

Exam in Machine Learning
DD2431

2008-12-18, kl 14.00 – 19.00

Aids allowed: *calculator, language dictionary.*

A Questions for pass or fail

Note: To pass the exam you must give the correct answer on almost *all* questions in this section. Only *one* error will be accepted, so be very careful not to make any unnecessary mistakes here.

A-1 Hypotheses Space

For a *decision tree* learning algorithm, what does the Hypotheses Space (H) contain?

- a) All nodes in the tree
- b) All possible trees
- c) All leaf nodes in the tree
- d) All combinations of the attributes
- e) All possible training patterns

b) All possible trees

A-2 Genetic Algorithms

Which three operations are central for a Genetic Algorithm?

- a) Injection — Increasing the chromosome size
- b) Selection — Preserving the best individuals
- c) Prevention — Preventing inconsistent combinations
- d) Crossover — Combining parts of good individuals
- e) Mutation — Random changes
- f) Recycling — Reusing bad combinations

Note: Answer with *all three* correct items.

b Selection, d Crossover and e Mutation

A-3 MAP and ML

Which one of these statements about MAP and ML hypotheses is correct (choose exactly one answer)? D are the observed data, and h are hypotheses.

- a) $h_{\text{MAP}} = \arg \max_i P(D|h_i)$ $h_{\text{ML}} = \arg \max_i P(h_i|D)$
- b) $h_{\text{MAP}} = \arg \max_i P(h_i|D)$ $h_{\text{ML}} = \arg \max_i P(D|h_i)$
- c) $h_{\text{MAP}} = \arg \max_i P(h_i, D)$ $h_{\text{ML}} = \arg \max_i P(D|h_i)$

b

A-4 Reinforcement Learning

In the standard setting for reinforcement learning, what does the term *policy function* refer to?

- a) A rule deciding which action to take for a given state
- b) A rule deciding which action to take for a given reward
- c) A rule deciding the value of a given state
- d) A rule deciding the value of a given state \times action pair
- e) A rule deciding the next state given the current state
- f) A rule deciding the next state given the current state \times action pair

a

A-5 Bias

What is restricted when an algorithm has *restriction bias*?

- a) The bias is minimized to improve generalization
- b) The order of the training patterns affect the result
- c) The number of training examples is limited
- d) The available hypotheses can not describe all possible classifications

d

A-6 Hypothesis Order

Which hypothesis is most *special*:

- a) Rainy
- b) Rainy \wedge Windy
- c) \star (always true)

b Rainy \wedge Windy

B Questions for higher grades

Preliminary number of points required for different grades:

$$20 \leq p \leq 24 \rightarrow A$$

$$16 \leq p < 20 \rightarrow B$$

$$12 \leq p < 16 \rightarrow C$$

$$6 \leq p < 12 \rightarrow D$$

$$0 \leq p < 6 \rightarrow E$$

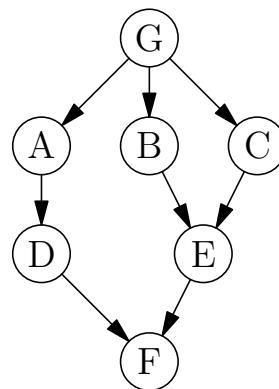
B-1 Hypothesis Order

(3p)

Consider a concept learning situation where the examples are individual persons. Given the following hypotheses:

- A) Women
- B) Men
- C) People living in Sweden
- D) Women older than 65
- E) Fredrik Reinfeldt¹
- F) Nobody
- G) Everyone

Draw a directed graph showing the partial order of these hypotheses (according to the general–special relationship). Explain for at least one of the edges what it means.



The arrow from *B* to *E* means that the concept of “being a man” is more general than the concept of “being Fredrik Reinfeldt”.

B-2 Genetic Algorithm

(3p)

You intend to use a *genetic algorithm* to train a feed-forward ANN. You intend to have a fixed network structure and only train the weight values. You have several hundred training examples on the form $x_1, x_2, x_3, x_4 \mapsto f$ where all values x_1, x_2, x_3, x_4 , and f are real numbers in the interval $[-1, 1]$.

