CSE848: Evolutionary Computation
Michigan State University
Assignment HA2: Home Assignment 2

Jaturong Kongmanee

1. There is much current research in producing autonomous vehicles that can be used on real roads. For each of the following capabilities that such a system should exhibit, state whether they are an optimization, modeling, or simulation problem. Explain your reasoning.

(a) Steering in the middle of the road.

- `Modeling` – In this scenario, we try to understand the problem by looking for what are the factors, e.g., speed, steering wheel angle, that affect the model's performance in steering. The model's input might be the visual and sensed data from cameras and sensors equipped with the car. The model's output might be the continuous or discrete values indicating whether the car is in the middle of the road.

(b) Avoiding a child that runs into the road.

- `Simulation` - The system's model is being tested with the assumed inputs that contain various behaviors of having a child running into the road. The goal is to see how well the created model reacts to these situations, and avoids a child on the road, then explores the possible flaws that happened to improve the model further.

(c) Recognizing a traffic sign in a video feed as the vehicle drives along.

- `Simulation` – The system's model is being tested with the new examples of never seen traffic sign input in a video feed to see how the created model accurately recognizes a traffic sign.

(d) Planning shortest, or quickest, route between two places.

- `Optimization` – In this situation, we have the model of measuring the distance between the two places, i.e., Euclidean distance. The desired output is specified that the route must have to be shortest between two places. Thus, the task is to find the possible best inputs.

(e) Learning to recognize traffic signs.

- `Modeling` – This is an example of classification. The system's model is trained to assign each input vector of traffic signs to one of a finite number of discrete categories of known target vectors.

2. List the three kinds of learning that can be discerned in Machine Learning. Give an example of each and explain, why it falls into that category of learning.

1. Supervised Learning – the training data comprises examples of input vectors along with their

corresponding target vectors.

An example of Supervised Learning:

- `Classification` – the problem consists of assigning each input vector to one of a finite number of discrete categories of target vectors.

Some variant of classification problem:

- `Sequence generation` – given an image, predict an image caption.

2. Unsupervised Learning – the training data consists of input vectors without any corresponding target values.

An example of Unsupervised Learning:

- `Clustering` – the problem consists of discovering groups of similar examples within the training data.

3. Reinforcement Learning – this kind of learning is concerned with the problem of finding suitable actions to take in a given situation/environment in order to maximize a reward. Note that the learning algorithm (agent) is not given examples of optimal outputs.

An example of Reinforcement Learning:

- `A mobile robot` – In this example, a mobile robot decides whether it should enter a new room to search for more trash to collect or start trying to find its way back to its battery recharging station. The agent's decision is based on the current charge level of its battery and how quickly and easily it has been able to find the recharger in the past.

This example shows the interaction between the agent and its environment, in which an agent finds suitable actions to take to maximize reward despite uncertainty about its environment.

reference: http://incompleteideas.net/book/the-book.html

3. The use of datasets in Machine Learning

(a) Explain the difference between the three uses of datasets common in Machine Learning:(i) Training, (ii) validation, and (iii) testing.

- `Training Set` - is used actually to train the algorithm.
- `Validation Set` - is used to evaluate and keep track of how well the algorithm is doing as it learns (i.e., overfitting).
- `Testing Set` - is used to finally test the model once it is ready for using.

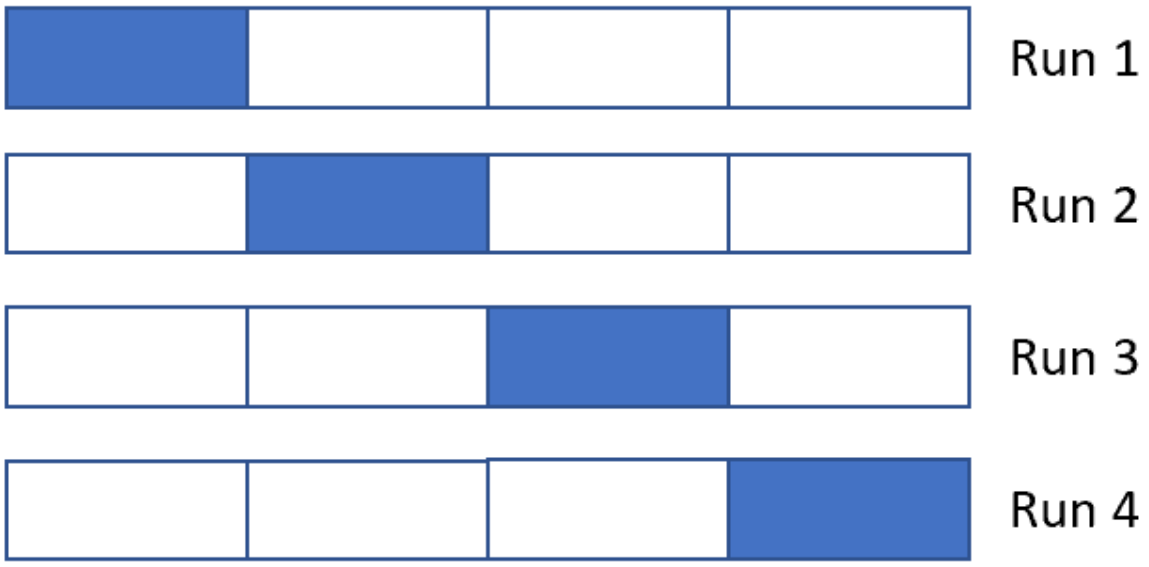
(b) Suppose your entire dataset has 10,000 samples. Roughly, how would you distribute your samplesamong them between training, validation and testing?

- The validation and testing sets should be big enough to represent the model's performance and be able to reflect expected data to get in the future. Also, the complexity of the model is another factor in defining the size of each data set. Thus, if the model is complex (i.e., deep

neural network), I will provide more training data and split the data with 9,000 (90%) in the training set, 500 (5%) in the validation set, and 500 (5%) in the testing set. However, if the model is simple, I will split the data with 7,000 (70%) in the training set, 1,500 (15%) in the validation set, and 1,500 (15%) in the testing set.

(c) Another method to use datasets in ML is by n-fold cross-validation. Make yourself familiar with this method and explain it briefly. What is the advantage of n-fold cross-validation compared to the previous method?

- This method involves splitting the available training data into *N* partitions (each partition is of equal size for the simplest case). For each run *i*, train a model (or a set of models) on the remaining *N-1* partitions and evaluate it on partition *i*. The performance scores of the *N* scores obtained from *N* runs are then averaged. N-fold cross-validation can be illustrated as the figure below, where $N = 4$.



- This method is useful when there is a limited number of training and testing data. It uses *(N-1)/N* of the available data for training and uses all of the data to evaluate performance.

In [ ]: