Pattern Recognition
Homework 4 Announcement

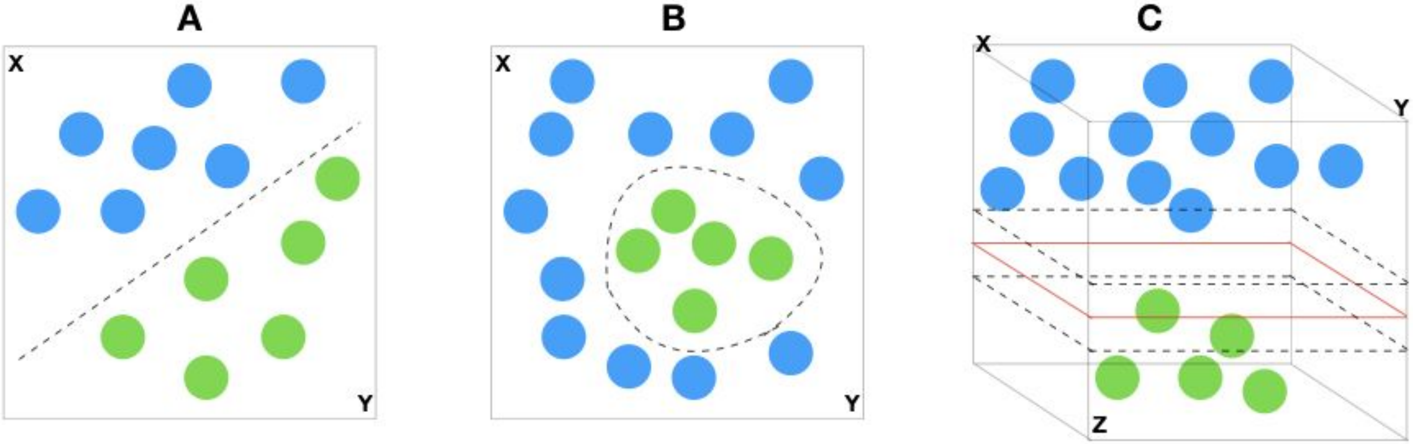
Lastest update: 2023.05.03 12:00

Homework 4

- Deadline: May. 17, Wed. at 23:59
  - Code assignment (50%)
    - Implement Cross-Validation and Grid Search for SVM training using only NumPy.
  - Questions (50%)
    - Write your answer in detail in the report.
- Question: Link
- Sample code: Link
- Dataset: Link

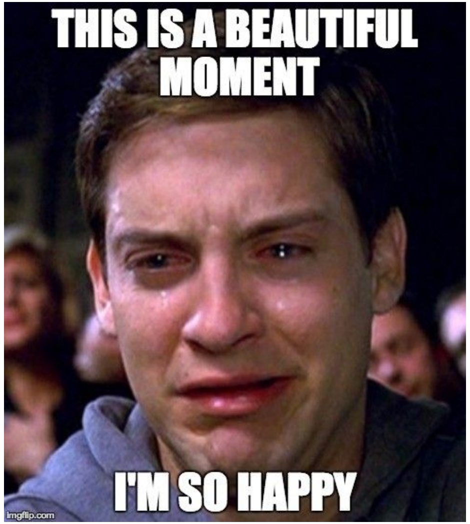
Support Vector Machines

- The Support Vector Classifier aims to identify the optimal hyperplane for separating distinct classes by maximizing the distance between the sample points and the hyperplane.



