Exam IN4050 Spring 2021

1 Search

1 Search

Name 2 differences between *running 20 different hillclimbing algorithms with different starting points* and *an Evolutionary Algorithm with a population of 20*. Assume we are working on a binary optimization problem (trying to find the best combination of 1's and 0's in a string of length N), and the EA applies:

-One-point crossover

-A mutation rate of 0.1

-Fitness-proportionate selection for parent selection

-Tournament selection for survivor selection

These EA settings may not all be of relevance to your answer, but you may mention it if any of them are relevant to the difference.

The two algorithms have the same starting and termination conditions. Explain any other assumptions you have to make about the two algorithms. Explain/justify how each of the two differences affect the search.

Maximum marks: 6

2 Evolutionary Algorithms

2

Individual number	Genotype	Fitness
1	0 1 1 1 1	1
2	0 0 0 0 1	2
3	1 0 0 1 1	3

(a) Selection in Evolutionary Algorithms - Fitness Proportional

Consider the individuals listed in the table on the left, from an evolutionary algorithm. Assume this is the full population, and you are about to do parent selection from it. Show your calculations and reasoning for the problems that follow.

What is the selection probability of each individual given *Fitness Proportional selection*? Note that the probability you calculate *is for a single selection*, not the probability that the individual gets selected into the pool of parents after N different selections.

Maximum marks: 2

(b) Selection in Evolutionary Algorithms - Fitness Sharing

Assume we want to do fitness sharing to reduce the potential for premature convergence. We will use a simple distance measure that simply counts the number of genes in two genotypes that are in the same location but have different values:

$$d(g1, g2) = \sum_{i=1}^{5}(abs(g1[i] - g2[i])$$

where g1 and g2 are the two genotypes being compared, g1[i] denotes the i'th gene of g1, and abs() indicates the absolute difference between the two elements. We use a niche size σ_{share} of 3, and set α=1. What is the modified fitness value of each individual f'(i) assuming the original fitness value was the one you calculated with Fitness Proportional selection? (6p)

Maximum marks: 6

(c) Selection in Evolutionary Algorithms - Selection Pressure

Referring to the way fitness values changed as you performed the fitness sharing, describe what happened to the *selection pressure* in this EA (and *why* this happened).

Maximum marks: 4

3 Features

3

	spam	chars	lines breaks	'dollar' occurs. numbers	'winner' occurs?	format	number
1	no	21,705	551	0	no	html	small
2	no	7,011	183	0	no	html	big
3	yes	631	28	0	no	text	none
4	no	2,454	61	0	no	text	small
5	no	41,623	1088	9	no	html	small
...							

The table shows the first entries from a data set of e-mails. The data set will be used for recognizing spam mails. In addition to the label, 'spam'/'no spam', 6 attributes are extracted from each e-mail. The attribute 'format' has two possible values: 'text' and 'html', the attribute 'numbers' has three possible values, 'small', 'big', and 'none', while ' "winner" occurs' can take the values 'yes' or 'no'.

(a) Categorical and numerical features

Explain in 2-4 sentences what is meant by categorical features and numerical features in machine learning. You may use the data set as example.

Maximum marks: 3

(b) Classifiers and features

We have considered several kinds of classifiers including

- Decision Trees
- *k*NN
- Perceptron
- Logistic regression
- Feed-forward neural networks (Multi-layer perceptrons)

Some of them use categorical features while others use numerical features. Which of them can handle categorical features without further preprocessing the data, and which of them can handle numerical features without preprocessing? You do not have to justify your answers.

Maximum marks: 3

(c) Preprocessing categorical attributes

Suppose you will apply a classifier which expects numerical features to a data set where some attributes are categorical. How can this data set be preprocessed to fit the classifier? You may use the spam data set as example.

Maximum marks: 3

(d) Further preprocessing

Suppose you are to prepare the spam data set for a classifier taking numerical features. Are there other preprocessing steps you would take? Explain shortly which steps you would take and why.

Maximum marks: 3

(e) Preprocessing numerical data

Suppose instead you will apply a classifier which expects categorical features to a data set where some attributes are numerical. How can this data set be preprocessed to fit the classifier? You may use the spam data set as example.

Maximum marks: 3

4 Perceptron

4(a) Classifying

We will use a perceptron for classification. There are two classes, 0 and 1. The data points have two features, e.g., the point $\mathbf{x}_1 = (x_{1,1}, x_{1,2}) = (1, 1)$. We add a bias term $x_{j,0} = -1$ to each data point $\mathbf{x}_j$. Assume the current weights are $\mathbf{w} = (w_0, w_1, w_2) = (0.1, 0.1, 0.2)$. How will the perceptron classify $\mathbf{x}_1$?

