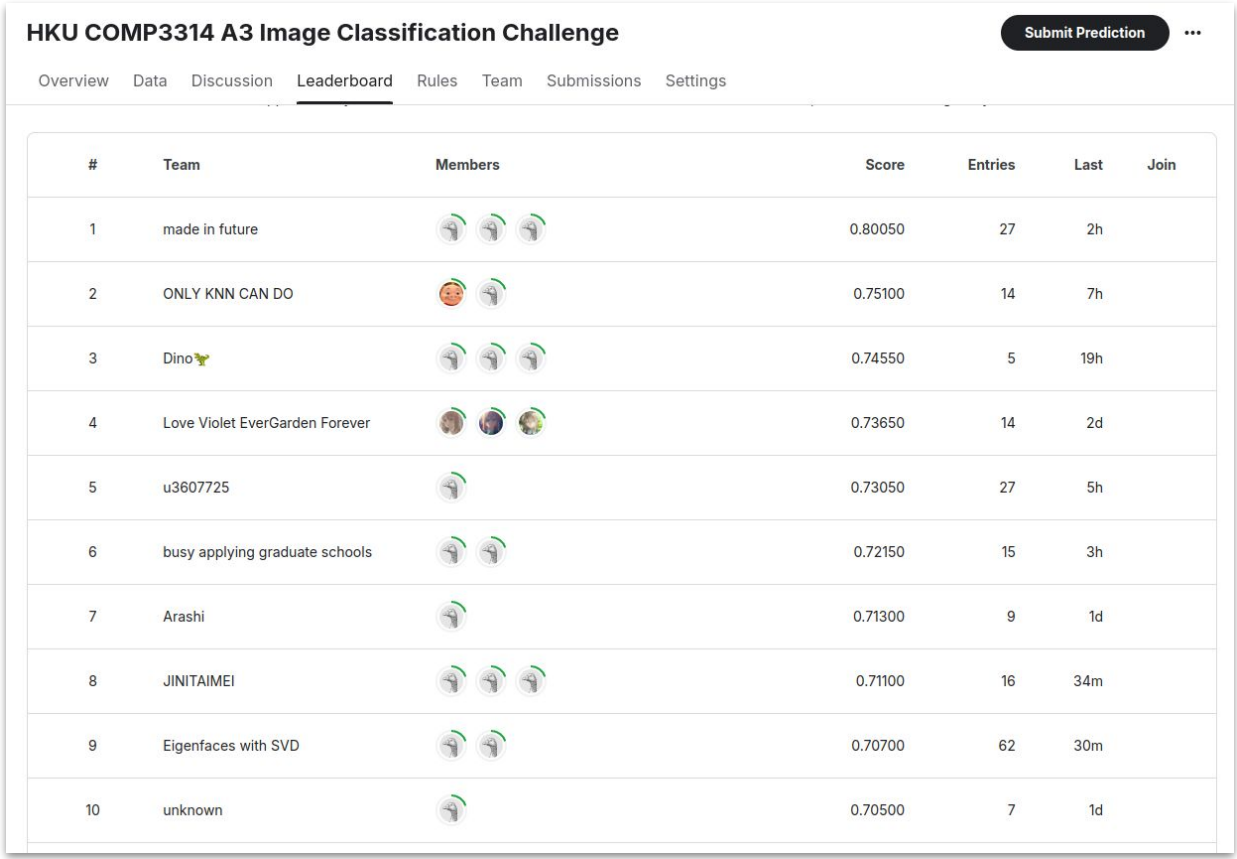
COMP3314 Tutorial 4

Tutorial for Assignment 4

TA for COMP3314

Kaggle competition statistics

- Deadline:
  - Nov 24 Sun, 23:59
- Up till Nov 22 Fri:
  - 115 Entrants
  - 48 Participants
  - 30 Teams
  - 371 Submissions
- 2 days to go!