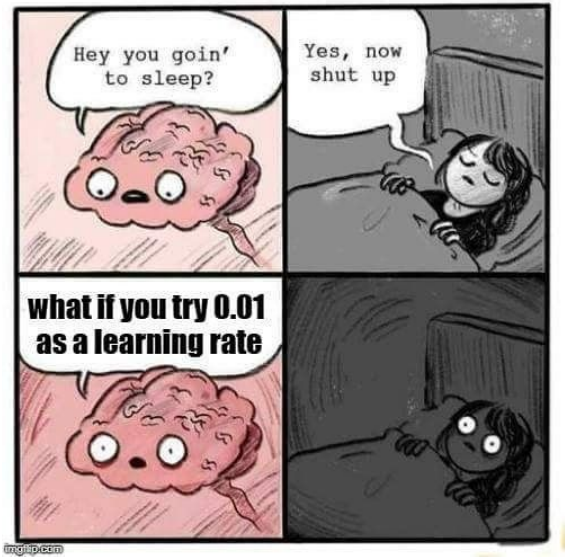
Support Vector Machines

- Since SVM involves complex mathematical operations, you are allowed to use the SVM implementation available in the sklearn library instead of implementing it from scratch.



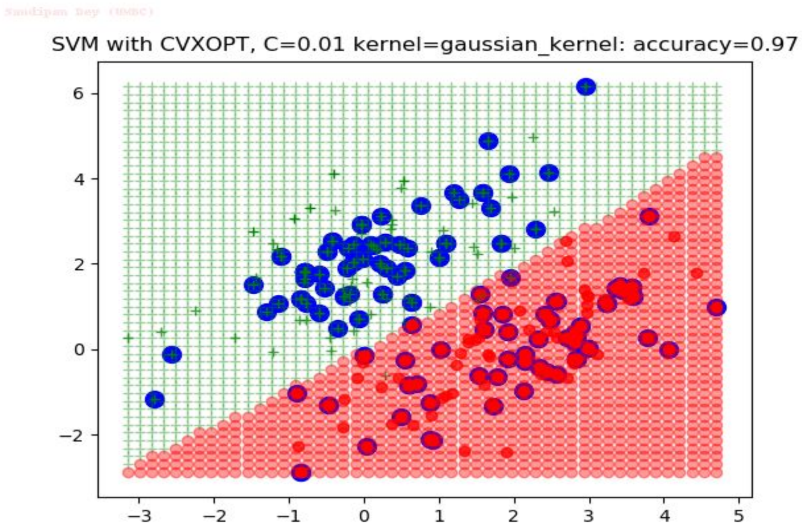
Grid Search and K-Fold Cross-Validation

- There are many hyperparameters in SVM.
- For this homework, you need to perform grid search and cross-validation to find the best hyperparameters for SVM on the provided dataset.



Grid Search

- Suppose we want to find the optimal values of two hyperparameters for an RBF kernel SVM, namely **C** and **gamma**.
- There are numerous combinations that need to be considered!
- Interactive demo

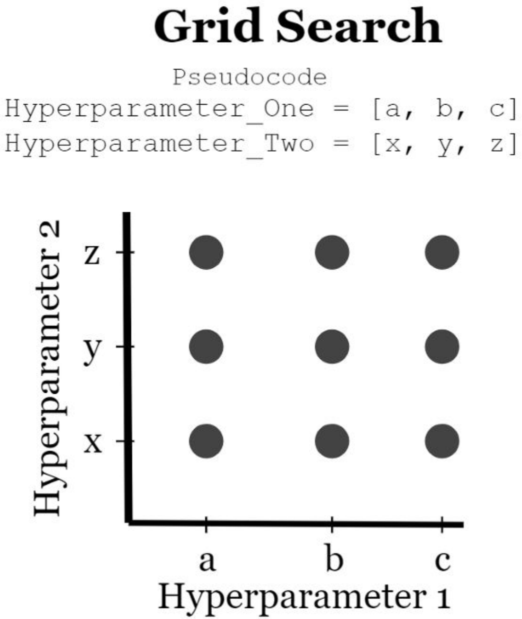


Grid Search

- Grid search explores all possible hyperparameter combinations and selects the best set of hyperparameters based on the model's performance.

```
C = [0.1, 1, 10] #3 values
gamma = [0.01, 0.1, 1, 10] #4 values
# There are totally 12 combinations for tuning
```

