

Exam in Machine Learning
Suggested solution
DD2431

2007-12-20, 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 *all* questions in this section. Be very careful not to make any unnecessary mistakes here.

A-1 Hypotheses Space

What is the size of the Hypotheses Space (H) for a single layered ANN with three inputs and two outputs?

- a) 2
- b) 3
- c) 8
- d) 2^6
- e) ∞

The weights are real numbers, hence: e) ∞

A-2 Genetic Algorithms

Which three operations are central for a Genetic Algorithm?

- a) Inversion — Reducing the population size with time
- b) Selection — Preserving the best individuals
- c) Restriction — Preventing inconsistent combinations
- d) Crossover — Combining parts of good individuals
- e) Mutation — Random changes
- f) Postponing — Avoiding changing improving individuals

Note: Answer with *all three* correct items.

b Selection, d Crossover and e Mutation

A-3 Bayes Classifier

A *naive Bayes classifier* is characterized by the following (choose exactly one answer):

- a) All class values are considered to have the same prior probability.
- b) All data attributes are considered conditionally independent given the class.
- c) One data attribute is selected randomly and used for classification.
- d) The classifier maintains a model of the joint likelihood of attributes given the class.

b

A-4 Reinforcement Learning

In the standard setting for reinforcement learning, what information does the *agent* receive from the environment in each time step?

- a) State
- b) Q-value
- c) Value
- d) Reward
- e) Policy
- f) Discount

Note: Answer with *all* correct alternatives.

a State and d Reward

A-5 Bias

An inductive bias preferring *simpler hypotheses* tends to

- a) Improve generalization for unseen examples
- b) Improve accuracy for training examples
- c) None of the above
- d) Both of the above

a (generalization improves, accuracy gets worse)

A-6 Hypothesis Order

Which hypothesis is most *general*:

- a) Rainy
- b) Rainy \wedge Windy

a Rainy

B Questions for higher grades

Preliminary number of points required for different grades:

$$\begin{aligned}0 \leq p < 6 &\rightarrow E \\6 \leq p < 12 &\rightarrow D \\12 \leq p < 16 &\rightarrow C \\16 \leq p < 20 &\rightarrow B \\20 \leq p \leq 24 &\rightarrow A\end{aligned}$$

B-1 MAP and ML

(3)

Describe the difference between Maximum A Posteriori (MAP) and Maximum Likelihood (ML) estimate. How are they defined? When are they used? Give an example (no formulas) where the MAP and ML estimates differ.

- – MAP: hypothesis with highest conditional probability given data: $h_{MAP} = \arg \max_i P(h_i|D)$
- ML: hypothesis with highest likelihood of generating the observed data: $h_{ML} = \arg \max_i P(D|h_i)$
- ML is used when all hypotheses have the same prior probability, MAP when they have different prior probability.
- Example: Any situation with two hypotheses h_1 and h_2 in which h_1 has a much higher probability than h_2 ; for example diagnosis of a very rare disease from tests with an error rate higher than the prior probability of the disease.

B-2 AdaBoost

(3)

Describe AdaBoost. What is the requirement on the used weak classifier to guarantee convergence?

- Different instances of the same classifier are iteratively trained with the data:
 1. The weak classifier is first trained with the original data (equal weights to all data points) to get the first instance of the weak classifier, C_1 .
 2. The current instance of the weak classifier, C_t , is used to classify the data points. A weight α_t is computed which is dependent on how well C_t does on the training data.
 3. The data points are reweighted so that the data examples that were wrongly classified by C_t get higher weights.
 4. The weak classifier is now trained again with the weighted data to get C_{t+1} .
 5. Repeat 2-4 until desired error level of combined classifier is reached, or a specified number of times T .

The combined classifier C^* is a linear combination of all weak classifiers C_t , weighted by α_t .

- The weak classifier must have a performance better than chance.

B-3 Candidate-Elimination

(3p)

CANDIDATE-ELIMINATION can be used to find the set of consecutive integers, consistent with the training examples. Each training example is a single integer, together with information about if it is in the interval or not.

Show, for each new training example, what the sets S and G will contain during training with these examples:

$$\langle 1, - \rangle, \langle 5, - \rangle, \langle 9, - \rangle, \langle 6, + \rangle$$

The examples are presented in this order (from left to right).