Maximum marks: 4

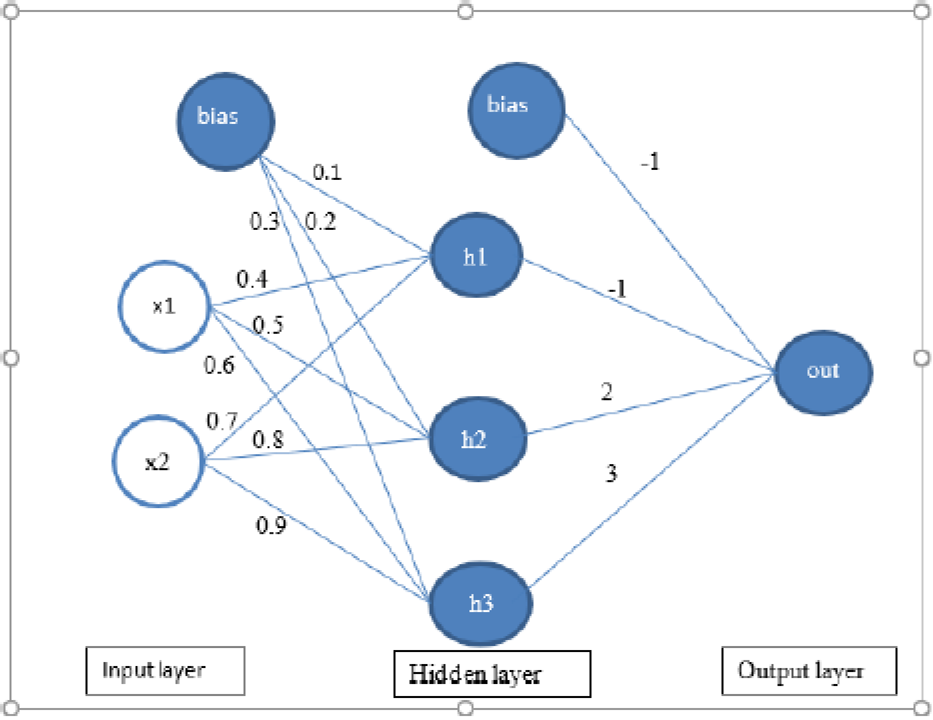
4(b) Training

Assume that the correct class for $\mathbf{x}_1 = (1, 1)$ is $t_1 = 0$. We are training the perceptron sequentially. How will the weights $\mathbf{w} = (w_0, w_1, w_2) = (0.1, 0.1, 0.2)$ be updated from observing $(\mathbf{x}_1, t_1)$? Assume a learning rate of 0.2.

Maximum marks: 4

5 Neural Networks and Backpropagation

5



The neural network.

(a) Value 1

Consider the network in the figure. Assume that the bias terms are set to -1 for both layers. Moreover, we assume that the logistic function (sigmoid) is used as activation in the hidden layer, and that the learning rate is 0.1. When the input is $(x_1, x_2) = (2, -1)$, what is the output from the hidden node h3?

Maximum marks: 4

(b) Value 2

Assume that the network is used for regression with no activation function in the output layer. What is the final output when the input is $(x_1, x_2) = (2, -1)$?

Maximum marks: 4

(c) Learning

Assume that the correct prediction for $(x_1, x_2) = (2, -1)$ is $t = 5$. We are training the network sequentially. What will the updated weight from the hidden node h3 to the out-node be after backpropagation?

Maximum marks: 5

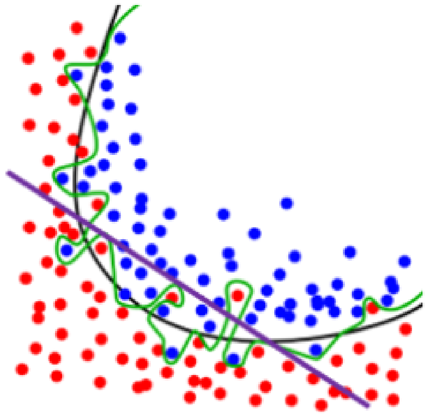
(d) Backpropagation

What will the updated weight from the x2-node to the hidden node h3 be after backpropagation?

Maximum marks: 5

6 Overfitting and Bias

6



Figure

(a) Overfitting

What do we mean by overfitting in machine learning and why can it be a problem? You may refer to the figure for an example.

Maximum marks: 3

(b) Inspection

Suppose you are training a supervised classifier and you suspect it to overfit. How would you go about to find out whether it is overfit?

Maximum marks: 3

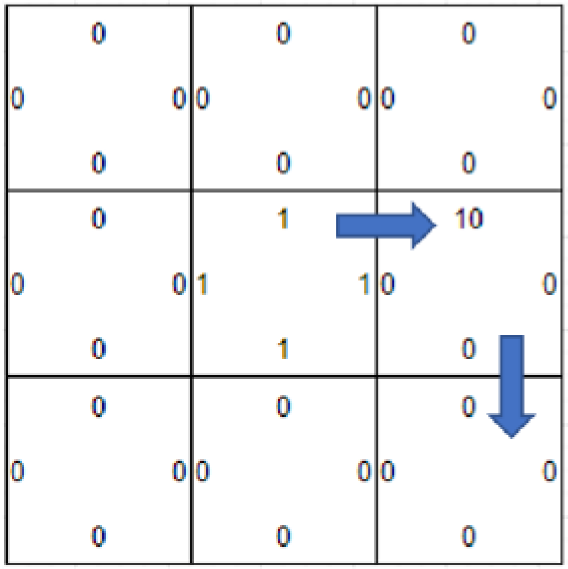
(c) Inductive bias

The percetron algorithm and logistic regression have a similar inductive bias. What is it? Use it to describe what is meant by inductive bias.