C\gamma	0.01	0.1	1	10
0.1	[0.1, 0.01]	[0.1, 0.1]	[0.1, 1]	[0.1, 10]
1	[1, 0.01]	[1, 0.1]	[1, 1]	[1, 10]
10	[10, 0.01]	[10, 0.1]	[10, 1]	[10, 10]



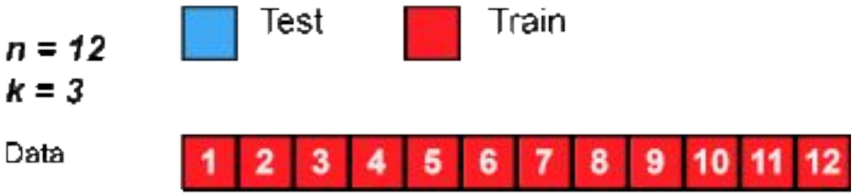
Grid Search

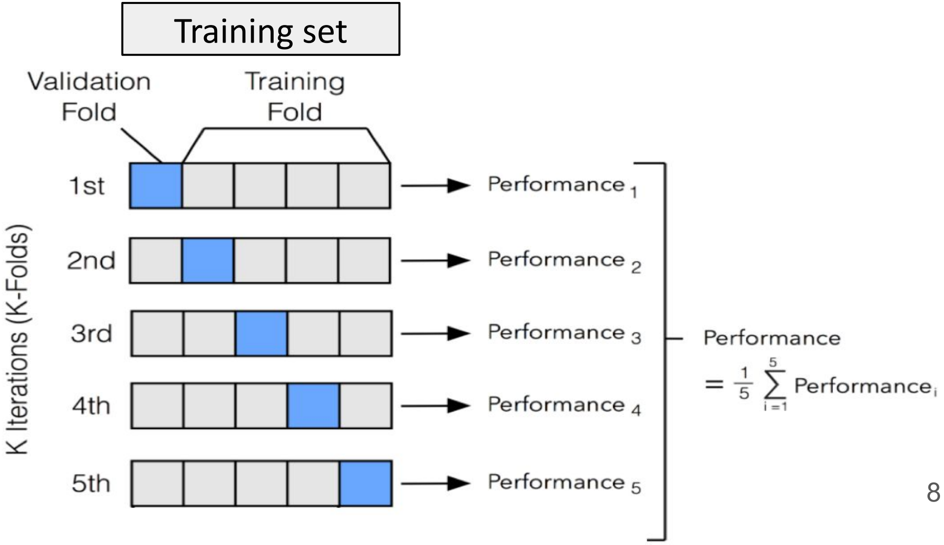
Pseudocode

```
Hyperparameter_One = [a, b, c]
Hyperparameter_Two = [x, y, z]
```

K-Fold Cross-Validation

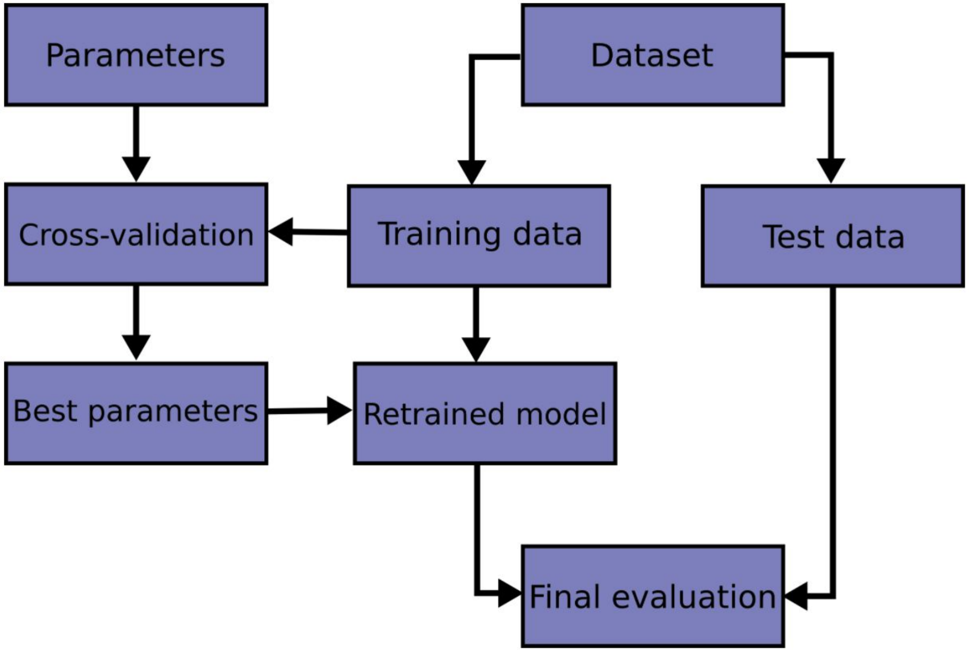
- We divide the dataset into K subsets.
  - One subset is used for validation.
  - The remaining K-1 subsets are combined to form the training set.
- This process is repeated K times, with each subset used once as the validation data.