Describe what the concepts used in the genetic algorithm corresponds to *in this specific example*. In particular, describe:

- a) What constitutes *individuals* and how are they represented in the form of *chromosomes*?
- b) How will you make a *crossover*?
- c) How will you define the *fitness function*?

- a) Individuals are different sets of weights. They can be represented simply as an array of numbers.
- b) Crossover is used to form new individuals, for example by picking $x_i, i \in \{1, 2, 3, 4\}$ from one or the other of two parent-individuals at random.
- c) Individuals are evaluated by testing the corresponding network on *all training examples* and giving it a fitness value depending on the total error.

¹Fredrik Reinfeldt is the swedish prime minister; he was born in 1965.

B-3 Christmas Worries

(3p)

Arne is worried that he may not get his Christmas gift this year. There are two independent reasons for this; one is that Santa Claus might think he has been too naughty this year. The second reason for his worry is that he may have forgotten to e-mail his wish-list to Santa.

Arne estimates that there is a 30% probability that he has been too naughty, and a 20% probability that he forgot to post the wish-list.

- a) How unpredictable is the situation that he does not get his Christmas gift, that is, that he has been too naughty or he forgot to post the wish-list (or both), calculated as an entropy measured in bits?

Notation: G — gift, W — forgot wish-list, N — naughty

$$\begin{aligned}P(G) &= P(\neg W) \cdot P(\neg N) = 0.8 \cdot 0.7 = 0.56 \\ \text{ent}(\neg G) &= -P(\neg G) \log_2(P(\neg G)) - P(G) \log_2(P(G)) = \\ &= -0.44 \log_2(0.44) - 0.56 \log_2(0.56) \approx 0.9896\end{aligned}$$

- b) Arne realizes that he can hack into the e-mail server of Santa Claus and check if he actually received his wish-list. What is the expected information gain from finding out if he did post the wish-list?

Treat the cases W and $\neg W$ separately.

$$\begin{aligned}P(W) &= 0.2 \\ \text{ent}_W(G) &= 0 \\ P(\neg W) &= 0.8 \\ \text{ent}_{\neg W}(G) &= -0.3 \log_2(0.3) - 0.7 \log_2(0.7) = 0.8813\end{aligned}$$

Expected entropy after measuring W :

$$P(W) \cdot 0 + P(\neg W) \cdot 0.8813 = 0.8 \cdot 0.8813$$

$$\text{Expected gain} = 0.9896 - 0.8 \cdot 0.8813 = 0.2846$$

B-4 VC-dimension

(3p)

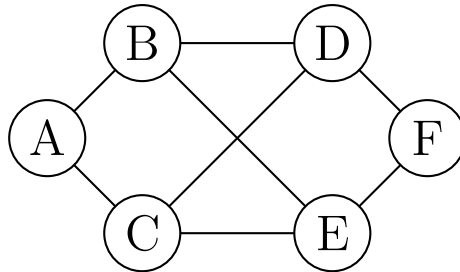
Consider a *concept learning task* where each example is a point in \mathcal{R}^2 (a two dimensional plane). What is the VC-dimension for a hypotheses space where *each hypothesis is a circle* where points inside the circle are considered part of the concept? The motivation for your answer is important.

Three points can be scattered (if placed in a triangle). Four points can not be scattered.
Answer: $VC(H) = 3$

B-5 Labyrinth

(3p)

Consider the following “labyrinth” where the labelled nodes denote positions. Position A is the *goal state* where you exit the labyrinth.



Given that each move (along any of the edges in the graph) gives a reward of -1 (i.e. a punishment), what is the value of being in each of the positions A, B, \dots, F when following an optimal policy? Use a discount factor (γ) of 0.9 and the normal definition of “value” used in reinforcement learning.

As usual, you must show how you arrived at your result.

$$\begin{aligned} V(A) &= 0 \\ V(B) = V(C) &= r + \gamma V(A) = -1 + 0.9 \cdot 0 = -1 \\ V(D) = V(E) &= r + \gamma V(B) = -1 + 0.9(-1) = -1.9 \\ V(F) &= r + \gamma V(D) = -1 + 0.9(-1.9) = -2.71 \end{aligned}$$

B-6 Support Vector Machines

(3p)

A support vector machine uses a minimization method for selecting some of the training data points to be *support vectors*.

