1. [10%] General AI Knowledge

For each of the statements below, fill in the box T if the statement is always and unconditionally true, or fill in the box F if it is always false, sometimes false, or just does not make sense.

a) If A is one of B's k-nearest-neighbors for a given value of k, then B must be one of A's k-nearest-neighbors.

b) SVM can only classify data that is linearly separable.

c) Assuming Boolean attributes, the depth of a decision tree, built using common algorithms such as ID3 (Iterative Dichotomiser 3), can never be larger than the number of training examples.

d) Every Boolean function can be represented by some Bayesian network.

e) Naive Bayes is a linear classifier.

f) A Markov process is a random process in which the future is independent of the present, given the past.

g) A single perceptron cannot compute the XOR function.

h) For reinforcement learning, we need to know the transition probabilities between states before we start.

i) In supervised learning, the examples given to the learner are not labeled.

j) A perceptron is guaranteed to learn a given linearly separable function within a finite number of training steps.

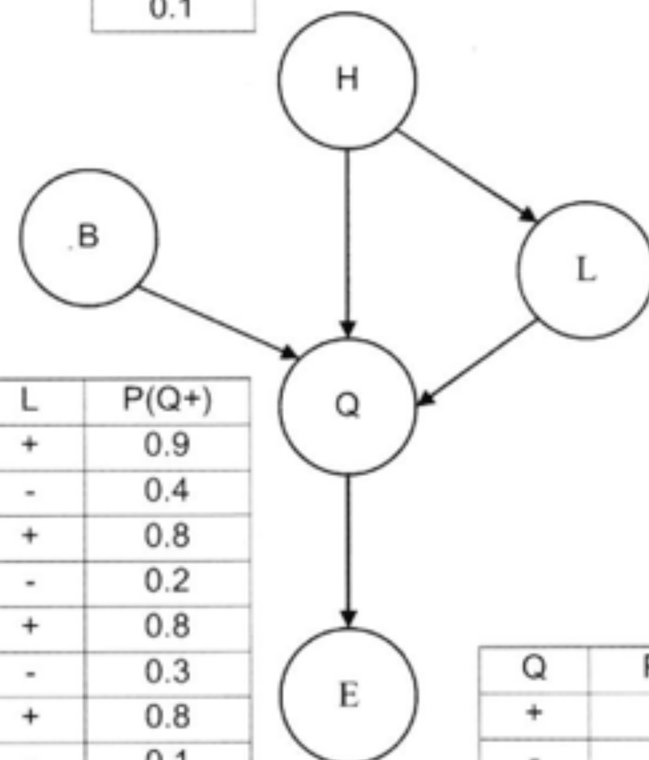
2. [20%] Bayesian Networks

In the network below, the Boolean variables have the semantics: B: Brilliant, H: Honest, L: LotsOfFriends, Q: Qualified, E: Elected.

P(H+)
0.1

P(B+)
0.5

H	P(L+)
+	0.3
-	0.9



B	H	L	P(Q+)
+	+	+	0.9
+	+	-	0.4
+	-	+	0.8
+	-	-	0.2
-	+	+	0.8
-	+	-	0.3
-	-	+	0.8
-	-	-	0.1

Q	P(E+)
+	0.6
-	0.1

2A. [6%] Which of these, if any, are asserted by the structure of the network (leaving aside the conditional probability tables (CPTs))?

1.	P(B, L) = P(B) P(L)
2.	P(E | Q, L) = P(E | Q, L, H)
3.	P(Q | B, H) = P(Q | B, H, L)

2B. [7%] Calculate the value of P(B+, H+, L-, Q+, E-). Show your work.

