Project Topics:

You are expected to first of all choose a project topic that you are really interested in. Please pick something that you can get excited and passionate about. Meanwhile, you should also consider the difficulty of each project. Once you chose the topic from the following pool of five topics, it can be very helpful for you to look up existing research on relevant topics by searching related keywords on Google and Google Scholar (http://scholar.google.com).

Topics:

Topic 1: Building your Future Sales predictor [Finance]

Suggested proposals: In this project, you will be expected to use either genetic algorithms or regression machine learning algorithms like Kernel Ridge Regression, XGBoost and Recurrent Neural Network to forecast the total amount of products sold in every shop in the future.

Data: https://www.kaggle.com/competitions/competitive-data-science-predict-future-sales

Topic 2: Building your birds classification mode [Ecology]

Suggested proposals: In this project, you will be expected to use Convolutional Neural Network to predict the species of each bird.

Data: http://www.vision.caltech.edu/datasets/cub_200_2011/

Topic 3: YouTube Spam Classifier [Entertainment]

Suggested proposals: In this project, the task is to predict whether a comment about a YouTube video is spam or not. Here are some examples of spam and not spam:

- Spam: “Huh, anyway check out this you[tube] channel: kobyoshi02”
- Not spam: “Still watching this 2 years later?”

Your goal is to train a classifier to predict whether a comment is spam or not.

Hint: You need to train classifiers using the training data, and then predict on the test data. You are free to choose the feature extraction method and classifier algorithm.

Data: The training data is in the text file youtube_spam_train.csv. This CSV file contains the comment text, comment author names, date, video IDs, and classes. The class labels are 0 for “not spam”, and 1 for “spam”. The testing data is in the text file youtube_spam_test.csv, and includes everything except the class labels.

Topic 4: Music Genre Classification [Music] [extra bonus: original score * 0.05 (upper bound = 100)]

Suggested proposals: In this project, the task is to develop a deep learning project to automatically classify different musical genres from audio files. You are required to classify these audio files using their low-level features of frequency and time domain.

Hint: Some of these methods could be used: 1) Multiclass SVM; 2) K-means clustering; 3) K-nearest neighbors; 4) Convolutional neural networks.

Data: The GTZAN genre dataset has 10 classes (10 music genres), each containing 100 audio tracks. Please divide the last 20 audio tracks of each genre as the test set.
https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification?select=Data

Topic 5: Skin Cancer Classification [Life Science] [extra bonus: original score * 0.07 (upper bound = 100)]

Suggested Proposal: In this project, you will be expected to use Convolutional Neural Network to predict the labels of dermoscopic images among eight different diagnostic categories.

Hint: Please note that the given training dataset contains class sizes that are not evenly distributed, whereas the testing images are evenly divided into the eight categories. (100 images per class)

Data: See the attached files.

Topic 6: Controlling your agents with Reinforcement Learning [Robotics] [extra bonus: original score * 0.1 (upper bound = 100)]

Suggested proposals: In this project, you will be expected to use reinforcement learning algorithms to control the agent to play games in a 2D box.

Environment: https://www.gymlibrary.dev/environments/box2d/

Evaluation:

The project will be evaluated based on:

- 70%: The technical correctness and quality of your work. (Does your technical material make sense? Are the submitted codes written completely by the students themselves? Are your proposed algorithms existing or with minor improvements? Compared to the state-of-the-art (surveyed by you), how is your performance? Do the students have insights?)
- 30%: Final report. (Do the students clearly describe the problem and the techniques for solving the problem? How much is each student's contribution?)

Report Guidelines:

- Background: What problem are you tackling? What is the setting you are considering?
- Method: Which algorithm (e.g., genetic algorithm, Logistic regression, Convolutional neural network, reinforcement learning) is used and why?

- Experiments: Describe the experiments that you have run, including the prediction accuracy or some analyses that you have done.
- Contributions: describe what each team member contributed to the project and how much.

You are also required to submit your codes packaged into a .zip file.

Academic dishonesty:

We will be checking you code against other submissions in the class for logical redundancy. If you copy others' codes and submit with only minor changes, we will know. We trust you all to submit your own original work, instead of copying others'. Please do not try to fool us; otherwise, we will apply the strongest disciplinary actions.

Getting help:

If you find yourself have difficulties and problems, contact the teaching assistant for help. Piazza forums, emails, and the offline Tutorial session are there for your support. Please do not hesitate to use them. We expect the project to be rewarding and inspirational for you, but not demoralizing.