Frontpage IN3050

IN3050 - Introduction to Artificial Intelligence and Machine Learning

Spring 2022

Tuesday, May 31
09:00 AM - 01:00 PM (4 hours)

All questions should be answered!
Each question is assigned a weight which is indicated.
The maximum number of points for the whole set is 100 points.

Permitted materials: None
An on-screen calculator is available.

You may answer in English, Norwegian, Danish or Swedish.

1 Search and Optimization

The search methods *gradient ascent/descent* and *hill climbing* are somewhat similar: Both work by taking a solution and iteratively change it a bit, moving step by step towards a local optimum. But there are also some important differences.

a. Describe how gradient ascent and hill climbing differ with regards to how they change the current solution into the next solution (which is hopefully a bit closer to the optimum).
b. For certain problems we can use hill climbing but not gradient ascent/descent. What kinds of problems are these? Why can't we use gradient ascent/descent there?

Fill in your answer here

Maximum marks: 6

2(a) Evolutionary Algorithms - Representations

Briefly describe a problem where candidate solutions are best represented as permutations. Why are other representations less relevant/efficient for this problem?

Fill in your answer here

Maximum marks: 3

2(b) Evolutionary Algorithms - Representations

Consider the solution [1 3 2 5 4] to a problem where solutions are permutations of the numbers from 1 to 5. Suggest a mutation operator that is appropriate for permutations and show an example of how it may transform the solution.

Fill in your answer here

Maximum marks: 3

3(a) Exploration and Exploitation in EA

What do we mean by the terms “exploration” and “exploitation” in EAs (and other search methods)?

Fill in your answer here

Maximum marks: 3

3(b) Exploration and Exploitation in EAs

Name a way to change an EA so that it explores more

Fill in your answer here

Maximum marks: 3

3(c) Exploration and Exploitation in EAs

What do we mean by selection pressure in an EA? How is selection pressure related to exploration/exploitation?

Fill in your answer here

Maximum marks: 3

4 k Nearest Neighbors (kNN)

Describe shortly the main steps in the *k* nearest neighbors algorithm for supervised classification. (Answer 1-2 sentences for each step.)

Fill in your answer here

Maximum marks: 7

5(a) Linear regression predict

We consider a model for linear regression. We have trained a model with the weights $\mathbf{w} = (w_0, w_1, w_2) = (0, -1, 1)$. Each observation **x** has two features x_1 and x_2. In addition, we have added one bias feature in the first position, x_0=-1. Which value will this model predict for the observation $\mathbf{p} = (p_0, p_1, p_2) = (-1, 3, 1)$?

Fill in your answer here

Maximum marks: 3

5(b) Linear regression training

We assume that the model is trained sequentially by gradient descent and that it is updated after each observation. The target value for **p** is 0. What are the new weights after being trained on **p**? Assume a learning rate of 0.1. Make additional assumptions if needed, and state them clearly. In particular, which loss function do you use?

Fill in your answer here

Maximum marks: 5

6(a) Logistic regression

Consider a logistic regression model for classification, where the positive class is given by 1 and the negative class by 0. We have trained a model with the weights $\mathbf{w} = (w_0, w_1, w_2) = (0, -1, 1)$. Each observation **x** has two features, x_1 and x_2. In addition we have added a bias feature in the first position, x_0=-1. Which class will this model predict for the observation $\mathbf{p} = (p_0, p_1, p_2) = (-1, 3, 1)$? State reasons for your answer.

Fill in your answer here

Maximum marks: 5

6(b) Logistic regression training

We assume that the model is trained sequentially by gradient descent and that it is updated after each observation. The target value for **p** is 0. What are the new weights after being trained on **p**? Assume a learning rate of 0.1. Make additional assumptions when needed and state them clearly. What type of loss function gives this update? (Only the name of the function, not the formula).

Fill in your answer here

Maximum marks: 4

7(a) Activation in hidden layer

Which sort of problems could we have solved with a multi-layer perceptron (MLP) if we did not include an activation function in the hidden layer?