2C. [7%] Calculate the probability that a candidate is brilliant or not given that she is honest, does not have lots of friends, and gets elected. That is, calculate P(B | H+, L-, E+). Show your work. (You need to give both P(B+ | H+, L-, E+) and P(B- | H+, L-, E+))

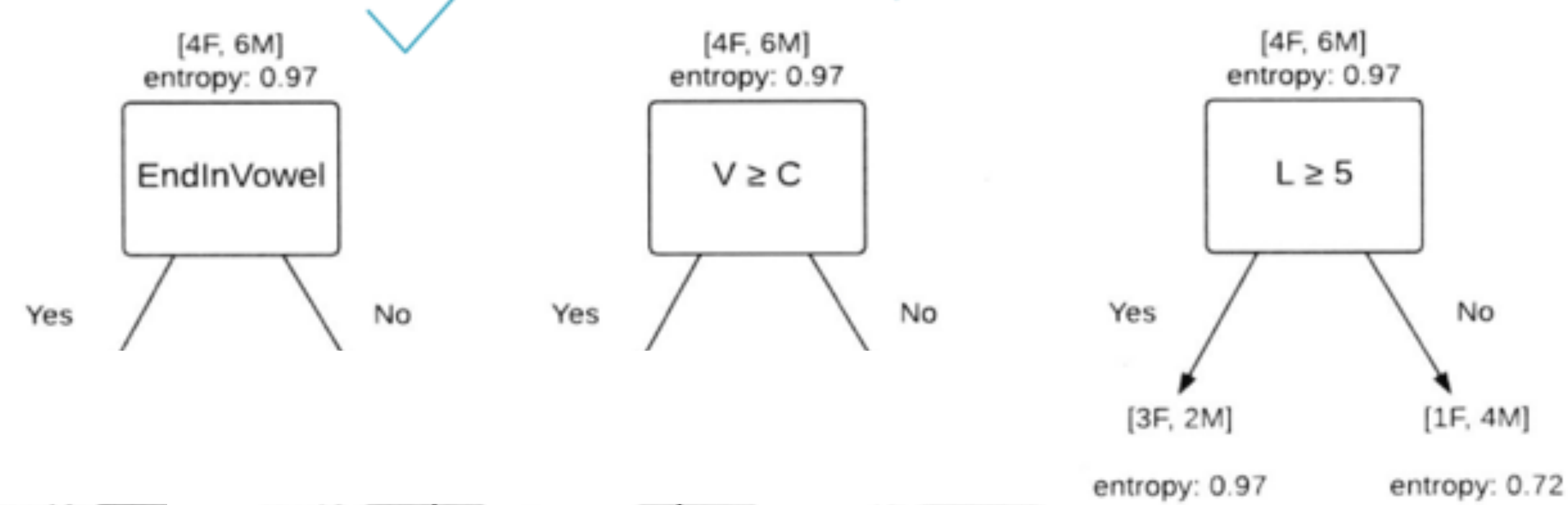
3. [23%] Decision Tree Learning

You are given the task of learning to classify first names by gender. You are given a list of names labeled as female (F) or male (M) and you want to learn a classifier based on decision tree learning.
For a given name, let us define **L** as its length, **V** as its number of vowels and **C** as its number of consonants.
We will consider that A-E-I-O-U-Y are vowels. The other letters are consonants.
You decide to use the following features to predict the classes:
- *EndInVowel*: The name ends in a vowel.
- *V ≥ C*: The name has more vowels than consonants.
- *L ≥ 5:* The name contains 5 letters or more.

Name	Feature			Gender
	EndInVowel	V ≥ C	L ≥ 5	
Annie	Yes	Yes	Yes	F
Brad	No	No	No	M
Carl	No	No	No	M
Daisy	Yes	Yes	Yes	F
Eleanor	No	Yes	Yes	F
Fernando	Yes	No	Yes	M
Gary	Yes	Yes	No	M
Hans	No	No	No	M
Isis	No	Yes	No	F
Jerry	Yes	No	Yes	M

With 4 Female names and 6 Male names, the entropy of the decision in bits is 0.97.

3A. [8%] Consider the following decision trees, splitting on (EndInVowel), (V ≥ C), (L ≥ 5). The (L ≥ 5) tree has been filled out. Complete the values for the other features, including entropy.



