Name and Student ID:

Machine Learning BLG527E, January 9, 2018, 120mins, Final Exam
Signature:
Duration: 120 minutes.
Closed books and notes. Write your answers neatly in the space provided for them. Write your name on each sheet. Good Luck!

Q1 (25)	Q2 (25)	Q3(25)	Q4(25)	TOTAL (100)

ANSWERS

QUESTIONS

Q1) [25pts]

Given an HMM $\lambda = (\pi, A, B)$ with state transition probability matrix A, emission probabilities B, initial state probabilities π, and two states and two symbols red and green,

$$\pi = [0.2\ 0.8]^T \qquad A = \begin{vmatrix} 0.8 & 0.2 \\ 0.9 & 0.1 \end{vmatrix} \qquad B = \begin{array}{cc} \text{red} & \text{green} \\ \end{array}$$

	red	green	
B=	0.7	0.3	State1
	0.4	0.6	State2

What is the Pr(O| λ) where O = {red,red }

Hint: The forward variables in an HMM are calculated as follows:

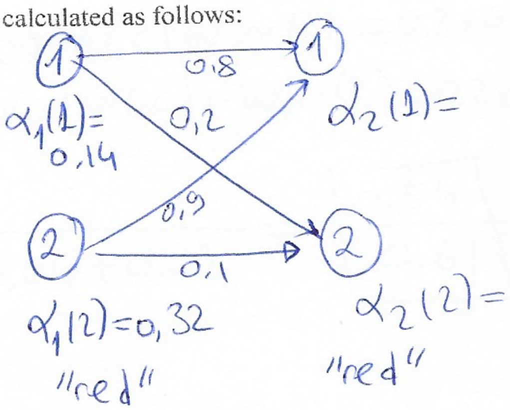
$$\alpha_t(i) \equiv P(O_1 \cdots O_t, q_t = S_i \mid \lambda)$$

Initialization:

$$\alpha_1(i) = \pi_i b_i(O_1)$$

Recursion :

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(O_{t+1})$$



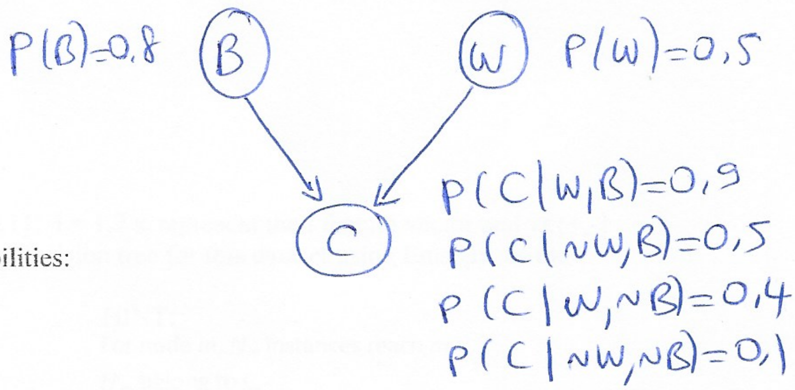
$\alpha_1(1) = \pi_1 b_1(\text{red}) = 0,2 \times 0,7 = 0,14$

$\alpha_1(2) = \pi_2 b_2(\text{red}) = 0,8 \times 0,4 = 0,32$

$\alpha_2(1) = (\alpha_1(1) a_{11} + \alpha_1(2) a_{21})\, b_1(\text{red})$
$= (0,14 \times 0,8 + 0,32 \times 0,9) \times 0,7 = 0,28$

$\alpha_2(2) = (\alpha_1(1) a_{12} + \alpha_1(2) . a_{22})\, b_2(\text{red})$
$= (0,14 \times 0,2 + 0,32 \times 0,1) \times 0,4 = 0,024$

$\Pr(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) = 0,28 + 0,024 = \boxed{0,304}$



Name and Student ID:

Q2) [25pts]

You are given the following probabilities:

P(B) = 0.8, P(W) = 0.5,

P(C|W,B) = 0.9, P(C|˜W,B) = 0.5,

P(C|W,˜B) = 0.4, P(C|˜W,˜B) = 0.1.

Given that the coffee you drink is good (C), compute the probability that the beans are good quality.

$$P(B|C) = \frac{P(C,B)}{P(C)} = \frac{P(C,B,W) + P(C,B,NW)}{P(C)}$$

$$P(C) = P(C,B,W) + P(C,B,NW) + P(C,NB,W) + P(C,NB,NW)$$

$$P(C,B,W) = P(B).P(W).P(C|W,B) = 0,8 \times 0,5 \times 0,9 = 0,36$$

$$P(C,B,NW) = P(B).P(NW).P(C|NW,B) = 0,8 \times 0,5 \times 0,5 = 0,2$$

$$P(C,NB,W) = P(NB).P(W).P(C|W,NB) = 0,2 \times 0,5 \times 0,4 = 0,04$$

$$P(C,NB,NW) = P(NB)P(NW).P(C|NW,NB) = 0,2 \times 0,5 \times 0,1 = 0,01$$

$$P(B|C) = \frac{0,36+0,2}{0,36+0,2+0,04+0,01} = \boxed{\frac{0,56}{0,61}} \approx 0,92$$

Name and Student ID:

Q3)[25pts]
In the table below, x_1, x_2 and $x_i \in \{0,1\}$, $i = 1,2$ x_i represent the i feature vector and $y \in \{+,-\}$ represents the class label. Generate a decision tree for this dataset using Entropy as the impurity measure.

