Quiz5

1. Suppose you are a witness to a night-time hit-and-run accident involving a taxi in Athens. All taxis in Athens are blue or green. You swear, under oath, that the taxi was blue.

   Extensive testing shows that, under the dim lighting conditions, discrimination between blue and green is 75% reliable. You are told that 9 out of 10 Athenian taxis are green.

   What is the probability of the taxi being blue given that it looked blue?

   Hint: distinguish carefully between the proposition that the taxi *is* blue and the proposition that it *looks* blue.

   Answer: The relevant aspect of the world can be described by two random variables: B means the taxi was blue, and LB means the taxi looked blue. The information on the reliability of colour identification can be written as:

   $P(LB|B) = 0.75$ & $P(\neg LB|\neg B) = 0.75$

   We need to know the probability that the taxi was blue, given that it looked blue:

   $P(B|LB) \propto P(LB|B)P(B) \propto 0.75P(B)$
   $P(\neg B|LB) \propto P(LB|\neg B)P(\neg B) \propto 0.25(1 - P(B))$

   Thus, we cannot decide the probability without some information about the prior probability of blue taxis, P(B). For example, if we knew that all taxis were blue, i.e., P(B) = 1, then obviously P(B|LB) = 1. Given that 9 out of 10 taxis are green, and assuming the taxi in question is drawn randomly from the taxi population, we have P(B) = 0.1. Hence,

   $P(B|LB) \propto 0.75 \times 0.1 \propto 0.075$
   $P(\neg B|LB) \propto 0.25 \times 0.9 \propto 0.225$

   $P(\neg B|LB) = 0.225 / (0.075 + 0.225) = 0.75$
   $P(B|LB) = 0.075 / (0.075 + 0.225) = 0.25$

2. Identify the hypothesis and the evidence in the problem statement above.

   Answer:
   Hypothesis – taxi is blue
   Evidence – taxi looked blue

3. In a Markov chain, with the transition matrix A, what does the square of the transition matrix (A^2) represent?

   Answer: it is equivalent to the probability of taking two steps/transitions from each state to each other state and to itself as well.

4. What are the three problems that we work through in the context of HMM?

   Answer:
   - Observation likelihood
   - Decoding hidden state sequence
   - Learning hmm parameters

5. What is the time complexity of the brute force method?
   a. O(N)
   b. $O(N^2T)$
   c. $O(N^T)$
   d. O(NT)

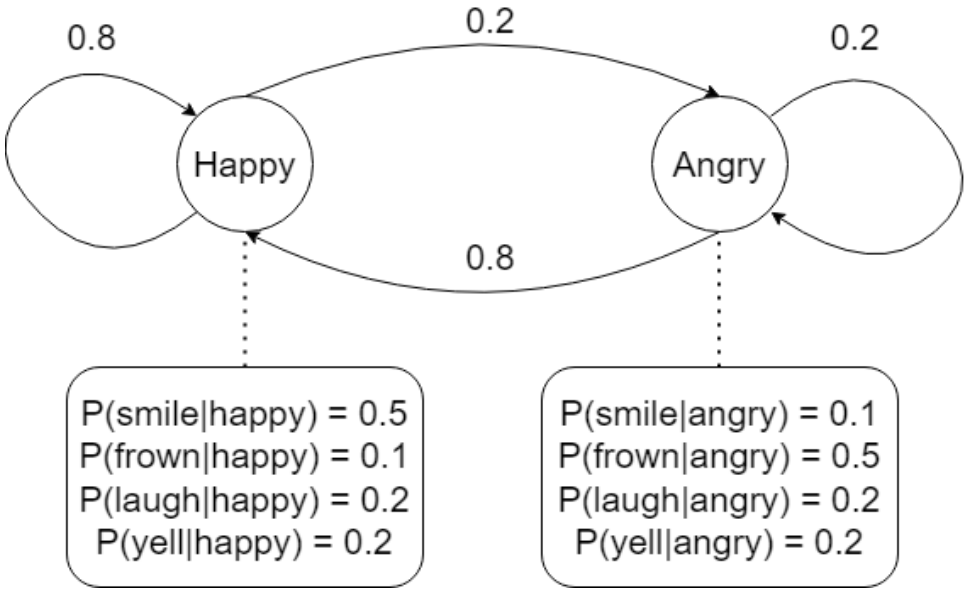
6. What is the time complexity of the forward trellis method/Viterbi algorithm?
   a. $O(N^T)$
   b. $O(N^2T)$
   c. O(NT)
   d. O(N)

7. Under what conditions is an event independent of another event? (Describe in terms of Bayes' theorem)

   Answer: if any of the following three conditions are true then an event A is independent of another event B:
   - $P(A \cap B) = P(A)P(B)$
   - $P(A|B) = P(A)$
   - $P(B|A) = P(B)$

8. You have a friend X who is happy some days and angry on other days. You can only observe when they smile, frown, laugh, or yell but not their actual emotional state.