K-Fold Cross-Validation

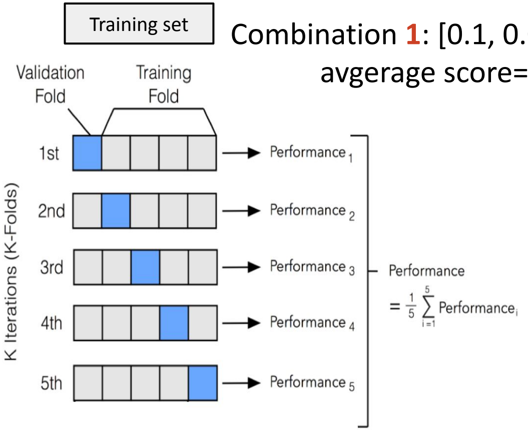
- Below is the workflow for k-fold cross-validation.



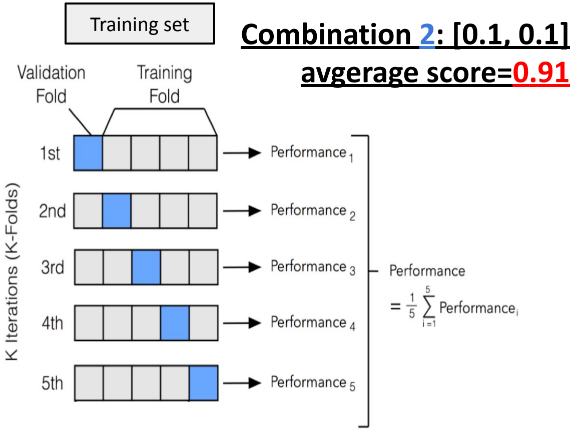
```
C = [0.1, 1, 10] #3 values
gamma = [0.01, 0.1, 1, 10] #4 values
```

Grid Search and K-Fold Cross-Validation

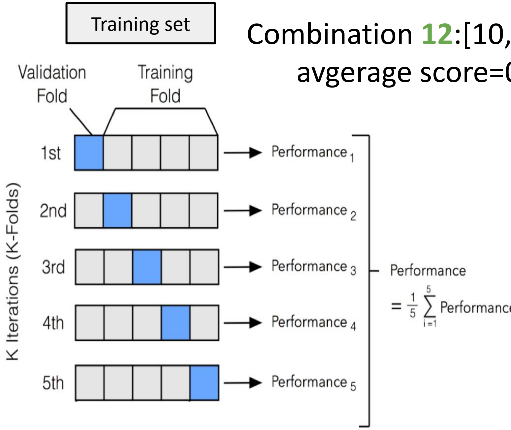
- We can experiment with 12 (3 x 4) combinations of hyperparameters as defined on page 7.
- For each combination, we can apply K-fold cross-validation and calculate the average performance.
- Find the combination that yields the best performance.