Id	x_1	x_2	y
1	0	0	-
2	0	1	+
3	1	0	+
4	1	1	-

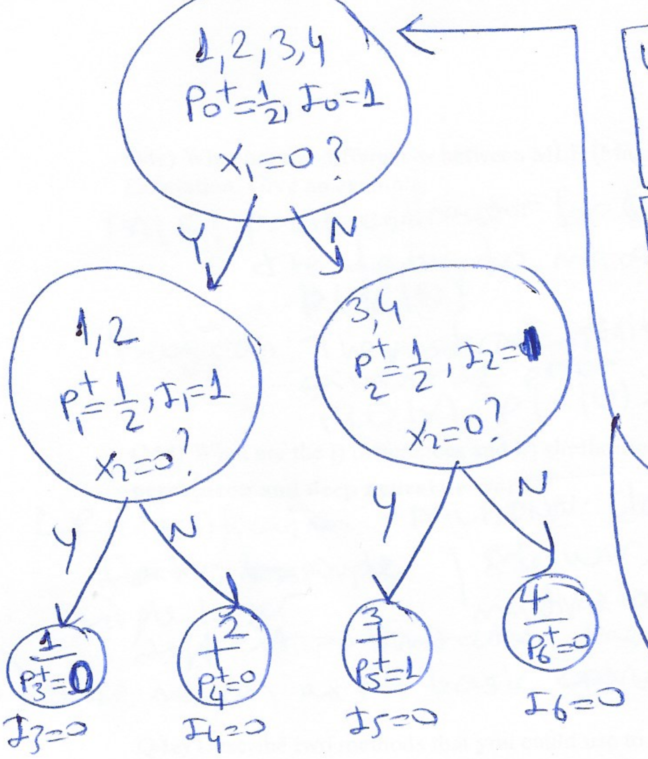
HINT:
For node m, N_m instances reach m, N^i_m belong to C_i

$$\hat{P}(C_i \mid \mathbf{x}, m) \equiv p^i_m = \frac{N^i_m}{N_m}$$

Node m is pure if p^i_m is 0 or 1

Measure of impurity is entropy

$$\mathcal{I}_m = -\sum_{i=1}^{K} p^i_m \log_2 p^i_m$$



When $p_m^- = p_m^+ = \frac{1}{2}$

$I_m = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$

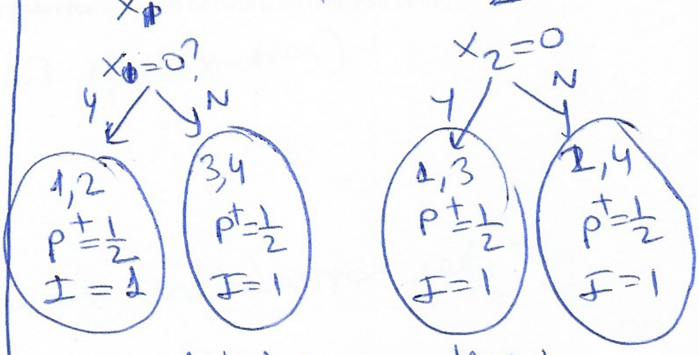
When $p_m^- = 0$, $p_m^+ = 1$ or $p_m^- = 1$, $p_m^+ = 0$

$I_m = -0\log 0 + 1\log 1 = 0$

Since there are only 2 classes we'll only show p_m^+ at the decision tree nodes

At node 0 (root) we can split according to

x_1 OR x_2



Splitting according to x_1 or x_2 gives the same entropy value. I chose X1 to split first (could have chosen x_2 also)

Name and Student ID:

Q4) [25pts]

Q4a). Given a dataset with N=2million instances and a neural network to classify it, which cross validation methods would you use to determine the number of hidden units in the neural network?

Kfold, 5x2, Bootstrapping OR

Divide whole data into Train, Val, Test sets.
Take a subset of Train, train, classify the remaining training data.
Add some misclassified instances to train set from training data
Continue while validation error keeps decreasing

Q4b) Give two examples of kernels that are used with SVM (Support Vector Machine) classifiers.

Linear, Quadratic, Polynomial, RBF (Gaussian)

Q4c) What are the differences between MLE (Maximum Likelihood Estimation) and Bayesian Estimation. Give an example.

MLE: point estimator for the parameter computed based on the data, to maximize the (log) likelihood $P(X|\theta)$

Bayesian: The posterior distribution for the parameter computed given the prior distribution & data
$P(\theta|X) = P(X|\theta) \cdot P(\theta) / P(X)$.

e.g. Gaussian mean
MLE: $m = \sum x_i / N$
Bayesian: $a \cdot m + b \cdot \mu_0$ where μ_0 is the mean of the prior distribution

Q4d) What are the i) differences and ii) similarities between **logistic regression, multilayer perceptron and deep neural networks**?

i) LR: classification | MLP, DNN: classification & regression
faster, ~~simple~~ | slower, more complex
single layer | multiple layer, DNN: lots of layers
different optimization methods, activation fns.

ii) Use neuron as the basic computation unit

Q4e) Describe two methods that you could use to **regularize** a neural network so that you could prevent overfitting.

Weight decay (L1 or L2 regularization)
weight elimination
Dropout
Hints
Early stopping based on validation set.