   Let us start on day 1 in the happy state (there are no observations made on day 1). There can be only one state transition per day. It can be either to the happy state or the angry state. The HMM is shown below;



What is $P(o_2 = \text{frown})$? (Mention only the numeric value of the answer)

Answer:
$P(o_2 = \text{frown}) = P(o_2 = \text{frown}|s_2 = \text{Happy}) + P(o_2 = \text{frown}|s_2 = \text{Angry})$
$= P(\text{Happy}|\text{Happy}) * P(\text{frown}|\text{Happy}) + P(\text{Angry}|\text{Happy}) * P(\text{frown}|\text{Angry})$
$= (0.8 * 0.1) + (0.2 * 0.5)$
$= 0.08 + 0.1$
$= 0.18$

9. For the same HMM model, what is $P(s_2 = \text{Happy} \mid o_2 = \text{frown})$?

   Answer: This conditional probability cannot be calculated directly. Hence, we apply Bayes' rule to solve as follows;

   $P(s_2 = \text{Happy}|o_2 = \text{frown}) = (P(o_2 = \text{frown}|s_2 = \text{Happy}) * P(s_2 = \text{Happy})) / P(o_2 = \text{frown})$
   $= (P(\text{Happy}|\text{Happy}) * P(\text{frown}|\text{Happy})) / 0.18$

   [Note: $P(o_2 = \text{frown}) = 0.18$ from previous question]

   $= (0.8 * 0.1) / 0.18$
   $= 0.08 / 0.18$
   $= 0.4444$

10. Select the correct statement(s).
    a. XOR can be modelled using an SLP
    b. A hidden layer of MLP automatically learns new helpful features for the task
    c. A neuron computes a linear function followed by an activation function
    d. Weights can be initialized at zero and the learning will be correct

    Answer:

a. XOR is not linearly separable and hence cannot be modelled by an SLP
b. Each hidden layer aggregates learning of new features of the dataset
c. A neuron learns the linear relationship of input to output with the weights and then passes the predicted output through an activation function to gauge whether or not to fire the signal through to the next layer
d. Initializing all the weights with zeros leads the neurons to learn the same features during training.
In fact, any constant initialization scheme will perform very poorly. Consider a neural network with two hidden units, and assume we initialize all the biases to 0 and the weights with some constant a. If we forward propagate an input (x_1, x_2) in this network, the output of both hidden units will be sigmoid($ax_1 + ax_2$). Both hidden units will have identical influence on the cost, which will lead to identical gradients. Thus, both neurons will evolve symmetrically throughout training, effectively preventing different neurons from learning different things.

11. Can XNOR gate be modelled using a single layer perceptron?
    a. Yes
    b. No

    Answer: XOR is not linearly separable and hence cannot be modelled by an SLP

12. Write the optimizing process of an MLP model from the beginning in 5 steps.

    Answer:
    a. Initialize weights and bias
    b. Pass inputs through perceptron
    c. Calculate error
    d. Update weights through gradient decent and repeat till convergence

13. The key difference between a single-layer perceptron (SLP) and multi−layer perceptron (MLP) is:
    a. SLPs only work with linearly separable data, whereas MLPs can be trained to solve non−linearly separable problems
    b. SLPs use threshold activation functions whereas MLPs use linear activation functions
    c. SLPs can be trained to solve character recognition problems, whereas MLPs can be trained to solve handwriting recognition problems

    Answer:
    a. A single perceptron fails to solve the problem which is linearly inseparable and is capable of outputting a linear equation in the form of a model. So, to solve a non-linear problem, we add multiple perceptrons to the network.
    b. Both can use any type of activation function
    c. SLPs would not be able to model character regonition since the data is not linearly separable

14. Let f be some function so that $f(w_0, w_1)$ outputs a number. For this problem, f is some arbitrary/unknown smooth function that estimates the cost.
Suppose we use gradient descent to try to minimize $f(w_0, w_1)$ as a function of w_0 and w_1.
Which of the following statements are true? (Check all that apply.)
    a. If the first few iterations of gradient descent cause $f(w_0, w_1)$ to increase rather than decrease, then the most likely cause is that we have set the learning rate to too large a value
    b. No matter how w_0 and w_1 are initialized, so long as learning rate is sufficiently small, we can safely expect gradient descent to converge to the same solution
    c. If the learning rate is too small, then gradient descent may take a very long time to converge
    d. If w_0 and w_1 are initialized at a local minimum, then one iteration will not change their values

    Answer:
    a. If alpha were small enough, then gradient descent should always successfully take a small downhill move and decrease f(w0, w1) at least a little bit. If gradient descent instead increases the objective value, that means alpha is too large (or you have a bug in your code!).
    b. This is not true, depending on the initial condition, gradient descent may end up at different local optima.
    c. If the learning rate is small, gradient descent ends up taking an extremely small step on each iteration, and therefor can take a long time to converge
    d. At a local minimum, the derivative (gradient) is zero, so gradient descent will not change the parameters.

15. Suppose that for some linear regression problem (say, predicting housing prices), we have some training set, and for our training set we managed to find some w_0, w_1 such that the cost function estimated by mean square errors, $J(w_0, w_1) = 0$.
Select all statements that must be true.
    a. We can perfectly predict the prices of even new houses that we have not yet seen
    b. For these values of w_0 and w_1 that satisfy $J(w_0, w_1) = 0$, we have that $y'(x^{(i)}) = y^{(i)}$ for every training example $(x^{(i)}, y^{(i)})$
    c. Our training set can be fit perfectly by a straight line, i.e., all of our training examples lie perfectly on some straight line
    d. For this to be true, we must have $w_0 = 0$ and $w_1 = 0$ so that $y'(x) = 0$

    Answer:
    a. Since the cost function evaluates to zero, we can be certain that the model is overfitting on the training set and is not a good generalisation of the entire dataset and hence will not be able to predict the values of new houses from the testing dataset correctly
    b. For the training data all predicted outputs are exactly equal to the target outputs which leads to the cost function being zero
    c. Again, the overfitting means that the training data all seem to lie on the best fit line that the model has learnt

d. If J(w0, w1) = 0 that means the line defined by the equation "y = w0 + w1x" perfectly fits all of our data. There's no particular reason to expect that the values of w0 and w1 that achieve this are both 0 (unless y(i) = 0 for all of our training examples).