Combination 1: [0.1, 0.01]
avgerage score=0.8



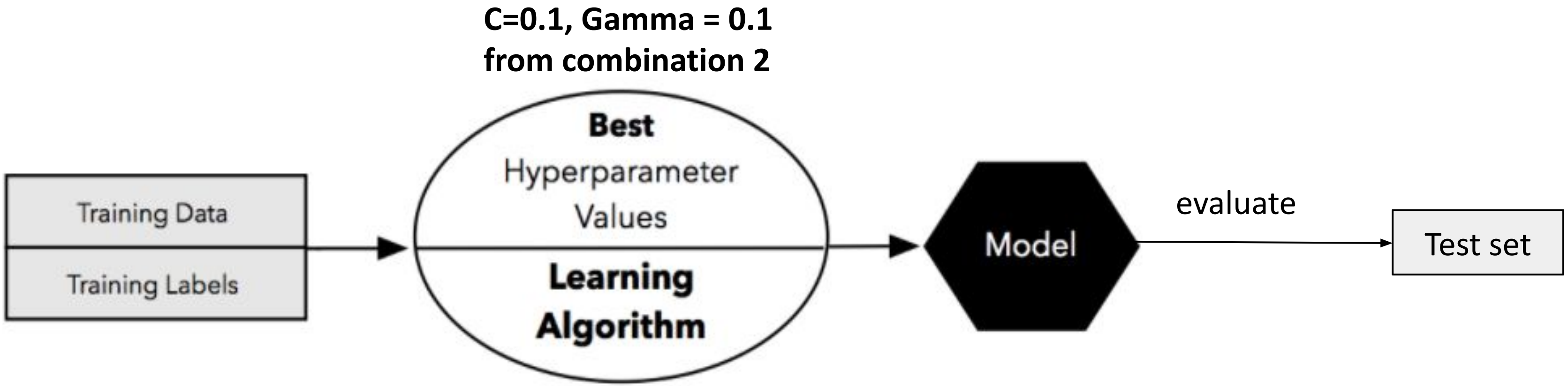
Combination 2: [0.1, 0.1]
avgerage score=0.91



Combination 12:[10, 10]
avgerage score=0.75

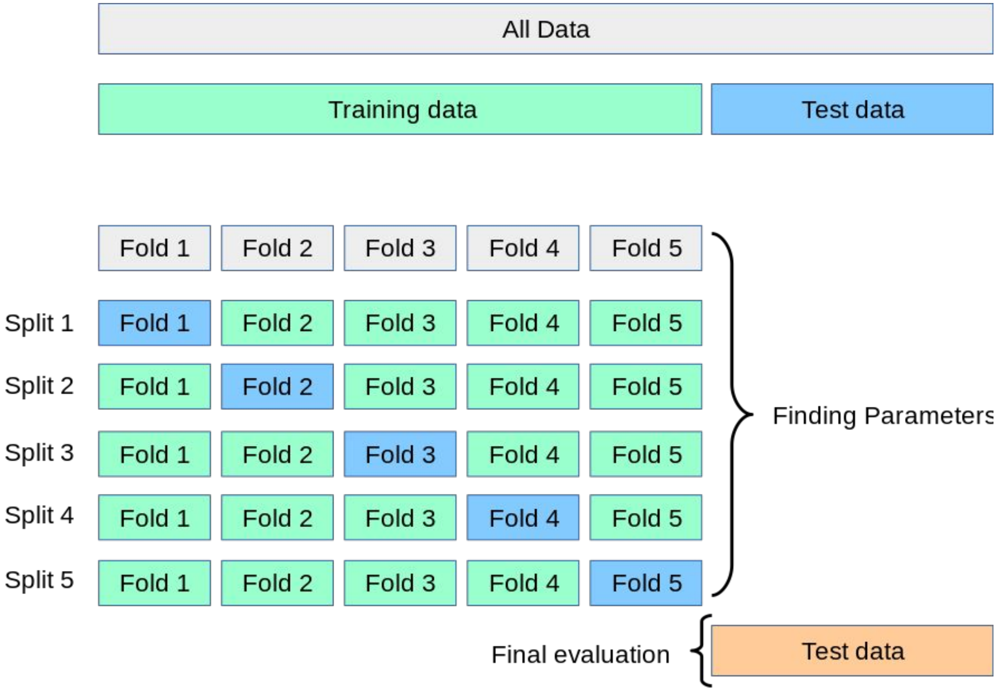
Grid Search and K-Fold Cross-Validation

- Finally, train your model on the entire training set using the best hyperparameters, and evaluate the model's performance on the test set.



Questions (50%)

1. (10%) Implement K-fold data partitioning using numpy.



All Data

Training data | Test data

Fold 1 Fold 2 Fold 3 Fold 4 Fold 5

Split 1 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5

Split 2 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5

Split 3 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5

Split 4 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5

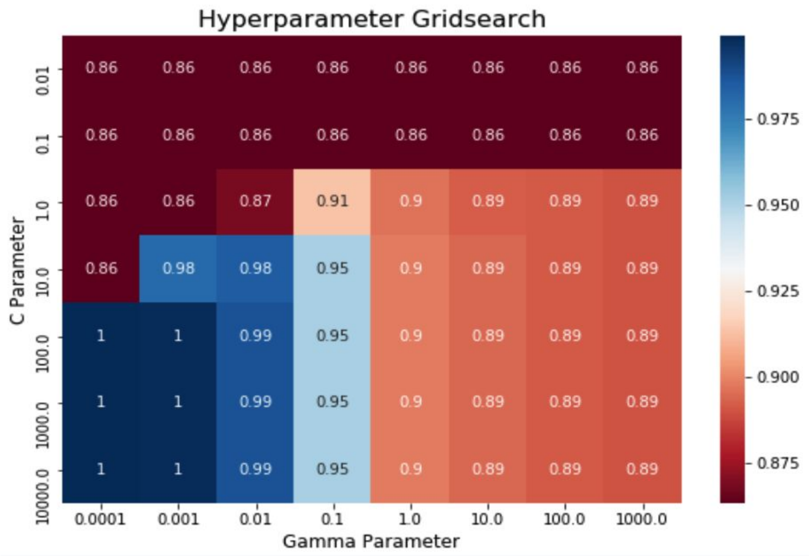
Split 5 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5

Finding Parameters

Final evaluation Test data

Questions (50%)

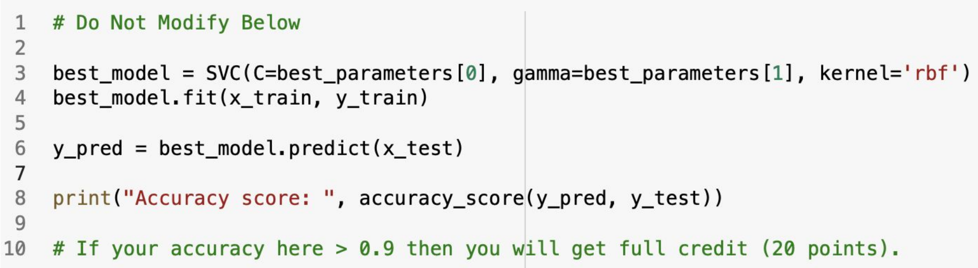
2. (10%) Perform a grid search on the hyperparameters **C** and **gamma** to identify the optimal values using cross-validation implemented by NumPy.
3. (10%) Plot the results of your SVM's grid search.



	0.0001	0.001	0.01	0.1	1.0	10.0	100.0	1000.0
0.01	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86
0.1	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86
1.0	0.86	0.86	0.87	0.91	0.9	0.89	0.89	0.89
10.0	0.86	0.98	0.98	0.95	0.9	0.89	0.89	0.89
100.0	1	1	0.99	0.95	0.9	0.89	0.89	0.89
1000.0	1	1	0.99	0.95	0.9	0.89	0.89	0.89
10000.0	1	1	0.99	0.95	0.9	0.89	0.89	0.89

Questions (50%)

4. (20%) Train your SVM model using the best hyperparameters found in Q2 on the entire training dataset, then evaluate its performance on the test set. Print your testing accuracy.