Maximum marks: 3

7 Reinforcement Learning

7



(a) Reinforcement Learning - Q learning

Consider the Reinforcement Learning grid world on the left. Each square is a state, and numbers represent Q-values associated with moving to the neighbor state in the given direction. As normal, the Q-values are estimates, that we want to update by exploring the environment.

We initialize an agent in the center state, and let it explore according to its policy. In this world, **every move from one state to another yields a reward of 1**. We want to train the agent with a **learning rate μ=0.1** and **discount factor** γ**=0.1**.

Using its policy, the agent performs first the action "right", then the action "down", as shown in the figure.

Calculate the updated Q-value (using the formula from the lecture slides or text book) for the center state, that is Q(center state, right), as indicated by the top arrow. You should calculate the updated Q-value according to the (off-policy) Q-learning algorithm.

Maximum marks: 4

(b) RL - SARSA

Assuming we again start from the same initial state and Q-values, as in the previous task, and perform the same actions, calculate the updated Q-value for state-action pair (center, right) according to the (on-policy) SARSA algorithm.

Like before, every move from one state to another yields a reward of 1. We again want to train the agent with a **learning rate μ=0.1** and **discount factor** γ**=0.1**.

Maximum marks: 4

(c) Reinforcement Learning Q vs SARSA

With reference to the two previous tasks, explain why the updated Q-values from Q-learning and SARSA are different: Why does on-policy and off-policy learning give different results here?

Maximum marks: 4

(d) Reinforcement Learning - Policy Why

Consider again the grid world on the left.

Epsilon-greedy and softmax are two different policies that aim to balance exploitation and exploration. Name an advantage of using softmax over epsilon-greedy.

Suppose we are considering to apply epsilon greedy with epsilon equal to 0.1, giving a quite small chance of choosing a non-optimal action. Why will softmax *still* give a much smaller probability than epsilon-greedy of choosing the "down" action in the right state (the one where the downward facing arrow begins)? You don't have to calculate the probability, but give a reason referring to how softmax and epsilon greedy calculate their probabilities.

Maximum marks: 5

8 Unsupervised Learning

8(a) Unsupervised Learning - Making Data

Assume you have a collection of data on two different animals, cats and dogs, measuring 1) their weight (between 0 and 100 kg), and 2) their train-ability (how easy it is to train the animals – ranging from 0, impossible to train to 100, very easy to train). You want to try to use a K-means classifier to discriminate between cats and dogs based on weight and trainability.

You will here make up some data on the cats and dogs, and in the following question use this data to explain how K-means clustering can proceed to divide your data into two parts.
The data you make up should be submitted as drawings (plots with x- and y-axes). It is not necessary to make up the exact numbers/coordinates for your datapoints - hand-drawn plots with approximate data positions are enough. You should include at least 4 datapoints representing cats and 4 datapoints representing dogs.

a) Make data (including values for both features and the correct label) that k-means can perfectly divide into 2 clusters, separating cats and dogs. (2p)
b) Make a new set of data that k-means *can not* perfectly divide into two clusters, separating cats and dogs. (2p)

Maximum marks: 4

8(b) Unsupervised Learning - Forming clusters

Referring to the cat and dog data you made up (the data you think *can* be separated), describe how the k-means algorithm forms clusters. You do not have to show all calculations, but explain how (and why) the cluster centers move around, and how (and why) the clustering changes as you iterate.

Maximum marks: 2

8(c) Unsupervised Learning - Why

Referring to the two datasets you made up about cats and dogs, why did k-means succeed on one dataset, but not the other?

Maximum marks: 3

9 L-systems

9 L-systems

Bracketed L-Systems are useful when visualizing the resulting string of an L-System. Consider the following string resulting from a bracketed L-System: F[+F][-F].

Assume 'F' is interpreted as a draw command. The symbols '+' and '-' are interpreted as 'turn' being either a positive or negative turn of 45 degrees respectively. The brackets '[' and ']' are used for storing and retrieving the position and orientation (the state) when drawing the L-System. Here the state is effectively pushed and popped on a stack.
Can you draw the resulting tree from this string? Indicate the position where you started drawing by adding an S to the starting point. (A positive turn command assumes rotation in a counter-clockwise direction)

You should deliver three separate drawings below:

a) The resulting tree after the first draw command (2p)

b) The resulting tree after the first and second draw command (2p)

c) The final drawing (2p)

Maximum marks: 6