HKU COMP3314 A3 Image Classification Challenge

Submit Prediction

Overview Data Discussion Leaderboard Rules Team Submissions Settings

#	Team	Members	Score	Entries	Last	Join
1	made in future		0.80050	27	2h	
2	ONLY KNN CAN DO		0.75100	14	7h	
3	Dino🦖		0.74550	5	19h	
4	Love Violet EverGarden Forever		0.73650	14	2d	
5	u3607725		0.73050	27	5h	
6	busy applying graduate schools		0.72150	15	3h	
7	Arashi		0.71300	9	1d	
8	JINITAIMEI		0.71100	16	34m	
9	Eigenfaces with SVD		0.70700	62	30m	
10	unknown		0.70500	7	1d	

Kaggle competition reminder

- Sign-up Form: https://forms.gle/KGjk5SN2pqWcCvhX7
  - Submission method: Google Forms
  - Everyone shall submit
  - You must fill this form, otherwise we cannot associate your submission
- PDF Report: Document your process, findings, and methodology (30 pts)
  - Submission method: Moodle
  - Everyone shall submit
- Jupyter Notebook: Share your code in a runnable notebook (20 pts)
  - Submission method: Moodle
  - Everyone shall submit
- Prediction CSV: Submit your predictions on Kaggle for scoring (50 pts)
  - Submission method: Kaggle

Assignment 4 overview

- Linear regression (30 points)
- Spam email classification with MLP (30 points)
- K-means clustering for color compression (40 points)

Q1: Linear regression (30 points)

Year	Advertisement Cost (x_i)	Annual Profit (y_i)
2020	20	60
2021	40	65
2022	60	75
2023	80	85

- Given the data of cost and revenue, try to fit a linear function y = w_0 + w_1 * x to the data with linear regression
- Your tasks:
  - First, compute the solutions of w_0 and w_1 manually
    - Formulate the loss function
    - Compute the gradients w.r.t. w_0 and w_1
    - Solve for w_0 and w_1 by setting the gradients to zero
  - Then, implement the linear regression model using scikit-learn and compare the results

Q2: Spam email classification with MLP (30 points)

- Given a dataset of emails, you are to build a spam classifier with MLP
- The code for downloading and preprocessing the data has been provided
- Your tasks:
  - Build MLP neural networks with different configurations
  - Train and evaluate the classifiers with the given dataset
  - Ensemble the classifiers and evaluate the ensemble classifier

Q3: K-means clustering for color compression (40 points)

- Implement the update steps for K-means
- Implement random initialization for K-means
  - Run your K-means algorithm and visualize color compression results
- Implement the FPS initialization method
  - Run your K-means algorithm and visualize color compression results
- Compare the results with random and FPS initialization
  - Qualitatively by visualization
  - Quantitatively with MSE
- FPS algorithm
  - Starts from a randomly selected sample as the first center
  - Then in each iteration, it selects the sample that is the farthest from the set of selected centers

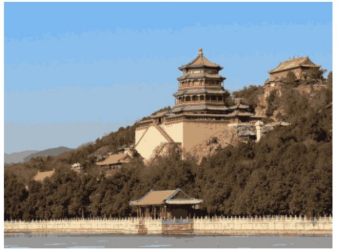
Q3: K-means clustering for color compression (40 points)

Original Image

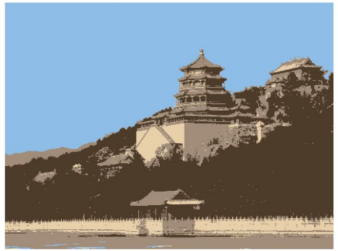
Random Init, 4 Clusters

Random Init, 8 Clusters

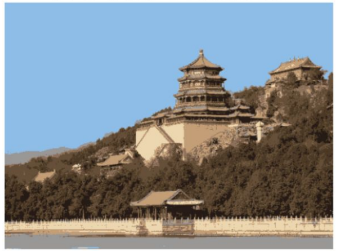
Random Init, 16 Clusters



FPS Init, 4 Clusters



FPS Init, 8 Clusters



FPS Init, 16 Clusters

Q & A