3B. [6%] Calculate the information gain for splitting on each of the 3 features. Show formulas and steps clearly.

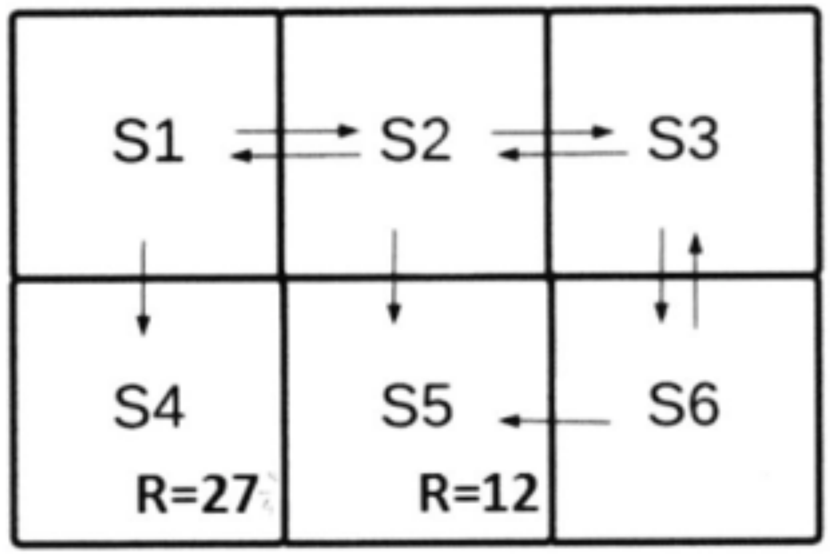
3C. [2%] Which attribute should you split on first? Justify your answer.

3D. [7%] For the second level of the tree, you decide to use the following rule:
- split on attribute ($V \geq C$) *if it was not split on first*
- split on attribute ($L \geq 5$) *otherwise*

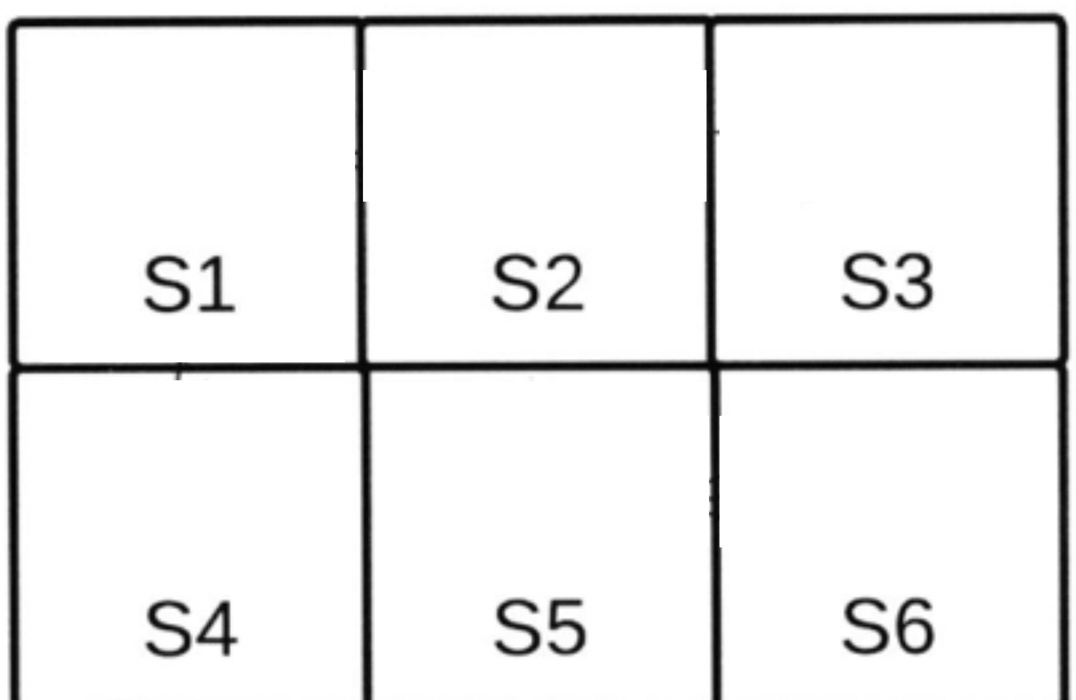
Draw the entire decision tree.

