TU Delft Delft University of Technology

Faculty of Architecture and the Built Environment

Final Exam

Course: Machine Learning for the Built Environment (GEO5017) – 5 EC

Location: CT-IZ 2.02 (in Civil Engineering)

Time: 12 April (Tuesday) 2022, 09:00 - 11:30

Instructors: Liangliang Nan, Shenglan Du, Nail Ibrahimli

1. This exam is open-book. Printed books, handouts, lecture slides, assignment reports, and any notes you've taken are allowed, and **you must answer the questions alone** without any help from anybody else.
2. **Electronic devices are forbidden** on your person, including cell phones, iPads, headphones, and laptops. Turn your cell phone off and leave all electronics at the front of the room.
3. This exam consists of two parts:
   (a) 20 multiple-choice questions (40 points). Each question has a single correct answer.
   (b) 6 open questions (60 points).
4. The total number of points is 100, and the minimum number of points to pass this exam is 55 (i.e., 55%).
5. If you think a question is ambiguous, mark what you think is the best answer. As always, we will consider written regrade requests if your interpretation of a question differed from what we intended.
6. Before starting,
   (a) check if all pages and questions are present. If not, ask the teacher.
   (b) fill out your **student ID** and **full name**.
   (c) read the instruction at the beginning of each part beginning before you answer the questions.
7. The exam questions are confidential and breaching this confidentiality agreement can be penalized.

Student ID: _____ **Name**: _____

Part 1 Multiple choice questions*[40 points]*

- *Every question has 4 choices and only 1 correct answer*
- *Every question counts for 2 points and the grading is based on*
  - *Answer is correct: 2 points*
  - *No answer is provided or 'not sure' is indicated: 0 point*
  - *Answer is wrong: -1 point (to discourage random guessing)*

1. "A computer program is said to learn from experience E with respect to some task T and some performance measure P if its performance on T, as measured by P, improves with experience E". Suppose we feed a learning algorithm a lot of historical weather data and have it learned to predict the weather. What would be a reasonable choice for P?
   A. The process of the algorithm examining a large amount of historical weather data.
   B. The weather prediction task.
   C. The probability of it correctly predicting a future date's weather.
   D. None of these.

   Answer: _____

2. One of the problems below can be best addressed using a supervised learning algorithm, and the others with an unsupervised learning algorithm. Which of the following would you apply supervised learning to? (In each case, assume some appropriate dataset is available for your algorithm to learn from.)
   A. Examine a large collection of emails that are known to be spam emails, to discover if there are sub-types of spam emails.
   B. Take a collection of 1000 essays written on the Dutch Economy and find a way to automatically group these essays into a small number of groups of essays that are somehow "similar" or "related".
   C. Given historical data of children's ages and heights, predict children's height as a function of their age.
   D. Given genetic (DNA) data of 100,000 patients, find the common patterns in their genes that may relate to a certain disease.

   Answer: _____

3. In the following statements about logistic regression, which one is NOT correct?
   A. Logistic regression is a machine learning model for the task of classification.
   B. Logistic regression minimizes the number of erroneously classified objects of the training set.
   C. Logistic regression assumes the posterior probability of the class y is a logistic sigmoid of a linear function of the input data samples $\mathbf{x}$.
   D. Logistic regression is more robust to the data distribution of the input $\mathbf{x}$ compared to the standard linear classification (i.e., using least squares as its loss function).

   Answer: _____

4. Many machine learning models learn a discriminant function that directly maps from the input data $\mathbf{x}$ to the output class y. Such functions will generate one or several decision boundaries which partition the input feature space into discrete regions. Which one of the following classifiers is ONLY able to generate a linear decision boundary?

    A. A SVM classifier with a polynomial kernel of degree 2.

    B. A random forest classifier with 20 decision tree estimators.

    C. A neural network with 4 layers of perceptrons.

    D. A standard linear classifier.

    Answer: _____

5. Which of the following is FALSE about sigmoid activation?

    A. Sigmoid function smooth approximation of the step function.

    B. Sigmoid function is zero-mean non-linear function.

    C. Sigmoid functions can cause vanishing gradients problem.

    D. Sigmoid function brings non-linearity to the network.

    Answer: _____

6. Which of the following is FALSE about CNN?

    A. Convolutional filter operations are usually followed by activation functions

    B. Having more layers always leads to better results.

    C. Fully connected layers use extracted features to classify the data.

    D. Shallow networks can help with overfitting in the case of small training data.

    Answer: _____

Part 2 Open questions *[60 points]*

- *No point will be awarded without a valid explanation*
- *Keep your answers as concise/short as possible. The space below each question is more than sufficient to provide a complete answer.*

1. The following figure demonstrates four sets of (2D) feature points with different distributions. Please suggest the proper clustering method for each feature distribution, and justify your choice of the method. *[10 points]*

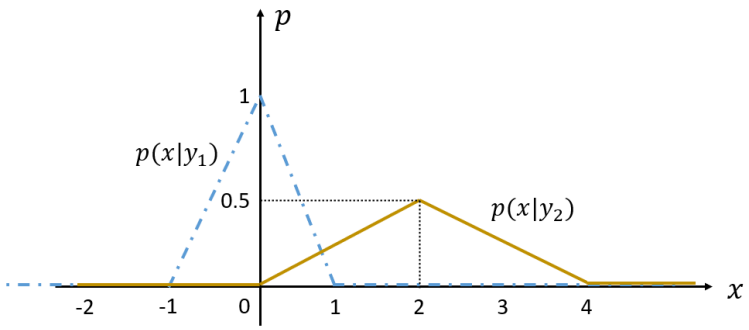
2. If the model used for regression is

$$y = a + b(x - 1)^2,$$

is it a multivariate linear regression problem? If your answer is "No", then what type of linear regression is it? Please explain your answer. *[10 points]*

3. Two class conditional probability density functions are given in the following figure. The first class y_1 is represented by a dashed blue line and the second class y_2 is represented with a solid yellow line. Two classes have equal prior probability:

$$P(y_1) = P(y_2) = \frac{1}{2}$$



(a) Use the Bayes' rule (so you can derive the class posterior probabilities), please answer to which class each of the following objects should be assigned? *[6 points]* of the following objects:

- $x_1 = -0.5$.
- $x_2 = 0.5$.
- $x_3 = 3$.

(b) What is the decision boundary of the Bayes classifier with such input data distribution? *[4 points]*

4. You want to train an SVM model on your point cloud dataset for the task of object classification.

   (a) Describe the high-level objective of the SVM classifier (i.e., you can use figures and mathematical formulations to illustrate the objective.) *[6 points]*

   (b) Describe why you need a train-test split when running your machine learning experiment, and how you will split the dataset. *[4 points]*