations: 15

(f) [3 Pts] Given the validation errors displayed above, which value of α should Ella choose?

- A. $\alpha = 0.1$
- B. $\alpha = 1$
- C. $\alpha = 10$
- D. α should be determined using ordinary least squares, not cross-validation

https://ds100.org/sp24/resources/assets/exams/su23/su23_final_sol.pdf

(e) [2 Pts] How many observations are in each validation fold of Ella's CV procedure?

Solution: In 4-fold cross-validation, each validation fold contains $1/4$ of the training data. This means that $60/4 = 15$ observations are in each validation fold.

(f) [3 Pts] Given the validation errors displayed above, which value of α should Ella choose?

- ○ A. $\alpha = 0.1$
- ○ **B.** $\alpha = 1$
- ○ C. $\alpha = 10$
- ○ D. α should be determined using ordinary least squares, not cross-validation

Solution: Ella should select the value of α that results in the lowest cross-validation error, defined as the mean error across all validation folds. To determine the cross-validation error for each α value, find the row-wise average of the four validation errors for that choice of α.

	Fold #1	Fold #2	Fold #3	Fold #4	CV Error
$\alpha = 0.1$	2	10	3	5	5
$\alpha = 1$	6	1	2	3	3
$\alpha = 10$	4	9	1	2	4

The best choice of hyperparameter is $\alpha = 1$.

Question: 12

(d) [2 Pts] Yash decides to use LASSO regression with the design matrix from part 3c(ii). He has 4 candidate values for the regularization parameter λ, and decides to use 5-fold cross-validation to find the best λ. How many validation errors will he need to calculate?

Number of Validation Errors = [ ]

data

x = 4 possible values
5-fold cross validation
→ for each λ, we need to train 5 times

(d) [2 Pts] Yash decides to use LASSO regression with the design matrix from part 3c(ii). He has 4 candidate values for the regularization parameter λ, and decides to use 5-fold cross-validation to find the best λ. How many validation errors will he need to calculate?

Number of Validation Errors = [ ]

Solution: 5 folds * 4 values of λ = 20 validation errors

Question: 13

The model has a very high validation error.

- ○ A. We should increase λ
- ○ B. We should decrease λ
- ○ C. Not enough information

→ Determine if model is complex or model is simple.

Question: 13

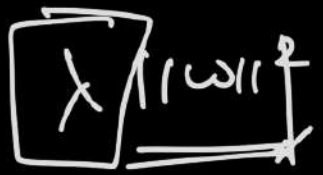
The model has a very high validation error.

- ○ A. We should increase λ
- ○ B. We should decrease λ
- ◉ C. Not enough information

→ Determine if model is complex or model is simple.

we can not determine if model is complex or simple.

Question: 13



The model has a very high validation error.

- ○ A. We should increase λ
- ○ B. We should decrease λ
- ○ C. Not enough information

→ Determine if model is complex or model is simple.

↳ training error is low. ⇒ Complex model ($\lambda \downarrow$ w ↑)

↳ training error is high ⇒ Simple model (to make it less $\lambda \uparrow$)

(ii) The model has a very high validation error.

- ○ A. We should increase λ
- ○ B. We should decrease λ
- ○ **C. Not enough information**

Solution: A high validation error could either mean our model is overfitting and too complex or underfitting and not complex enough. Based on this information alone, we cannot determine which of these two situations is occurring and, therefore, do not know how to adjust λ.

Question: 14

What are the advantages of k-fold cross validation relative to

(a) Validation set approach (Hold-out validation)

it does not give better indicator of generalised error

(b) Leave-one-out cross validation (LOOCV)

very expensive

What are the advantages of k-fold cross validation relative to

(a) Validation set approach

Solution. The estimate of the test error rate can be highly variable depending on which observations are included in the training and validation sets. □

(b) Leave-one-out cross validation (LOOCV)

Solution. LOOCV is a special case of k-fold cross-validation with $k = n$. Thus, LOOCV is the most computationally intense method since the model must be fit n times. □

Question: 15

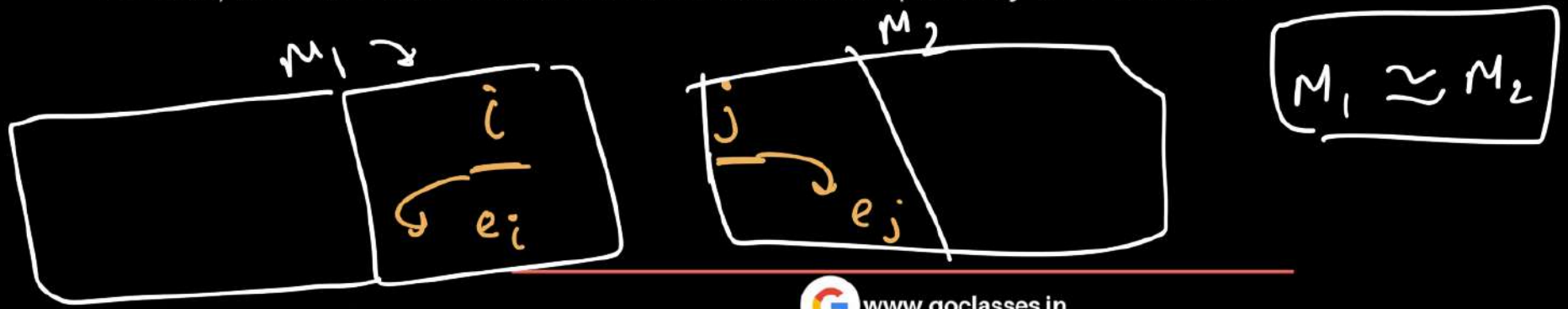
→ LOOCV

In n-fold cross-validation, each data point belongs to exactly one test fold, which makes the test folds independent. However, if the data in test folds i and j are independent, are the error estimates e_i and e_j on test folds i and j also independent?

A. True, because each test fold contains different data points.

B. False, because data points can appear in multiple training sets, making error estimates dependent.

C. True, because error estimates are calculated separately for each fold.



Answer: B

In n-fold cross-validation, each data point belongs to exactly one test fold, which makes the test folds independent. However, if the data in test folds i and j are independent, are the error estimates e_i and e_j on test folds i and j also independent?

A. True, because each test fold contains different data points.

B. False, because data points can appear in multiple training sets, making error estimates dependent.

C. True, because error estimates are calculated separately for each fold.

★ **SOLUTION:** False. Since a data point appears in multiple folds the training sets are dependent and thus test fold error estimates are dependent.

MSQ Question: 16

(c) [3 pts] You train a linear classifier on 10,000 training points and discover that the training accuracy is only 67%. Which of the following, done in isolation, has a good chance of improving your training accuracy?

- ○ Add novel features
- ○ Use linear regression
- ○ Train on more data
- ○ Train on less data

MSQ Question: 16

(c) [3 pts] You train a linear classifier on 10,000 training points and discover that the training accuracy is only 67%. Which of the following, done in isolation, has a good chance of improving your training accuracy?

- ○ Add novel features
- ○ Use linear regression
- ○ Train on more data
- ○ Train on less data

(c) [3 pts] You train a linear classifier on 10,000 training points and discover that the training accuracy is only 67%. Which of the following, done in isolation, has a good chance of improving your training accuracy?

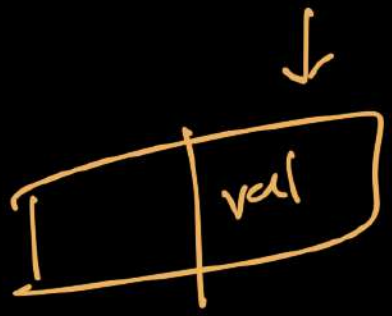
- ● Add novel features
- ○ Use linear regression
- ○ Train on more data
- ● Train on less data

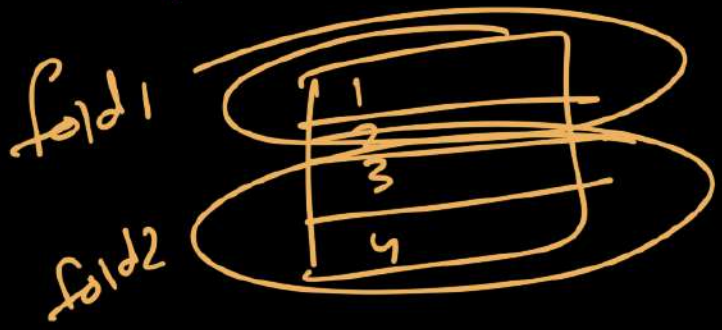
Question: 17

(g) [1 Pt] **Picking Hyperparameters**: Connie is now using cross validation to pick a hyperparameter for her regularization term above. She provides you with the errors she has computed, along with the fitted parameters for each fold and the training data points. Assume that the training and validation sets for each fold are consistent across all three choices of λ. What value of λ should Connie choose?