4. [17%] Markov Decision Process

Consider the 6-state Markov Decision process below. The goals with rewards are in state S4 and S5. At each state, the possible transitions are **deterministic** and indicated by the arrows. You get a reward of R_4=27 if you get to the goal S4 and a reward of R_5=12 if you get to the goal S5.

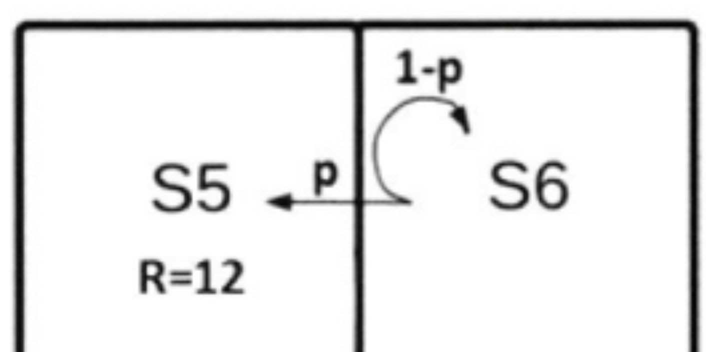


4A. [7%] Consider a discount factor of $\gamma = 2/3$. On the figure below, show the optimal value V* for each state and the arrows corresponding to the set of optimal actions.



4B. [5%] What values of γ would result in a different optimal action in S2? Indicate which policy action changes.

4C. [5%] In this question, you consider only states S5 and S6. The transition is no longer deterministic. When going to S5 from S6, you have a probability p of succeeding and a probability 1-p of tripping, and staying in state S6. What is the optimal value V* at state S6 if the discount factor $\gamma = 2/3$ and $p = 1/4$?



5. [20%] Neural Networks

5A. [4%] How many weights does a 2-layer feed-forward neural network with 5 input units, 3 hidden units and 2 output units contain, including the biases (dummy input weights)? Show your work.

5B. [4%] True or False.

