Artificial Intelligence

1. Define ‘artificial intelligence’, ‘machine learning’, ‘deep learning’

2. Compare the following three different tree-search strategies

   A. Exhaustive search

   B. Heuristic search

   C. Monte Carlo tree search

   D. Local search

3. Symbolize the followings using predicate logic.

   A. If city A is north of city B and city C is north of CityA, then city C is north of city B.

   B. No cat likes to swim.

   C. Some cats swim.

   D. Harry’s father likes wine.

4. Describe and compare the following two different approaches to artificial intelligence

   A. Top-down (knowledge-based)

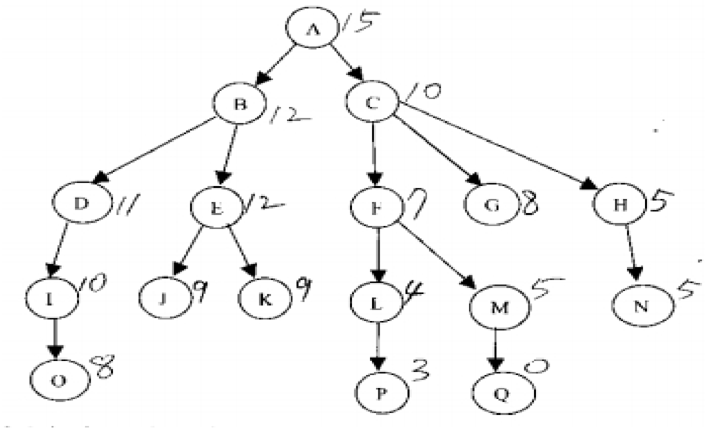
   B. Bottom-up (data-oriented)

5. Given the following search tree (st. the number next to each node represents cost of the node),

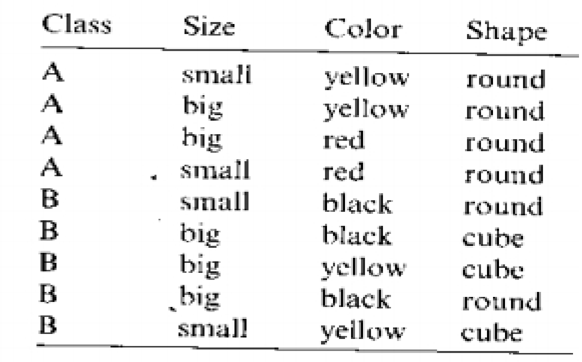
   A. Show the list of nodes visited, using breadth-first search

   B. Show the list of nodes visited, using hill-climbing search

   C. Show the list of nodes visited, using best-first-search

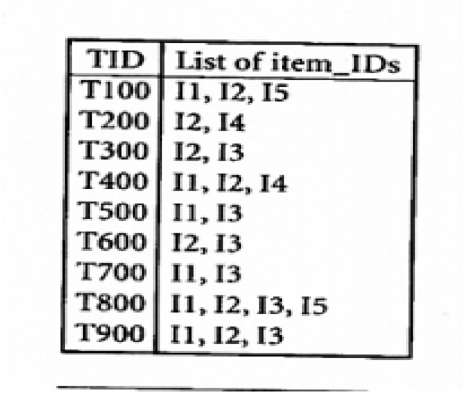


6. Build a decision tree (ID3 or C4.5) based on the following dataset.



Class	Size	Color	Shape
A	small	yellow	round
A	big	yellow	round
A	big	red	round
A	small	red	round
B	small	black	round
B	big	black	cube
B	big	yellow	cube
B	big	black	round
B	small	yellow	cube

7. Given the following transactional dataset, extract all possible association rules. (minimum support is 2)



TID	List of item_IDs
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

8. Explain the followings.

A. Definition of VC (Vapnik-Chervonenkis) dimension

B. The VC dimension of a linear classifier for n dimensional input space

9. According to the following classification rule, the class of the input x is predicted as the one with the maximum posterior probability.

$$y(x) = \underset{i}{\text{argmax}} P(\text{class} = i|x)$$

A. For binary classification problems, prove that the above classification rule is optimal. In what sense is it optimal?

B. Is th above classification rule still optimal for multiclass classification? Justify your answer.

10. Consider the following two-dimension samples of training data for naive Bayes classifiers.

x_1	x_2	class
5	4	1
6	5	1
6	5	2
6	6	1
6	6	2
6	6	2
7	6	1
7	6	2
7	6	2
8	7	2

A. Assume that x_1 and x_2 follows categorical distributions, and a naive Bayes classifier is trained using a maximum likelihood estimation method. Compute $p(\text{class} = 1|x_1 = 7, x_2 = 5)$ and $p(\text{class} = 2|x_1 = 7, x_2 = 5)$.

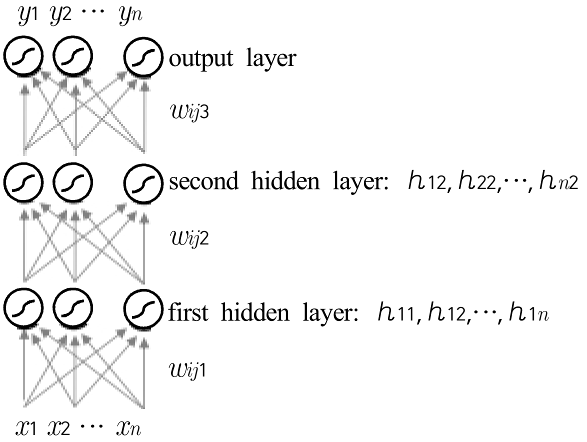
B. Assume that x_1 and x_2 follows Gaussian distributions, and a naive Bayes classifier is trained using a maximum likelihood estimation method. Compute the mean vector and covariance matrix for each class.