Fold Num	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
1	1	0.5	9
2	3	1.5	1

Fold Num	Training Data	θ_1	θ_2
1	Rows 1 and 2	-1	1
2	Rows 3 and 4	0	2





(g) [1 Pt] **Picking Hyperparameters**: Connie is now using cross validation to pick a hyperparameter for her regularization term above. She provides you with the errors she has computed, along with the fitted parameters for each fold and the training data points. Assume that the training and validation sets for each fold are consistent across all three choices of λ. What value of λ should Connie choose?

Fold Num	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
1	1	0.5	9
2	3	1.5	1

Fold Num	Training Data	θ_1	θ_2
1	Rows 1 and 2	-1	1
2	Rows 3 and 4	0	2

Your Answer:

Solution: $\lambda = 2$. Compute the mean error across the folds for each λ.

For $\lambda = 1$: $\frac{1+3}{2} = 2$
For $\lambda = 2$: $\frac{0.5+1.5}{2} = 1$
For $\lambda = 3$: $\frac{9+1}{2} = 5$

Since $\lambda = 2$ has the lowest mean cross validation error, Connie should choose $\lambda = 2$.

Question: 18

4 Points

Now you will run 10-fold cross-validation for training k-NN. For each of candidate k values, you train k-NN on all dataset but one of the 10 folds, then measuring an approximate validation error on the examples in that heldout fold.

When you have 5 different candidate k values to decide the best model, you will train k-NN total N1 times. The performance of individual models (with a specific k value) will be evaluated by the mean of N2 validation errors.

- ○ N1=10, N2=5
- ○ N1=5, N2=10
- ○ N1=45, N2=10
- ○ N2=45, N1=50
- ○ N1=50, N2=10
- ○ N1=10, N2=50

10 fold

hyper parameter: K

5

Total N_1 times train

for fixk :

avg over N_2 errors

Question: 18

4 Points

Now you will run 10-fold cross-validation for training k-NN. For each of candidate k values, you train k-NN on all dataset but one of the 10 folds, then measuring an approximate validation error on the examples in that heldout fold.

When you have 5 different candidate k values to decide the best model, you will train k-NN total N1 times. The performance of individual models (with a specific k value) will be evaluated by the mean of N2 validation errors.

- ○ N1=10, N2=5
- ○ N1=5, N2=10
- ○ N1=45, N2=10
- ○ N2=45, N1=50
- ✓ ○ N1=50, N2=10
- ○ N1=10, N2=50

10 fold

hyper parameter: K

5

Total N_1 times train

for fixk :

avg over N_2 errors

Question: 19

Regardless of your answer to the previous question, suppose Jordan uses k-fold cross-validation with α chosen from $0.1, 0.2, 0.4$ and $\mathcal{B}$ chosen from $32, 64, 128$. The average cross-validated loss is shown in the below table for each combination of α and $\mathcal{B}$.

$\mathcal{B}$	α		
	0.1	0.2	0.4
32	0.0022	0.7031	0.0370
64	0.0051	0.9075	0.0471
128	0.0018	0.6007	0.0157

Which of the following is the **most optimal pair** of α and $\mathcal{B}$? Fill in the blanks.

$\alpha =$ _____ and $\mathcal{B} =$ _____

Regardless of your answer to the previous question, suppose Jordan uses k-fold cross-validation with α chosen from $0.1, 0.2, 0.4$ and $\mathcal{B}$ chosen from $32, 64, 128$. The average cross-validated loss is shown in the below table for each combination of α and $\mathcal{B}$.

$\mathcal{B}$	α		
	0.1	0.2	0.4
32	0.0022	0.7031	0.0370
64	0.0051	0.9075	0.0471
128	0.0018	0.6007	0.0157

Which of the following is the **most optimal pair** of α and $|\mathcal{B}|$? Fill in the blanks.

$\alpha =$ _____ and $|\mathcal{B}| =$ _____

Solution: The most optimal pair to minimize the average CV loss is $\alpha = 0.1$ and $\mathcal{B} = 128$.

Question: 20

(2.0 pt) Let's say we pick three hyperparameters to tune with cross-validation. We have 5 candidate values for hyperparameter 1, 6 candidate values for hyperparameter 2, and 7 candidate values for hyperparameter 3. We perform 4-fold cross validation to find the optimal combination of hyperparameters, across all possible combinations.

In this cross-validation process, how many **random forests** will we train? Your answer can be left as a product of multiple integers, e.g. "1 * 2 * 3", or simplified to a single integer, e.g. "6". (These are not the correct answers to the problem).

4-fold

$\lambda \Rightarrow 5$

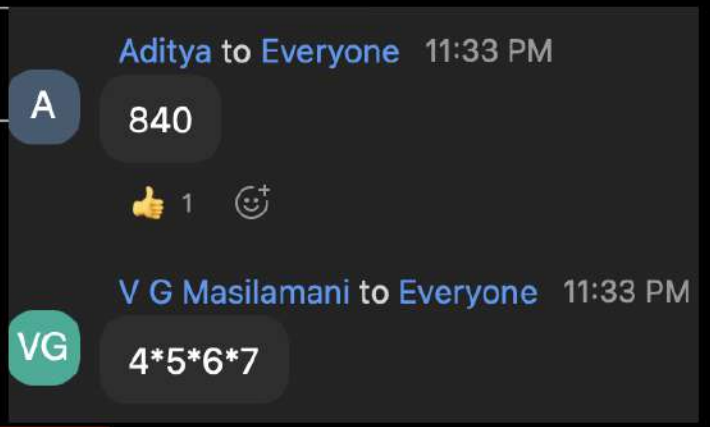
$\alpha \Rightarrow 6$

$\beta \Rightarrow 7$

(2.0 pt) Let's say we pick three hyperparameters to tune with cross-validation. We have 5 candidate values for hyperparameter 1, 6 candidate values for hyperparameter 2, and 7 candidate values for hyperparameter 3. We perform 4-fold cross validation to find the optimal combination of hyperparameters, across all possible combinations.

In this cross-validation process, how many **random forests** will we train? Your answer can be left as a product of multiple integers, e.g. "1 * 2 * 3", or simplified to a single integer, e.g. "6". (These are not the correct answers to the problem).

$4 * 5 * 6 * 7 = 840$



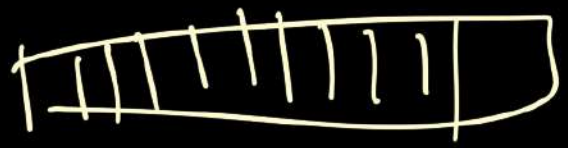
Question: 21

[2 Pts] Suppose we have m data points in our training set and n data points in our test set. In *leave-one-out* cross validation, we only use one data point for validation while the rest are used for training. Which of the following is *leave-one-out* cross validation equivalent to?