- Explain what is special about these particular points.
- Draw a figure illustrating classification in 2D with the support vectors clearly marked.
- Is it possible for *all* support vectors to be only positive or only negative examples? Explain why/why not.

- Support vectors are the points closest to the decision boundary.
- All support vectors are on equal distance from the decision boundary.
- No, the margins have to be supported from both sides.

B-7 Bayesian Belief Network

(3p)

In Sweden, Santa Claus makes personal visits to all children on Christmas Eve. To minimize the number of unnecessary door-openings, the family X would like to have a probabilistic model of visitors' identity. More specifically, the X:s need a model, which at the event of a door-knock or door-bell-ring returns the probability that Santa Claus is at the door.

Build a *Bayesian Belief Network* from which this probability can be inferred. Specifically: Write down the variables involved, where all variables are binary with the state space [true, false]. Draw the graph, where nodes correspond to variables, and directed edges correspond to conditional dependencies between variables. For all nodes with parents, write down the conditional probability distribution associated with this node.

The following information is provided:

- Santa Claus never comes on any other day than Christmas Eve.
- Santa Claus will visit the X:s sometime during Christmas Eve, but 11 scouts selling Christmas candy will also come that day.
- When Daddy X leaves the house to buy a newspaper, the chance that Santa Claus will come increases by a factor 10.
- Daddy X goes out to buy newspapers regularly, independently of which date it is.
- Santa Claus always knocks, but scouts selling Christmas candy are as likely to use the door bell as to knock.
- The X:s have a very bad door bell — it is broken half the time.

- Variables:

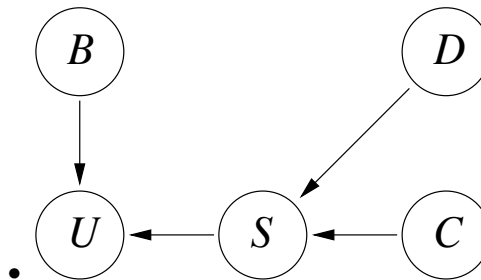
S = "The person at the door is Santa Claus (instead of a scout)"

C = "It is Christmas Eve"

D = "Daddy X is out to buy a newspaper"

U = "The door bell is used (instead of the door being knocked)"

B = "The door bell is broken"



- Conditional probabilities for S and U :

$$P(S|D, C) = 10/12$$

$$P(S|\neg D, C) = 1/12$$

$$P(S|D, \neg C) = 0$$

$$P(S|\neg D, \neg C) = 0$$

$$P(U|B, S) = 0$$

$$P(U|\neg B, S) = 0$$

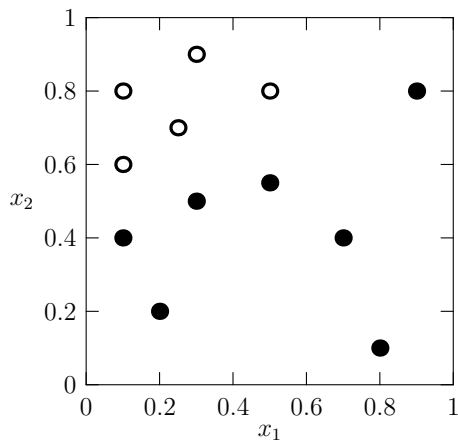
$$P(U|B, \neg S) = 0$$

$$P(U|\neg B, \neg S) = 1/2$$

B-8 Neural Network

(3p)

Manually calculate the weights and threshold value of a single layered neural network so that all the points in the figure are correctly classified. Filled circles should give the output 1, open circles should give 0. Clearly show how the output is computed from the input with a formula where your values are included.



Observation 1: There are three values to compute: two weights (one per input dimension) and one threshold.

Observation 2: The points can be separated by a line between the points $(0.0, 0.5)$ and $(1.0, 1.0)$.

Choose e.g. the weights $w_1 = 1$ and $w_2 = -2$ with threshold $\theta = -1$. The output is given from:

$$y(x_1, x_2) = \begin{cases} 1 & \text{when } x_1 - 2x_2 > -1 \\ 0 & \text{otherwise} \end{cases}$$

Good Luck!