Machine Learning

Note: In the ML portion of the exam there will be a set of **five** multiple choice questions as in Q1, and **two** short answer questions, each similar to Q2.

1) Multiple Choice Questions

(a) The i.i.d. assumption over the training and held-out (e.g. test) datasets is best described as:
   a. The datasets are both sampled for the same classes, in e.g. a classification problem.
   b. **The datasets are independent and drawn from the same probability distribution.**
   c. The datasets can be collected in any way, as long as they are sampled from the real-world.
   d. Training samples can also be present in other (held-out) datasets.

(b) Given a dataset with labels, we can use _____ learning.
   a. Reinforcement
   b. Meta
   c. **Supervised**
   d. Unsupervised

(c) The Vapnik-Chervonenkis dimension, or VC dimension, measures the _____ of a binary classifier.
   a. Overfitting
   b. Mutual Information
   c. **Capacity**
   d. Generalization

(d) Hyperparameters are best described as:
   a. Parameters of a model that must be optimized.
   b. Parameters that depend on the dataset used for training.
   c. Parameters of the learning algorithm.
   d. **Parameters that control a learning algorithm's behaviour but are not adapted by the learning algorithm itself.**

(e) Nonparameteric models are best described as:
   a. Models without parameters.
   b. Models without hyperparameters.
   c. **Models with a varying number of parameters, often based on dataset size.**
   d. Models trained without data.

2) You are given a dataset of 40,000 labelled images and tasked with performing supervised training of a machine learning model, while also tuning hyper-parameters of your model and evaluating the generalization performance of your final trained model with the tuned hyperparameters.

    a. Explain what datasets you would create, how you would create them, and roughly how large they would be.

Randomly sample three datasets (training, validation and test) from the 40,000 images without replacement. Training should be the largest, e.g. 30,000, with test and validation of relatively small but significant size, e.g. 5000 images each.

    b. How would you use one or more of the datasets you described above to tune the hyperparameters of the model? Explain why this is necessary.

i. For each set of hyperparameters we wish to attempt, we would train a model using the training set, and evaluate the performance using the validation set.

ii. We need a held-out dataset to test the generalization performance of a model trained with a candidate set of hyperparameters on dataset not seen during training.

    c. How would you use one or more of the datasets you described above to evaluate the generalization of the final model you trained with the tuned hyperparameters? Explain why this is necessary.

i. The final model would be trained on the training set with the tuned hyperparameters, and then evaluated on the test set.

ii. We need a held-out dataset to test the generalization performance of a model trained with the tuned set of hyperparameters. The validation set is not considered held-out since it is used to optimize/tune the hyperparameters. Only the test dataset is truly held-out in this instance and can let us evaluate the true generalization performance of our model and tuned hyperparameters.