- ◯ A. m-fold cross validation
- ◯ B. n-fold cross validation
- ◯ C. (m + n)-fold cross validation
- ◯ D. 1-fold cross validation

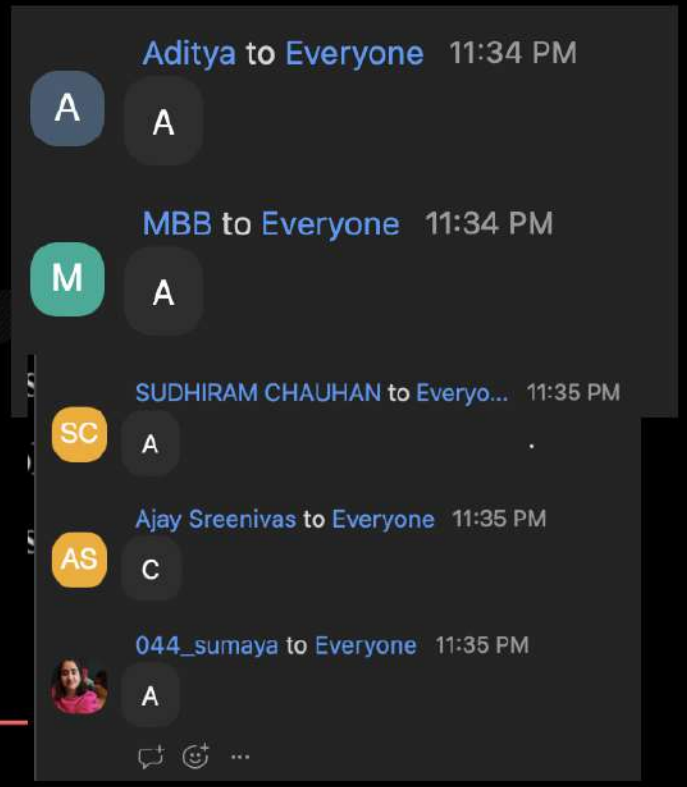
m : train

n : test



17. [2 Pts] Suppose we have m data points in our training set and n data points in our test set. In *leave-one-out* cross validation, we only use one data point for validation while the rest are used for training. Which of the following is *leave-one-out* cross validation equivalent to?

- **A. m-fold cross validation**
- B. n-fold cross validation
- C. (m + n)-fold cross validation
- D. 1-fold cross validation



https://ds100.org/fa20/resources/assets/exams/fa18/fa18finalsol.pdf

Question: 22

No. of examples = 7

v. (1 point) Suppose we use leave-one-out cross validation meaning we use 7-fold cross validation with a split of 6 to 1 between training set and validating set. Compute the average classification error over the 7-folds.

train Hard margin SVM

Figure 5: Training Data

https://nyu-ds1003.github.io/mlcourse/2021/exams/sp20/midterm_solutions.pdf

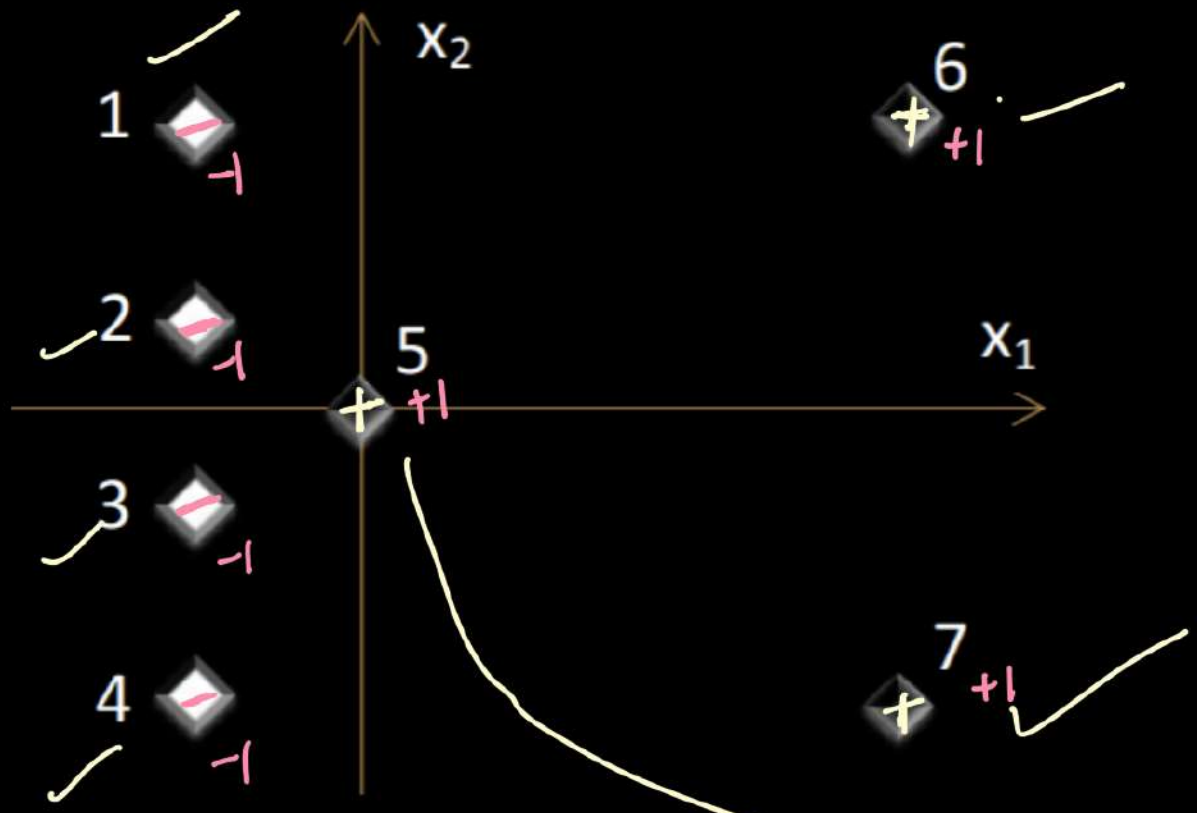


Figure 5: Training Data

will be misclassified in LOOCV

v. (1 point) Suppose we use leave-one-out cross validation meaning we use 7-fold cross validation with a split of 6 to 1 between training set and validating set. Compute the average classification error over the 7-folds.

Solution: 1/7

https://nyu-ds1003.github.io/mlcourse/2021/exams/sp20/midterm_solutions.pdf

Question: 23 True/False

(19) True/False: k-fold cross-validation with $k = 100$ is computationally more expensive (slower) than "leave-one-out" cross validation. (Assume that there are enough data points to divide the dataset evenly by k.)

(A) True

(B) False

https://courses.cs.washington.edu/courses/cse446/23au/exams/pastexams/22au-midterm-solutions.pdf

(19) True/False: k-fold cross-validation with $k = 100$ is computationally more expensive (slower) than "leave-one-out" cross validation. (Assume that there are enough data points to divide the dataset evenly by k.)

(A) True

(B) False

Solution:

The solution is (B)

Question: 24

(20) Assume we have a data matrix X. Which of the following is a true statement when comparing leave-one-out cross validation (LOOCV) error with the true error?

(A) LOOCV error is typically a slight underestimation of the true error of a model trained on X.

(B) LOOCV error is typically a slight overestimation of the true error of a model trained on X.

(C) LOOCV error is an unbiased estimator of the true error of a model trained on X.

https://courses.cs.washington.edu/courses/cse446/23au/exams/pastexams/22au-midterm-solutions.pdf

(20) Assume we have a data matrix X. Which of the following is a true statement when comparing leave-one-out cross validation (LOOCV) error with the true error?

(A) LOOCV error is typically a slight underestimation of the true error of a model trained on X.

(B) LOOCV error is typically a slight overestimation of the true error of a model trained on X.

(C) LOOCV error is an unbiased estimator of the true error of a model trained on X.

Solution:

The solution is (B)

Question: 25

9. Leave-one-out cross-validation (LOOCV) is a special case of k-fold cross-validation where:

(a) The training set contains all but one sample, and the remaining sample is used for testing.

(b) The training set contains only one sample, and the remaining sample is used for testing.

(c) The training set contains exactly one sample from each class, and the remaining samples are used for testing.

(d) The training set contains one sample from each fold, and the remaining folds are used for testing.

9. Leave-one-out cross-validation (LOOCV) is a special case of k-fold cross-validation where:

(a) The training set contains all but one sample, and the remaining sample is used for testing.

(b) The training set contains only one sample, and the remaining sample is used for testing.

(c) The training set contains exactly one sample from each class, and the remaining samples are used for testing.

(d) The training set contains one sample from each fold, and the remaining folds are used for testing.

Correct answers: (a)

Question: 26

(a) (4.0 points)

Recall that a random forest is created from a number of decision trees, with each decision tree created from a bootstrapped version of the original training set. One hyperparameter of a random forest is the number of decision trees we train to create the random forest.

Define T to be the number of decision trees used to create the random forest. Let's say we have two candidate values for T: $var1$ and $var2$. We want to perform $var3$ - fold cross-validation to determine the optimal value of T. Assume $var1$, $var2$, and $var3$ are integers.

i. (2.0 pt) In this cross-validation process, how many **random forests** will we train? Your answer should be in terms of $var1$, $var2$, and/or $var3$ and should be an integer.

ii. (2.0 pt) In this cross-validation process, how many **decision trees** will we train? Your answer should be in terms of $var1$, $var2$, and/or $var3$ and should be an integer.

(a) (4.0 points)

Recall that a random forest is created from a number of decision trees, with each decision tree created from a bootstrapped version of the original training set. One hyperparameter of a random forest is the number of decision trees we train to create the random forest.

Define T to be the number of decision trees used to create the random forest. Let's say we have two candidate values for T: $var1$ and $var2$. We want to perform $var3$ - fold cross-validation to determine the optimal value of T. Assume $var1$, $var2$, and $var3$ are integers.

i. **(2.0 pt)** In this cross-validation process, how many **random forests** will we train? Your answer should be in terms of $var1$, $var2$, and/or $var3$ and should be an integer.

2 * var3

ii. **(2.0 pt)** In this cross-validation process, how many **decision trees** will we train? Your answer should be in terms of $var1$, $var2$, and/or $var3$ and should be an integer.

(var1 + var2) * var3

Question: 27

5. Suppose we have a design matrix $\mathbb{X}$ comprising of n observations, d features, and an additional intercept term. We decide to use $\mathbb{X}$ to create, tune, and evaluate a regularized linear regression model with two regularization hyperparameters, λ_1 and λ_2. We have i different choices for λ_1 and j different choices for λ_2, so there are $i \cdot j$ possible combinations of λ_1 and λ_2.

   We set aside 20% of our data to use as a test set and perform k-fold cross-validation to tune our hyperparameters. We compute the cross-validation error of each of the $i \cdot j$ possible combinations of λ_1 and λ_2 values, and our goal is to find the combination of values with the lowest cross-validation error. All following answers should be expressed in terms of i, j, n, d, k, and/or constants only (except for part c).

   (a) [2 Pts] For a single combination of hyperparameter values, how many model parameters do we fit?

   (b) [2 Pts] How many observations are in the validation set of each fold?

   (c) [2 Pts] How many model parameters do we calculate in total? Your answer can be in terms of A, your answer to part a.

5. Suppose we have a design matrix $\mathbb{X}$ comprising of n observations, d features, and an additional intercept term. We decide to use $\mathbb{X}$ to create, tune, and evaluate a regularized linear regression model with two regularization hyperparameters, λ_1 and λ_2. We have i different choices for λ_1 and j different choices for λ_2, so there are $i \cdot j$ possible combinations of λ_1 and λ_2.

   We set aside 20% of our data to use as a test set and perform k-fold cross-validation to tune our hyperparameters. We compute the cross-validation error of each of the $i \cdot j$ possible combinations of λ_1 and λ_2 values, and our goal is to find the combination of values with the lowest cross-validation error. All following answers should be expressed in terms of i, j, n, d, k, and/or constants only (except for part c).

   (a) [2 Pts] For a single combination of hyperparameter values, how many model parameters do we fit?

   Solution: $k \cdot (d+1)$

   (b) [2 Pts] How many observations are in the validation set of each fold?

   Solution: $\frac{0.8n}{k}$

   (c) [2 Pts] How many model parameters do we calculate in total? Your answer can be in terms of A, your answer to part a.

   Solution: $i \cdot j \cdot k \cdot (d+1)$

Question:

35. **Extra credit:** Consider a nearest neighbor classifier that chooses the label for a test point to be the label of its nearest neighboring training example. What is its leave-one-out cross-validated error for the data in the following figure? ("+" and "−" indicate labels of the points).

Answer: _____

35. **Extra credit:** Consider a nearest neighbor classifier that chooses the label for a test point to be the label of its nearest neighboring training example. What is its leave-one-out cross-validated error for the data in the following figure? ("+" and "−" indicate labels of the points).

Answer: _____

Explanation: The solution is 2/5

Question:

15. [2.5 Pts] Aman and Ed built a model on their data with two regularization hyperparameters λ and γ. They have 4 good candidate values for λ and 3 possible values for γ, and they are wondering which λ, γ pair will be the best choice. If they were to perform five-fold cross-validation, how many validation errors would they need to calculate?

https://ds100.org/fa20/resources/assets/exams/fa18/fa18finalsol.pdf

15. [2.5 Pts] Aman and Ed built a model on their data with two regularization hyperparameters λ and γ. They have 4 good candidate values for λ and 3 possible values for γ, and they are wondering which λ, γ pair will be the best choice. If they were to perform five-fold cross-validation, how many validation errors would they need to calculate?

Solution: 60

Question:

13. Why is it important to use a different test set to evaluate the final performance of the model, rather than the validation set used during model selection?

(a) The model may have overfit to the validation set

(b) The test set is a better representation of new, unseen data

(c) Both a and b

(d) None of the above

https://courses.cs.washington.edu/courses/cse446/23au/exams/pastexams/23sp-midterm-solutions.pdf

13. Why is it important to use a different test set to evaluate the final performance of the model, rather than the validation set used during model selection?

(a) The model may have overfit to the validation set

(b) The test set is a better representation of new, unseen data

(c) Both a and b

(d) None of the above

Correct answers: (c)

MSQ Question:

(g) [4 pts] You are training a soft-margin SVM on a binary classification problem. You find that your model's training accuracy is very high, while your validation accuracy is very low. Which of the following are likely to improve your model's performance on the validation data?

- ○ A: Training your model on more data.
- ○ B: Adding a quadratic feature to each sample point.
- ○ C: Increasing the hyperparameter C.
- ○ D: Decreasing the hyperparameter C.

https://people.eecs.berkeley.edu/~jrs/189/exam/mids22.pdf

(g) [4 pts] You are training a soft-margin SVM on a binary classification problem. You find that your model's training accuracy is very high, while your validation accuracy is very low. Which of the following are likely to improve your model's performance on the validation data?

- ● A: Training your model on more data.
- ○ B: Adding a quadratic feature to each sample point.
- ○ C: Increasing the hyperparameter C.
- ● D: Decreasing the hyperparameter C.

A is true as training on more data reduces variance and decreases overfitting in general. B is false since polynomial features increase the variance and the risk of overfitting. C is false because increasing C enforces a harder margin constraint, leading to more overfitting. D is true because decreasing C allows for more slack, decreasing the risk of overfitting.

MSQ Question:

Which of the following statement(s) is(are) true for k-NN?

- ☐ k represents the number of classes.
- ☐ The final outcome of the algorithm may change with the distance measure.
- ☐ CrossValidation can be used to find the optimal k.
- ☐ As k increases, we overfit the data eventually.

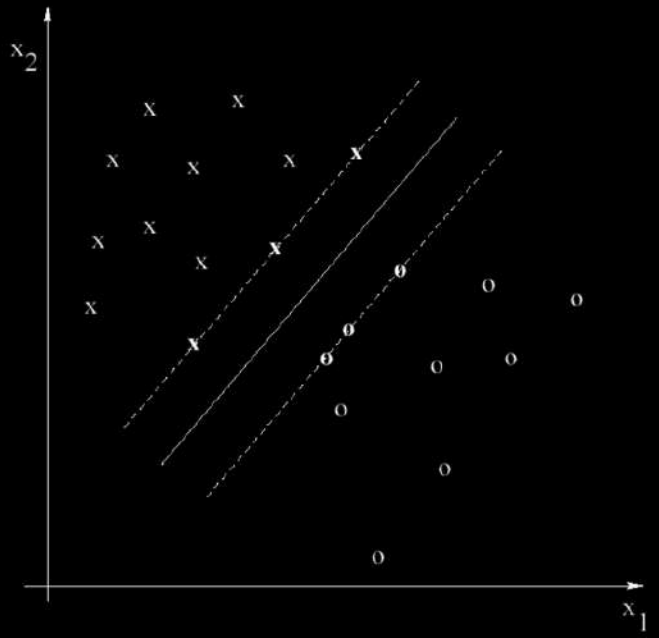
https://moodlearchive.epfl.ch/2022-2023/pluginfile.php/3076586/mod_resource/content/1/Mock_Exam_2022_Spring_Solutions.pdf

Which of the following statement(s) is(are) true for k-NN?