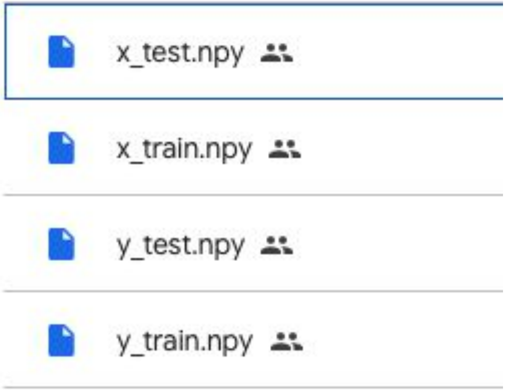


```
# Do Not Modify Below

best_model = SVC(C=best_parameters[0], gamma=best_parameters[1], kernel='rbf')
best_model.fit(x_train, y_train)

y_pred = best_model.predict(x_test)

print("Accuracy score: ", accuracy_score(y_pred, y_test))

# If your accuracy here > 0.9 then you will get full credit (20 points).
```

Accuracy	Your scores
acc > 0.9	20 points
0.85 <= acc <= 0.9	10 points
acc < 0.85	0 points

Dataset

- Training set: 7000
- Testing set: 3000
- 300 features, 2 labels

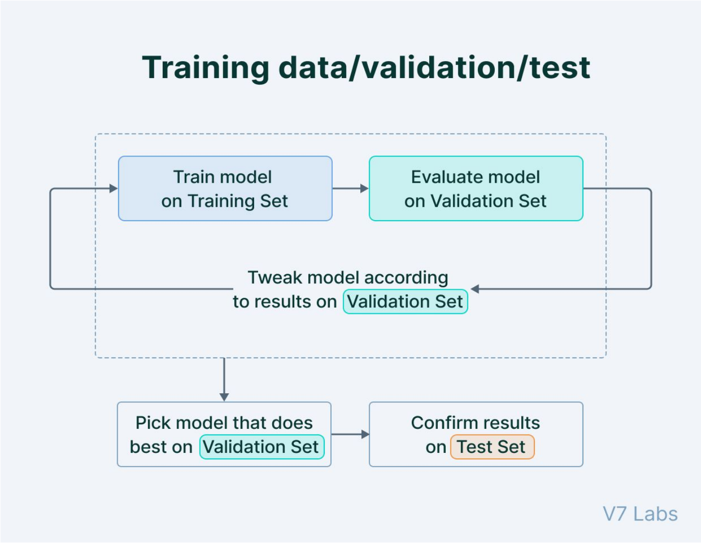


x_test.npy

x_train.npy

y_test.npy

y_train.npy



Training data/validation/test

Train model on Training Set

Evaluate model on Validation Set

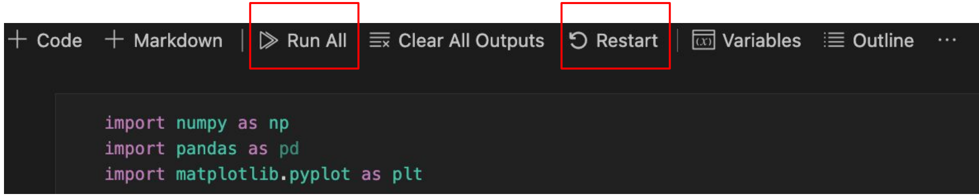
Tweak model according to results on Validation Set