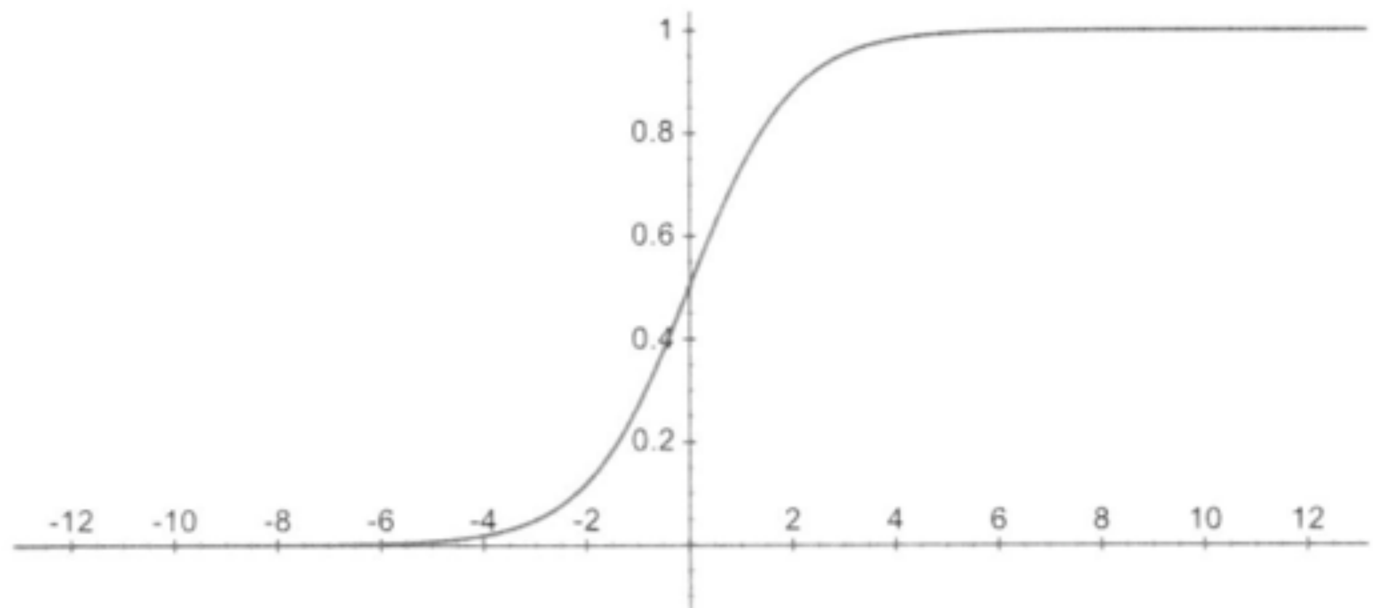
1. The back-propagation algorithm, when run until a minimum error is achieved, always converges to the same set of weights no matter what the initial set of weights is.
2. When choosing between two different neural network structures, we should always prefer the one with the lower error on the training set.

5C. [12%] Consider the neural network built out of units with real-valued inputs $X_1 \ldots X_n$, where the unit output Y is given by

$$Y = \frac{1}{1 + \exp(-(w_0 + \sum_i w_i X_i))}$$

Here we will explore the expressiveness of neural nets, by examining their ability to represent Boolean functions. Here the inputs X_i will be 0 or 1. The output Y will be real-valued, ranging anywhere between 0 and 1. We will interpret Y as a Boolean value by interpreting it to be a Boolean 1 if Y > 0.5, and interpreting it to be 0 otherwise.

The figure for $\frac{1}{1+e^{-x}}$ is:



Give 3 weights for a single unit with two inputs X_1 and X_2, that implements the logical OR function $Y = X_1 \vee X_2$ and the logical AND function $Y = X_1 \wedge X_2$, respectively.

Functions	w_0	w_1	w_2
Logical OR function $Y = X_1 \vee X_2$			
Logical AND function $Y = X_1 \wedge X_2$			

%] AI Applications.

1. [2%] Which statement is true about cognitive architectures?

a. A cognitive architecture is a hypothesis about the fixed structures that provide a mind.
b. A cognitive architecture tries to yield intelligent behavior in complex environments.
c. A generically cognitive architecture spans both the creation of artificial intelligence and the modeling of natural intelligence, at a suitable level of abstraction.
d. All of the above
e. None of the above

2. [2%] In the task of randomly assigning air marshals to flights using game theory, which argument allows us to use an incremental strategy for scaling-up?

a. The support set size is small: most variables are 0.
b. The full rewards matrix is sparse.
c. The computation can be parallelized.
d. All of the above
e. None of the above

3. [2%] Which method can be used to solve a problem in which the utility function is not known?

a. Reinforcement learning
b. Markov Decision Process
c. Perceptron learning
d. All of the above
e. None of the above

4. [2%] In Natural Language Processing, which of these algorithms takes advantage of grammars to represent sentences as trees?

a. Conditional Random Field (CRF)
b. Cocke-Younger-Kasami (CYK)
c. Hidden Markov Models (HMM)
d. All of the above
e. None of the above

5. [2%] In the minimax algorithm, which of the following is the most unrealistic in practice?

a. The knowledge of the utility values for the terminal states
b. The generation of the whole game tree
c. The assumption that the players are rational
d. All of the above
e. None of the above