- ☐ k represents the number of classes.
- ☒ The final outcome of the algorithm may change with the distance measure.
- ☒ CrossValidation can be used to find the optimal k.
- ☐ As k increases, we overfit the data eventually.

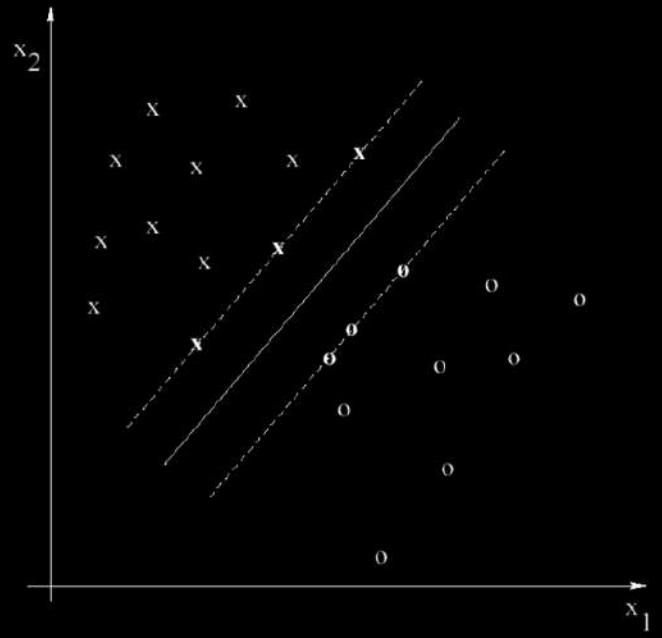
Question:

What is the leave-one-out cross-validation error estimate for maximum margin separation in the following figure ? (we are asking for a number)



http://www.ai.mit.edu/courses/6.867-f03/exams/midterm2001sols.pdf

What is the leave-one-out cross-validation error estimate for maximum margin separation in the following figure ? (we are asking for a number)



Answer: 0

Based on the figure we can see that removing any single point would not chance the resulting maximum margin separator. Since all the points are initially classified correctly, the leave-one-out error is zero.

Question:

[2 points] Suppose you have picked the parameter θ for a model using 10-fold cross validation. The best way to pick a final model to use and estimate its error is to

(a) pick any of the 10 models you built for your model; use its error estimate on the held-out data

(b) pick any of the 10 models you built for your model; use the average CV error for the 10 models as its error estimate

(c) average all of the 10 models you got; use the average CV error as its error estimate

(d) average all of the 10 models you got; use the error the combined model gives on the full training set

(e) train a new model on the full data set, using the θ you found; use the average CV error as its error estimate

[2 points] Suppose you have picked the parameter θ for a model using 10-fold cross validation. The best way to pick a final model to use and estimate its error is to

(a) pick any of the 10 models you built for your model; use its error estimate on the held-out data

(b) pick any of the 10 models you built for your model; use the average CV error for the 10 models as its error estimate

(c) average all of the 10 models you got; use the average CV error as its error estimate

(d) average all of the 10 models you got; use the error the combined model gives on the full training set

(e) train a new model on the full data set, using the θ you found; use the average CV error as its error estimate

★ **SOLUTION:** E

Question:

[3] Under which of the following conditions is ***k*-fold cross-validation** the *same as* **leave-one-out cross-validation**?

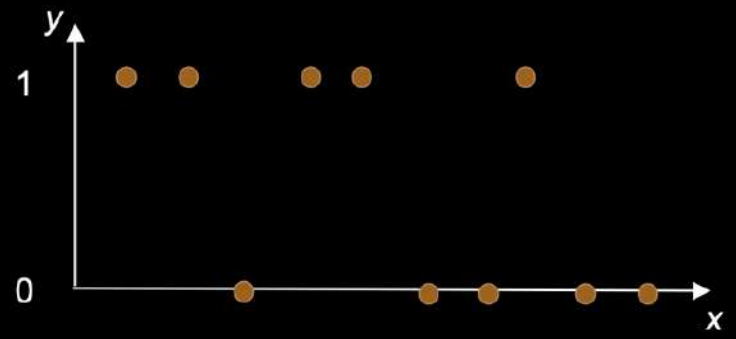
A. The training set and test set have the *same* number of examples
B. The training set and tuning set have the *same* number of examples
C. $k = 1$
D. $k = n$, where n is the total number of examples
E. None of the above

[3] Under which of the following conditions is ***k*-fold cross-validation** the *same as* **leave-one-out cross-validation**?

A. The training set and test set have the *same* number of examples
B. The training set and tuning set have the *same* number of examples
C. $k = 1$
D. $k = n$, where n is the total number of examples
E. None of the above

Question:

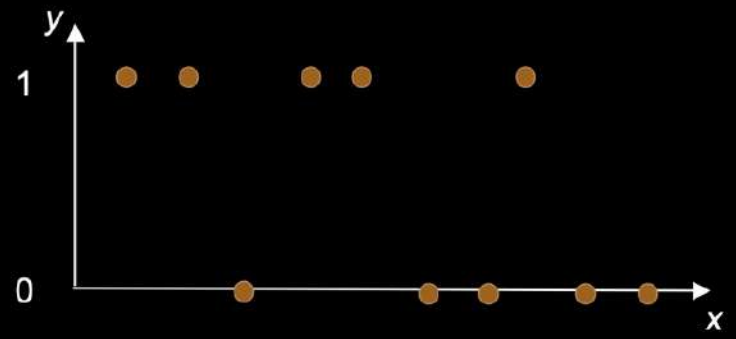
Suppose you are using a Majority Classifier on the following training set containing 10 examples where each example has one real-valued feature, *x*, and a binary class label, y, with value 0 or 1. Define this Majority Classifier to predict the class label that is in the *majority in the training set*, regardless of the input value. In case of ties, predict class 1.



(a) [3] What is the *training set accuracy*?

(i) 0%
(ii) 10%
(iii) 50%
(iv) 90%
(v) 100%

Suppose you are using a Majority Classifier on the following training set containing 10 examples where each example has one real-valued feature, *x*, and a binary class label, y, with value 0 or 1. Define this Majority Classifier to predict the class label that is in the *majority in the training set*, regardless of the input value. In case of ties, predict class 1.



(a) [3] What is the *training set accuracy*?

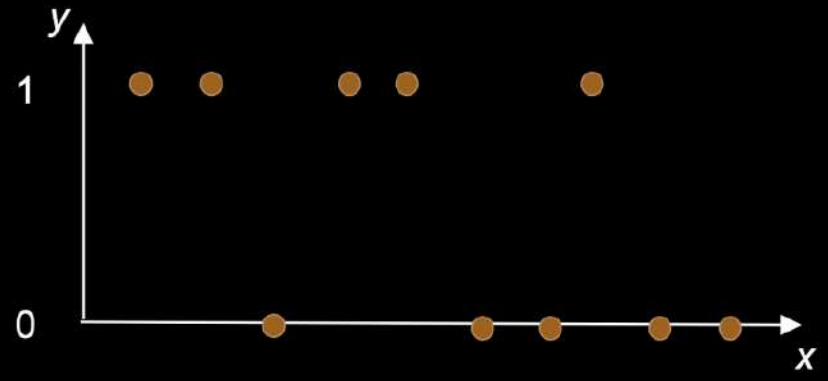
(i) 0%
(ii) 10%
(iii) 50%
(iv) 90%
(v) 100%

(iii) 5/10 = 50%

Question:

(b) [3] What is the *Leave-1-Out Cross-Validation accuracy*?