Pick model that does best on Validation Set

Confirm results on Test Set

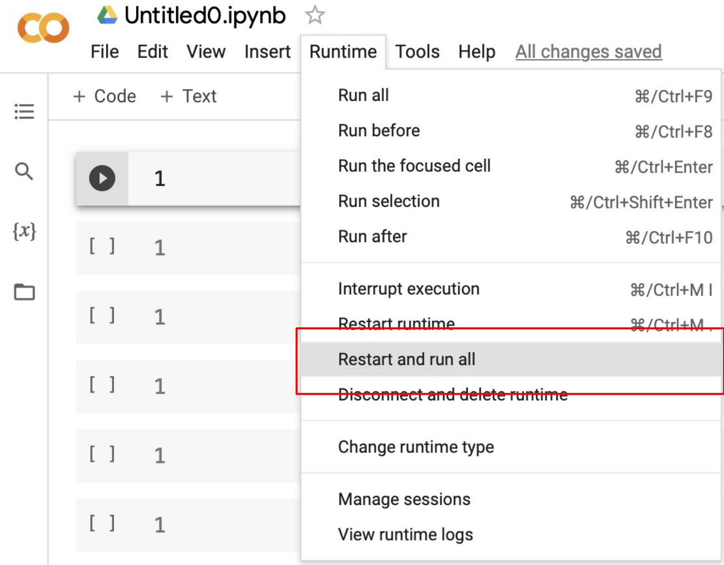
V7 Labs

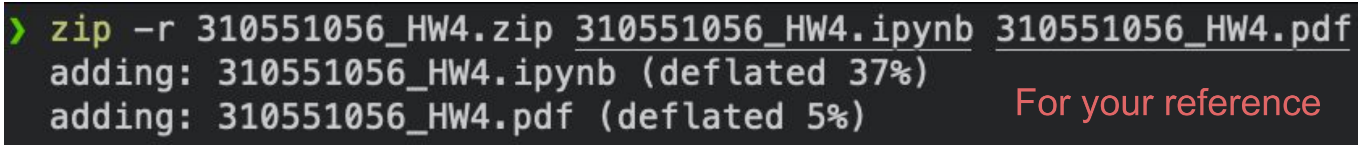
Submission



```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

- Compress your .ipynb and .pdf into a zip file and submit it on E3.
- Before submission:
    - Restart and run All
    - Save and submit the .ipynb (keep all cell outputs)
    - Get 0 points if you do not keep the cell outputs.
- <STUDENT ID>_HW4.zip
    - <STUDENT ID>_HW4.ipynb
    - <STUDENT ID>_HW4.pdf
- No csv file in HW4.





```
❯ zip -r 310551056_HW4.zip 310551056_HW4.ipynb 310551056_HW4.pdf
  adding: 310551056_HW4.ipynb (deflated 37%)
  adding: 310551056_HW4.pdf (deflated 5%)
```

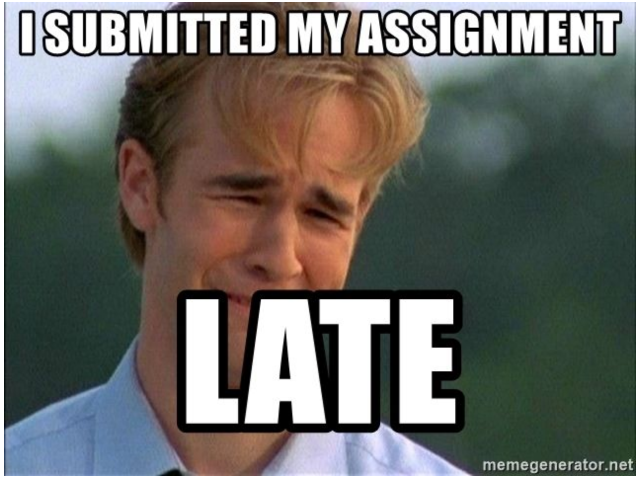
For your reference

Report

- Please write your report in English.
- Please follow the HW1 report template.
- You must type the answer and also screenshot at the same time for the coding part.
- **Answer each question as clearly as possible**. You will get an extra penalty for only the brief answer.

Late policy

- We will deduct a late penalty of 20 points per additional late day.
- If you get 90 points but delay for two days, you get 90 - (20 x 2) = 50 points!
- We only accept submissions that are up to 10 minutes late. Any submissions that are later than that will be considered late, regardless of the reason.



Reference

- K-fold Cross-Validation & Grid Search
- SVM hyperparameter tuning