11. There are two coins in an urn. A random experiment of drawing a coin from the urn and tossing it is repeated three times with replacement of the coins. Suppose that the result of the random experiment is $D = \{H, H, T\}$. Run one iteration of the E-step and M-step of the expectation maximization (EM) algorithm using D as a training data set. What are the values of $p(H|C_1)$ and $p(C_1)$? Assume the following initial probabilities.

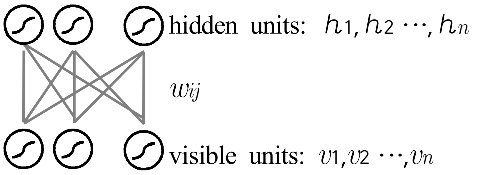
$p(H|C_1) = 0.6$ (probability of head when tossing coin C_1)
$p(H|C_2) = 0.4$ (probability of head when tossing coin C_2)
$p(C_1) = 0.5$ (probability of drawing coin C_1 from the urn)

12. Does the ID3 algorithm always produce a consistent decision tree for a training data set? Justify your answer.

13. Derive a weight update rule for the first hidden layer when the following artificial neural network is to be trained using the error back-propagation algorithm.


y1 y2 ⋯ yn
output layer
wij3
second hidden layer: h12, h22,⋯, hn2
wij2
first hidden layer: h11, h12,⋯, h1n
wij1
x1 x2 ⋯ xn

14. Consider the following restricted Boltzmann machine (RBM).

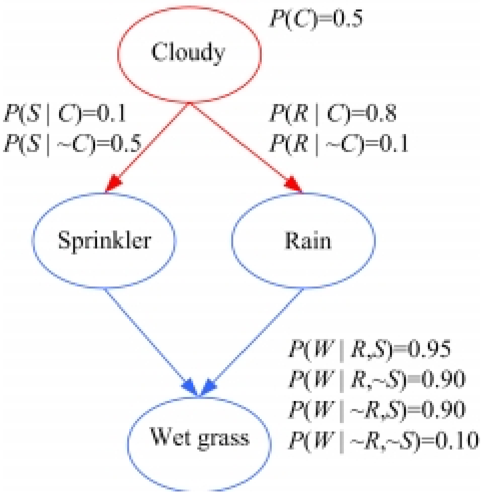


A. Derive a weight update rule when the objective function is maximum likelihood (ML).

B. Derive a weight update rule when the objective function is Kullback-Leibler (KL) divergence.

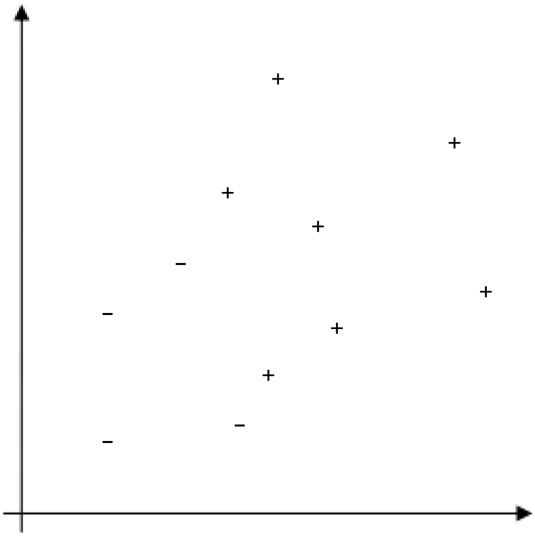
C. Derive a weight update rule when the objective function is contrastive divergence (CD).

15. Compute $p(C|W)$ in the following Bayesian belief network.



16. Find all support vectors in the following figure when C is equal to zero.

$$L_p = \frac{1}{2}\|w\|^2 + C\sum_t \xi^t - \sum_t \alpha^t[\gamma^t(w^T x^t + w_0) - 1 + \xi^t] - \sum_t \mu^t \xi^t$$



17. Define in your own words the following terms: state, state space, search technique, initial state, goal state, and objective function.

18. Describe why NLP is hard with your own example in terms of ambiguities in each phases of NLP.

19. Describe why knowledge representation and knowledge acquisition are called AI bottlenecks.

20. AI problems can be represented with state space, initial state, goal state, and objective function. Describe why search techniques are important AI techniques in terms of the representation of AI problem.

21. Consider the following sentences:

- John likes all kinds of food.

-Apples are food.

- Chicken is food.

-Anything anyone eats and isn't killed by is food.

- Bill eats peanuts and is still alive.

- Sue eats everthing Bill eats.

A. Translate these sentences into formulas in predicate logic.

B. Prove that John likes peanuts using backward chaining.

C. Convert the formulas of part **A** into clause form.

22. Design your document filtering system by using a statistical model.

23. Explain two major approaches to NLP and their pros and cons.

24. Explain statistical models for part-of-speech tagging, syntactic tagging, and machine translation by using noisy channel.

25. Statistical Part-of-Speech tagging can be defined as assigning an appropriate tag to each word of input sentence. Design an HMM based Part-of-Speech tagging model and explain the model.

26. Explain why search technique is one of important techniques in AI researches.

27. Explain the “Vanishing gradient problem” in deep neural networks and discuss possible solutions.

28. Discuss when bottlenecks are useful in deep neural networks with examples.

29. Define “skip connection” and discuss when it is beneficial. Show at least two architectures that have skip connections.

30. Explain why ReLU is less likely to suffer from the gradient vanishing problem than the sigmoid function.

31. Draw ReLU, LeakyReLU, and ELU with details such as intercepts and functions values. Point out one problem of ReLU and explain why the problem can be addressed by LeakyReLU or ELU.