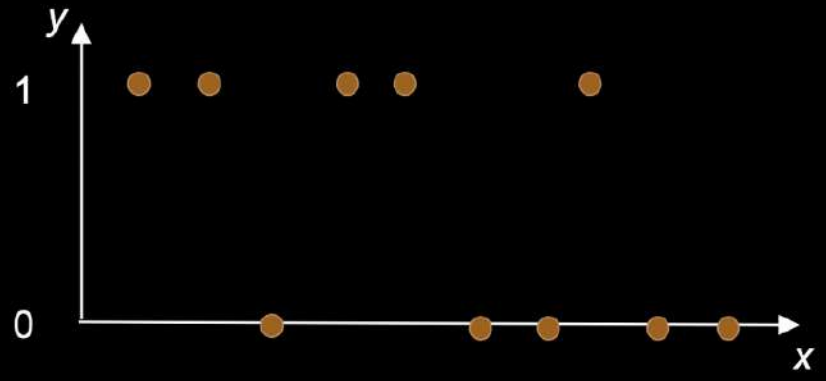
(i) 0%
(ii) 10%
(iii) 50%
(iv) 90%
(v) 100%



https://pages.cs.wisc.edu/~dyer/cs540/exams/exam1_f18-sol.pdf

(b) [3] What is the *Leave-1-Out Cross-Validation accuracy*?

(i) 0%
(ii) 10%
(iii) 50%
(iv) 90%
(v) 100%

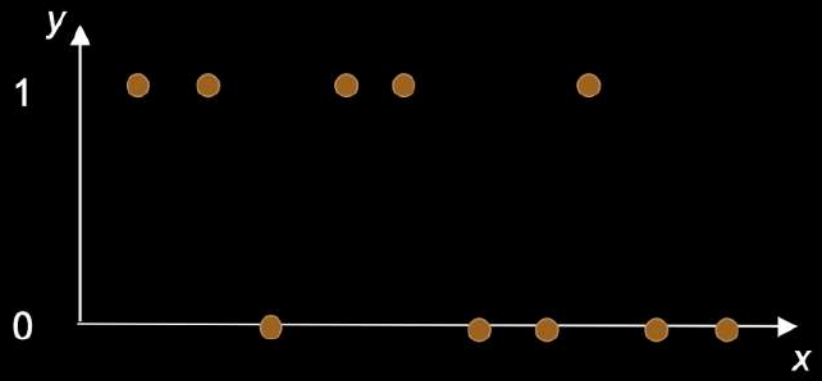


(i) 0% because each of the 10 test examples is classified incorrectly because the majority class of the other 9 is in the opposite class, so the average accuracy is 0%

Question:

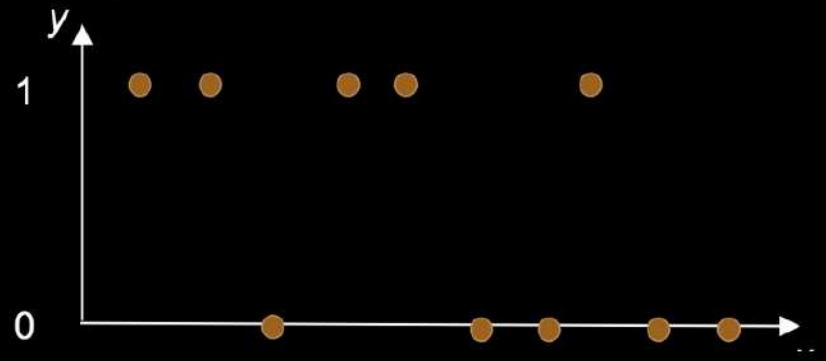
(c) [3] What is the *2-fold Cross-Validation accuracy*? Assume the leftmost 5 points (i.e., the 5 points with smallest x values) are in one fold and the rightmost 5 points are in the second fold.

(i) 0%
(ii) 20%
(iii) 50%
(iv) 80%
(v) 100%



(c) [3] What is the *2-fold Cross-Validation accuracy*? Assume the leftmost 5 points (i.e., the 5 points with smallest x values) are in one fold and the rightmost 5 points are in the second fold.

(i) 0%
(ii) 20%
(iii) 50%
(iv) 80%
(v) 100%



(ii) 20% because for each fold, only 1 of the 5 test examples is classified correctly, so the average accuracy on the two folds is 20%

Question:

Suppose we have a training dataset of 90 points, and a test set of 30 points, and want to know which λ value is best for a ridge regression model. Our candidate hyperparameters are $\lambda = 0.1$, $\lambda = 1$, and $\lambda = 10$.

Suppose we select the best choice of λ from the three choices available using **3-fold** cross validation. As mentioned in class, we can compute the optimal parameters for a ridge regression model with the expression $\vec{\beta} = (\mathbb{X}^T\mathbb{X} + n\lambda I)^{-1}\mathbb{X}^T\vec{y}$. Assume that we use this closed equation to fit the parameters for our model.

i. [2 Pts] During the entire process of selecting our best λ, how many total times will we evaluate the expression $(\mathbb{X}^T\mathbb{X} + n\lambda I)^{-1}\mathbb{X}^T\vec{y}$?

◯ 1 ◯ 2 ◯ 3 ◯ 6 ◯ 9 ◯ 30 ◯ 60 ◯ 90 ◯ 270

ii. [2 Pts] How many rows will be in $\mathbb{X}$ each time this expression is evaluated?

◯ 1 ◯ 2 ◯ 3 ◯ 6 ◯ 9 ◯ 30 ◯ 60 ◯ 90 ◯ 120
◯ It will vary each time. ◯ Not enough information.

Suppose we select the best choice of λ from the three choices available using **3-fold** cross validation. As mentioned in class, we can compute the optimal parameters for a ridge regression model with the expression $\vec{\beta} = (\mathbb{X}^T\mathbb{X} + n\lambda I)^{-1}\mathbb{X}^T\vec{y}$. Assume that we use this closed equation to fit the parameters for our model.

i. [2 Pts] During the entire process of selecting our best λ, how many total times will we evaluate the expression $(\mathbb{X}^T\mathbb{X} + n\lambda I)^{-1}\mathbb{X}^T\vec{y}$?

◯ 1 ◯ 2 ◯ 3 ◯ 6 ◯ **9** ◯ 30 ◯ 60 ◯ 90 ◯ 270

ii. [2 Pts] How many rows will be in $\mathbb{X}$ each time this expression is evaluated?

◯ 1 ◯ 2 ◯ 3 ◯ 6 ◯ 9 ◯ 30 ◯ **60** ◯ 90 ◯ 120
◯ It will vary each time. ◯ Not enough information.

Question:

Suppose we have a training dataset of 90 points, and a test set of 30 points, and want to know which λ value is best for a ridge regression model. Our candidate hyperparameters are $\lambda = 0.1$, $\lambda = 1$, and $\lambda = 10$.

As in the previous part, suppose we want to select the best λ from the three choices above using **3-fold** cross validation. To evaluate the MSE for a given $\vec{\beta}$, we use the sum of squares: $||\vec{y} - \mathbb{X}\vec{\beta}||_2^2$. Reminder that this expression is just another way of writing $\sum(\vec{y}_i - \vec{x}_i^T\vec{\beta})^2$.

i. [2 Pts] During the entire process of selecting our best λ, how many times will this expression get evaluated?

◯ 1 ◯ 2 ◯ 3 ◯ 6 ◯ 9 ◯ 30 ◯ 60 ◯ 90

ii. [2 Pts] How many rows will be in $\mathbb{X}$ each time this expression is evaluated?

◯ 1 ◯ 2 ◯ 3 ◯ 6 ◯ 9 ◯ 30 ◯ 60 ◯ 90 ◯ 120
◯ It will vary each time. ◯ Not enough information.

https://ds100.org/fa20/resources/assets/exams/fa19/fa19midterm2sol.pdf

Suppose we have a training dataset of 90 points, and a test set of 30 points, and want to know which λ value is best for a ridge regression model. Our candidate hyperparameters are $\lambda = 0.1$, $\lambda = 1$, and $\lambda = 10$.