Fill in your answer here

Maximum marks: 3

7(b) Step-function

Why don't we apply a step function as the activation function in the hidden layer of a multi-layer perceptron (MLP)?

Fill in your answer here

Maximum marks: 3

7(c) Batch vs SGD

When training a classifier by gradient descent, we distinguish between batch training and stochastic gradient descent. Explain the differences between the two. Explain also what we mean by mini-batch training.

Fill in your answer here

Maximum marks: 6

8 Random forest

According to Marsland, the main idea behind the random forest classifiers "is largely that if one tree is good, then many trees (a forest) should be better, provided that there is enough variety between them."

Which methods can be applied to provide variety between the trees?

Fill in your answer here

Maximum marks: 7

9 Recurrent neural networks

In Marsland's book (p. 71), we can read that recurrent neural networks (RNNs) "have been studied, but are not commonly used." This has changed a lot since the book was published. What is the main difference between recurrent neural networks and feed-forward networks (MLPs)? For which kind of problems are RNNs particularly suited?

Fill in your answer here

Maximum marks: 7

10

State/action	Q-value
State S1, action A1	0.4
State S1, action A2	0.4
State S2, action A1	0.0
State S2, action A2	0.8

(a) Learning Value Functions

For the next questions, consider a reinforement learning problem with two states and two possible actions. We are trying to learn a table of Q-values representing the utility of performing each action in each state. At some point in our learning, the table looks like shown on the left.

An alternative to learning Q-functions is to directly learn estimated values (V) for each state. Referring to the Q-table on the left, name one disadvantage of doing this.

Fill in your answer here

Maximum marks: 3

(b) V-values - Advantages

In the last task, you suggested a disadvantage to estimating state-values (V). What may be the *advantage* of estimating state-values (V) instead of values for state/action pairs (Q)?

Note, this is a general question on Reinforcement Learning. Here, you do not have to refer to the given Q-table.

Fill in your answer here

Maximum marks: 3

(c) Q-learning

Assume we are in state S1 and perform action A2. We observe a reward of 1 and notice that we next find ourselves in state S2.

Calculate the updated Q-value Q(S1,A2) assuming a discount factor of 0.9 and learning rate of 0.1. The update function for q-learning is given below, taken from the course textbook (Marsland). Show all steps of your calculation.

$$Q(s, a) \leftarrow Q(s, a) + \mu(r + \gamma \max_{a'} Q(s', a') - Q(s, a))$$

Fill in your answer here

Maximum marks: 3

(d) Reinforcement Learning Policies

Referring to the words exploration and exploitation, describe what is the difference between a greedy policy and an epsilon-greedy policy.

Fill in your answer here

Maximum marks: 3

11 Developmental Systems

How can developmental systems generate complex patterns from simple rules? Give a simple example, referring to one of the developmental systems discussed in this course's lecture.

Fill in your answer here

Maximum marks: 5

12 Hierarchical Clustering

We have performed 20 different blood tests on each of *n* patients and have collected all the results in a 20 x n matrix **X** such that **X**[i,j] is the result of the i'th blood test performed on the j'th patient.

We wish to see if the patients cluster into subgroups based on the blood test results, and decide to use hierarchical clustering. **What distance measure and linkage method would you choose in each of these cases?**

Case 1: We wish to cluster together patients with quite similar blood test results (i.e. the distance between two patients is small if their results for each blood test are numerically quite similar). Furthermore, every element of a cluster should have a small distance to every other element of the same cluster.

Case 2: We wish to cluster together patients with correlated blood test results (i.e. the distance between two patients is small if their results are low on the same blood tests and high on the same blood tests). Furthermore, each element of a cluster should have a small distance to at least one other element of the same cluster.

Fill in your answer here

Maximum marks: 6

13 Future Perspectives: Causality

a. What do we mean when we say machine learning models are good at learning correlations, but not causal relationships?
b. Give an example of a situation where a learned correlation is not enough to make a good decision, but where we also need information about causal relationships.

Fill in your answer here

Maximum marks: 6