Initially: $G = \{[-\infty, \infty]\}$ $S = \{\}$
after $\langle 1, - \rangle$: $G = \{[-\infty, 0], [2, \infty]\}$ $S = \{\}$
after $\langle 5, - \rangle$: $G = \{[-\infty, 0], [2, 4], [6, \infty]\}$ $S = \{\}$
after $\langle 9, - \rangle$: $G = \{[-\infty, 0], [2, 4], [6, 8], [10, \infty]\}$ $S = \{\}$
after $\langle 6, + \rangle$: $G = \{[6, 8]\}$ $S = \{[6]\}$

B-4 Christmas Bells

(3)

For a computer application she is developing, Maria needs the sound of Christmas bells, but she is concerned about potential problems with copyright and prefers to use a synthetic sound rather than recording real bells. She come up with the idea of using a *genetic algorithm* to find a time series (a short wave file) which produces the most

beautiful Christmas bell sound. Marias idea is to let people visiting her web portal listen to two sound samples and click on the one that sounds best.

Describe how you would implement Marias idea. In particular, describe:

- a) What constitutes individuals and how are they represented in the form of chromosomes?
- b) How will you generate new individuals?
- c) How will you decide which individuals to preserve?

- a) Individuals are different sounds, for example represented as a timeseries (sequence of numbers) or amplitudes of different frequencies.
- b) New sounds are created from old ones by crossing (i.e. mixing parts) or mutation.
- c) Individuals are evaluated by using without getting an explicit fitness value. Therefore it is natural to use tournament selection where random sounds from the population are drawn and the winning one (from the user judgement) is preserved.

B-5 Restaurant food

(3)

Arne is somewhat picky with what he eats and often finds the food they serve at the local restaurant un-eatable. He has noted that there are two independent reasons for this: sometimes the food is too cold and sometimes it contains too much salt. Arne estimates the probability that the food is too cold to be 20%, and the probability that it is too salt to 30%.

- a) How unpredictable is the situation that the food is un-eatable, i.e. too cold or too salt (or both), calculated as an entropy measured in bits?

Notation: E — eatable, C — too cold, S — too salt

$$P(E) = P(\neg C) \cdot P(\neg S) = 0.8 \cdot 0.7 = 0.56$$

$$\text{ent}(\neg E) = -P(\neg E) \log_2(P(\neg E)) - P(E) \log_2(P(E)) =$$

$$= -0.44 \log_2(0.44) - 0.56 \log_2(0.56) \approx 0.9896$$

- b) Arne has figured out a (somewhat risky) way of checking the temperature of the food before ordering. What is the expected information gain from finding out if the food is too cold?

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

$$P(C) = 0.2$$

$$\text{ent}_C(E) = 0$$

$$P(\neg C) = 0.8$$

$$\text{ent}_{\neg C}(E) = -0.3 \log_2(0.3) - 0.7 \log_2(0.7) = 0.8813$$

Expected entropy after measuring C :

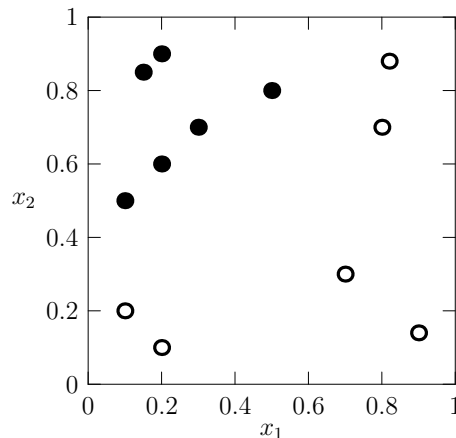
$$P(C) \cdot 0 + P(\neg C) \cdot 0.8813 = 0.8 \cdot 0.8813$$

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

B-6 Neural Network

(3)

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 $(0.0, 0.2)$ and $(0.8, 1.0)$.

Choose e.g. the weights $w_1 = -1$ and $w_2 = 1$ with threshold $\theta = 0.2$. The output is given from:

$$y(x_1, x_2) = \begin{cases} 1 & \text{when } -x_1 + x_2 > 0.2 \\ 0 & \text{otherwise} \end{cases}$$

B-7 VC-dimension

(3)

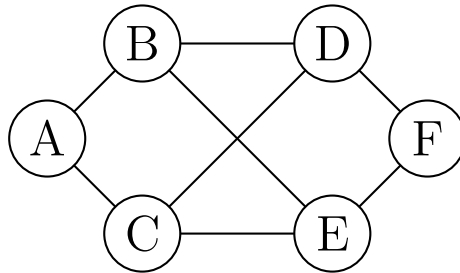
Consider a *concept learning task* where each example is a point in \mathcal{R}^2 (a two dimensional plane). What is the VC-dimension for a hypotheses space where *each hypothesis is a triangle* where points inside the triangle are considered part of the concept? The motivation for you answer is important.

Seven points can be scattered (if placed on a circle). Eight point can not be scattered (can't pick out four when there are four intermediate points).
Answer: $VC(H) = 7$

B-8 Labyrinth

(3)

Consider the following “labyrinth” where the labelled nodes denote positions. Position F 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.

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

Good Luck!