As in the previous part, suppose we want to select the best λ from the three choices above using **3-fold** cross validation. To evaluate the MSE for a given $\vec{\beta}$, we use the sum of squares: $||\vec{y} - \mathbb{X}\vec{\beta}||_2^2$. Reminder that this expression is just another way of writing $\sum(\vec{y}_i - \vec{x}_i^T\vec{\beta})^2$.

i. [2 Pts] During the entire process of selecting our best λ, how many times will this expression get evaluated?

◯ 1 ◯ 2 ◯ 3 ◯ 6 ◯ **9** ◯ 30 ◯ 60 ◯ 90

ii. [2 Pts] How many rows will be in $\mathbb{X}$ each time this expression is evaluated?

◯ 1 ◯ 2 ◯ 3 ◯ 6 ◯ 9 ◯ **30** ◯ 60 ◯ 90 ◯ 120
◯ It will vary each time. ◯ Not enough information.

https://ds100.org/fa20/resources/assets/exams/fa19/fa19midterm2sol.pdf

Question: True/False

(a) **(1.0 pt)** The test set is divided into k folds. For each fold of the test set, we use the entire training set to train the model, and use the given fold/subset of the test set for validation. The average error among all k folds is the cross-validation error.

True or False: This modification will result in overfitting.

- ◯ True
- ◯ False

(a) **(1.0 pt)** The test set is divided into k folds. For each fold of the test set, we use the entire training set to train the model, and use the given fold/subset of the test set for validation. The average error among all k folds is the cross-validation error.

True or False: This modification will result in overfitting.

- ● True
- ○ False

We shouldn't be using the test set for validation purposes; that defeats the purpose of cross-validation.

https://ds100.org/fa20/resources/assets/exams/fa20/fa20finalsol.pdf

Question: True/False

(b) (1.0 pt) We use normal k-fold cross-validation, but for each fold we only use half of the validation set for validation.

True or False: This modification will result in overfitting.

- ◯ True
- ◯ False

https://ds100.org/fa20/resources/assets/exams/fa20/fa20finalsol.pdf

(b) (1.0 pt) We use normal k-fold cross-validation, but for each fold we only use half of the validation set for validation.

True or False: This modification will result in overfitting.

○ True

● False

This will not cause overfitting, but it is essentially throwing away data; we could be training our model on more data without overfitting.

https://ds100.org/fa20/resources/assets/exams/fa20/fa20finalsol.pdf

Question: True/False

(c) (1.0 pt) We use normal k-fold cross-validation, but for each fold we use the entire training set for training.

True or False: This modification will result in overfitting.

- ◯ True
- ◯ False

The purpose of training on $k-1$ folds and using the remaining fold for validation is to not train and validate our model on the same fold. By making the modification proposed in the question, we would be doing just that.

https://ds100.org/fa20/resources/assets/exams/fa20/fa20finalsol.pdf

(c) (1.0 pt) We use normal k-fold cross-validation, but for each fold we use the entire training set for training.

True or False: This modification will result in overfitting.

- ● True
- ○ False

The purpose of training on $k-1$ folds and using the remaining fold for validation is to not train and validate our model on the same fold. By making the modification proposed in the question, we would be doing just that.

https://ds100.org/fa20/resources/assets/exams/fa20/fa20finalsol.pdf

Question:

Consider the following dataset with one real-valued input x and one binary output y. We are going to use k-NN with unweighted Euclidean distance to predict y for x.

X	Y
-0.1	-
0.7	+
1.0	+
1.6	-
2.0	+
2.5	+
3.2	-
3.5	-
4.1	+
4.9	+

(a) What is the leave-one-out cross-validation error of 1-NN on this dataset? Give your answer as the number of misclassifications.

(b) What is the leave-one-out cross-validation error of 3-NN on this dataset? Give your answer as the number of misclassifications.

https://www.cs.cmu.edu/~epxing/Class/10701/exams/final2003-solution.pdf

Consider the following dataset with one real-valued input x and one binary output y. We are going to use k-NN with unweighted Euclidean distance to predict y for x.

X	Y
-0.1	-
0.7	+
1.0	+
1.6	-
2.0	+
2.5	+
3.2	-
3.5	-
4.1	+
4.9	+

(a) What is the leave-one-out cross-validation error of 1-NN on this dataset? Give your answer as the number of misclassifications.

4

(b) What is the leave-one-out cross-validation error of 3-NN on this dataset? Give your answer as the number of misclassifications.

8

MSQ Question:

Mark all the statements that are true, about cross-validation.

- ☐ Increasing k in k-fold cross validation decreases the bias of the model.
- ☐ We use cross validation because it is less computationally intensive than regular validation.
- ☐ We use cross validation over regular validation because cross validation allows us to use the entire training data set to test the model
- ☐ Cross-validation cannot be used in production, as the model is cheating by looking at validation samples.

Mark all the statements that are true, about cross-validation.

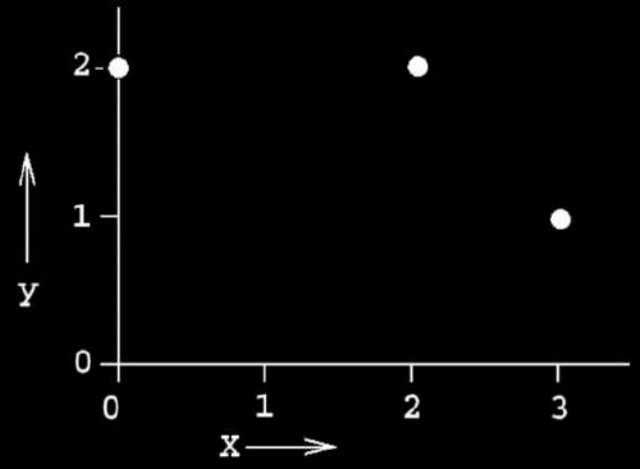
- ☐ **Increasing k in k-fold cross validation decreases the bias of the model.**
- ☐ We use cross validation because it is less computationally intensive than regular validation.
- ☐ **We use cross validation over regular validation because cross validation allows us to use the entire training data set to test the model**
- ☐ Cross-validation cannot be used in production, as the model is cheating by looking at validation samples.

Solution:

- True. Increasing k increases the size of the training set per fold and decreases the size of the validation set. Each model is trained on more data; therefore decreasing the bias of the model.
- False. Cross validation is more computationally intensive than regular validation.
- True.
- False.

Question:

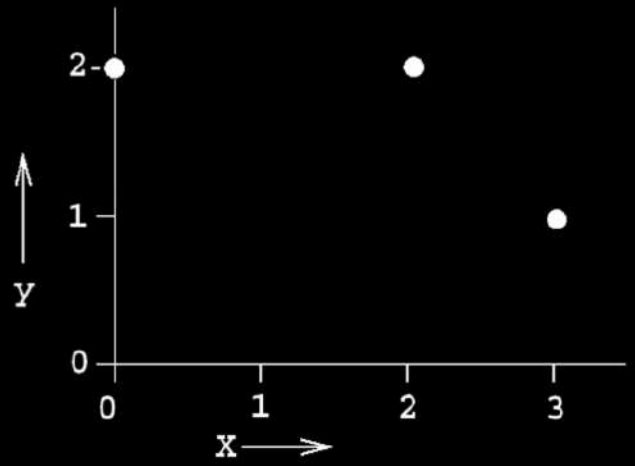
(c) [*5 pts*] Suppose you have this data set with one real-valued input and one real-valued output:



x	y
0	2
2	2
3	1

(c.1) What is the mean squared leave one out cross validation error of using linear regression ? (i.e. the mode is $y = \beta_0 + \beta_1 x$)

(c) [5 *pts*] Suppose you have this data set with one real-valued input and one real-valued output:



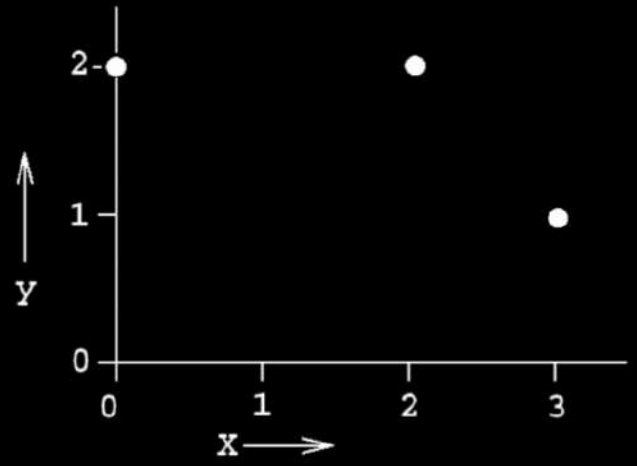
x	y
0	2
2	2
3	1

(c.1) What is the mean squared leave one out cross validation error of using linear regression ? (i.e. the mode is $y = \beta_0 + \beta_1 x$)

Answer: $\frac{2^2+(2/3)^2+1^2}{3} = 49/27$

Question:

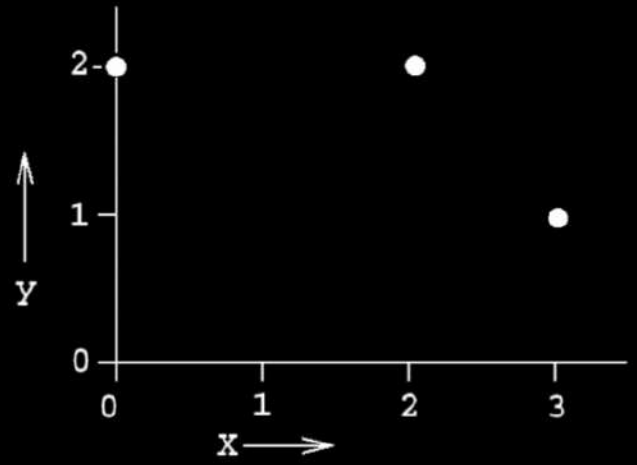
(c) [5 *pts*] Suppose you have this data set with one real-valued input and one real-valued output:



x	y
0	2
2	2
3	1

(c.2) Suppose we use a trivial algorithm of predicting a constant $y = c$. What is the mean squared leave one out error in this case? (Assume c is learned from the non-left-out data points.)

(c) [*5 pts*] Suppose you have this data set with one real-valued input and one real-valued output:



x	y
0	2
2	2
3	1

(c.2) Suppose we use a trivial algorithm of predicting a constant $y = c$. What is the mean squared leave one out error in this case? (Assume c is learned from the non-left-out data points.)

Answer: $\frac{0.5^2+0.5^2+1^2}{3} = 1/2$

Question:

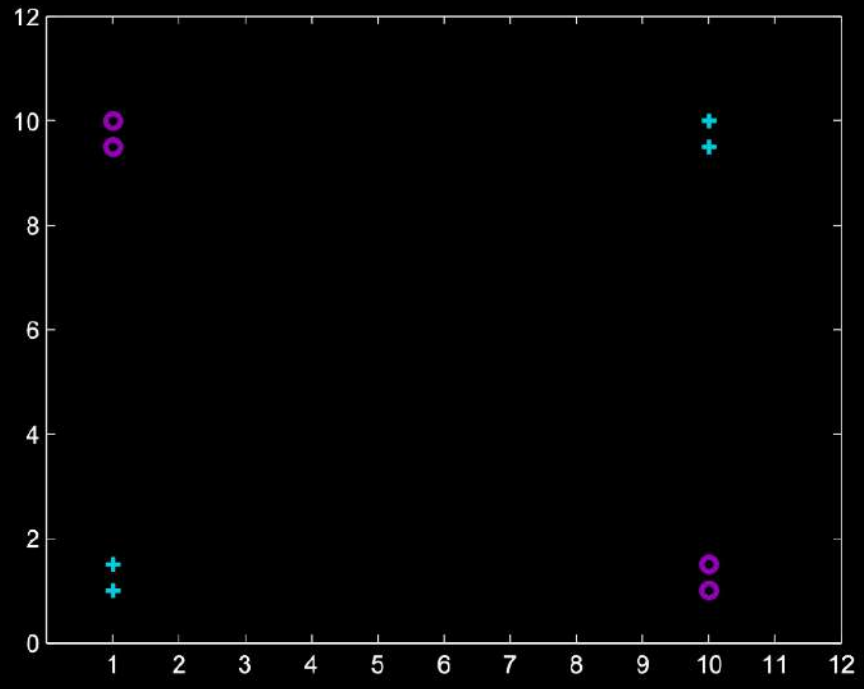
Given the 2D dataset, which of the following is true regarding the performance of 1-nearest neighbor (1-NN) and Support Vector Machines (SVM) in terms of leave-one-out cross-validation error (LOO error)?

A) 1-nearest neighbor (1-NN) has lower LOO error than SVM.

B) SVM has lower LOO error than 1-NN.

C) Both 1-NN and SVM have the same LOO error.

D) It's impossible to compare the LOO error between 1-NN and SVM for this dataset.



Question:

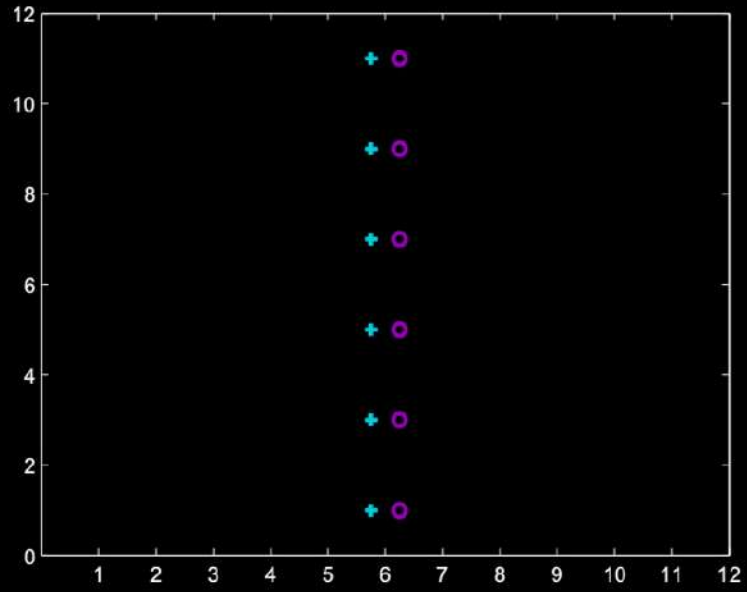
Given the 2D dataset, which of the following is true regarding the performance of 1-nearest neighbor (1-NN) and Support Vector Machines (SVM) in terms of leave-one-out cross-validation error (LOO error)?

A) 1-nearest neighbor (1-NN) has lower LOO error than SVM.

B) SVM has lower LOO error than 1-NN.

C) Both 1-NN and SVM have the same LOO error.

D) It's impossible to compare the LOO error between 1-NN and SVM for this dataset.



GO CLASSES

Question:

Consider a scenario where we use leave-one-out cross-validation (LOOCV) with Support Vector Machines (SVM) for binary classification. Given a dataset $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)\}$, the cross-validation error for a sample $(\mathbf{x}_n, y_n)$ is defined as:

$$e_n = \begin{cases} 1 & \text{if } \hat{y}_n \neq y_n \\ 0 & \text{otherwise} \end{cases}$$

where $\hat{y}_n$ is the predicted label for $\mathbf{x}_n$ based on the model trained without the sample $\mathbf{x}_n$. The overall cross-validation error is given by:

$$E_{\text{CV}} = \frac{1}{N} \sum_{n=1}^{N} e_n$$

Which of the following statements is true regarding the relationship between the overall cross-validation error E_{CV} and the number of support vectors K?

A. $E_{\text{CV}} \geq \frac{K}{N}$

B. $E_{\text{CV}} = \frac{K}{N}$

C. $E_{\text{CV}} \leq \frac{K}{N}$

D. E_{CV} is independent of K

https://exams.skule.ca/exams/ECE421H1_20201_631619146406Final_Winter_2020.pdf