AMARTYA DUTTA

Advanced Machine Learning: Final Exam

Section A

A1.) Model 1 is too simple because it is a linear model. ~~as~~ Compared to that the data points to be fitted are generated by a function of the 4th power. Thus, model 1 will end up underfitting on the data while mode 2 will fit better to the data.

Thus, model 1 will have a higher bias than model 2

Ans: (B) Higher

A2) For Bayes classifier, no of params:

$$3 \times (2^{10}) = 3072$$

For Naive Bayes classifier: $3 \times (10+1) = 33$

Ans: (D) 3072, 33

A3.) Since it is a One vs Rest approach, the no of classifiers needed = no of classes. But when classes = 2, it is binary classification, so 1 will do.

∴ 5 & 1

Ans. ~~(D) 5, 1~~ (B) 5, 1

A4) Since we want to carry out the Expectation-Maximization algorithm, therefore we want to maximise the log likelihood

$$\because \ln p(X|\theta) = \mathcal{L}(q,\theta) + KL(q\|p)$$

~~Therefore, we want to~~

$$= \sum_z q(z) \ln \left\{ \frac{p(X,z|\theta)}{q(z)} \right\}$$

$$- \sum_z q(z) \ln \left\{ \frac{p(z|X,\theta)}{q(z)} \right\}$$

in order to maximise $\ln p(X|\theta)$

→ we need to maximise $\mathcal{L}(q|\theta)$
→ we need to minimise $KL(q\|p)$

~~Ans: (c) Maximize, minimize~~

Ans (c) Maximize, minimize

A5.) The probability distribution of the latent variables is hard to find and calculate. That is why an approximate value is calculated in the EM algorithm. The probability distribution q handles this by estimating the posterior.

Ans: (B) $p(Z|X,\theta)$

Section: B

B1.) In leave-one-out cross validation 1 sample point is used as validation while remaining n-1 samples are used for training. Thus it is a special case of k-fold cross validation where k is no. of training samples

Ans: True

B2.) Since, the test set data has been used to select model parameters, the test set error is biased over the training set error

Ans: False

B3.) During gradient-descent for a logistic regression model, it may get stuck at a local minima as well.

Ans: False

B4.) The convolution operations are linear transformations & it is the non-linear activation functions that introduce non-linearity. Without them, convolutional layers cannot do non-linear classification.

Ans: False

B5.) When there is a small training set - cross validation is indeed used to train a model and tune the parameters.

Ans: True

Section C

1)

Q1) Logistic regression is usually limited to a binary classification problem.
But logistic regression can be extended to multi-class classification using a one-vs-rest approach. ~~Where~~ In this approach, a seperate classifier is trained for each class.

Q2) A small training but a high validation error implies that the model is overfitting. Thus, it has a high variance and low bias.
Thus, L2 regularization should be increased.

Q3) Two ways to deal with non-linearly seperable data using SVM are:

i) Since SVM is a linear classifier, it introduces the concept of hyperplane which projects the otherwise non-linearly seperable data into a higher dimension where it is linearly seperable.

ii) Combines soft margin and kernel trick to modify the decision boundary for non-linearly seperable cases.

Q2)

Q.1) $f_x(x;\theta) = \theta x^{-\theta-2}$

$$L(\theta) = \prod_{i=1}^{n} f_x(x;\theta)$$

$$= \theta^n \prod_{i=1}^{n} x_i^{-\theta-2}$$

$$\ln(L(\theta)) = n\ln\theta + (-\theta-2)\sum_{i=1}^{n} \ln x_i$$

$$\frac{\partial \ln(L)}{\partial\theta} = \frac{n}{\theta} - \sum_{i=1}^{n} \frac{1}{x_i}$$

$\because \quad \frac{\partial^2 \ln(L)}{\partial\theta^2} = \frac{-n}{\theta^2} < 0 \qquad [\theta > 0]$

$\therefore \hat{\theta}_{MLE} = \frac{\partial \ln(L\theta)}{\partial\theta} \quad \frac{\partial \ln(L\theta)}{\partial\theta} = 0$

$$\Rightarrow \hat{\theta}_{MLE} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Q2) $p(\theta \mid x_1, x_2, \ldots x_n)$

$$= \frac{p(x_1, x_2 \ldots x_n \mid \theta) \cdot p(\theta)}{p(x_1, x_2 \ldots x_n)} \rightarrow \text{normalization}$$

$$\therefore p(x_1, x_2 \ldots x_n \mid \theta) \cdot p(\theta)$$
$$= \theta^n \prod_{i=1}^{n} x_i^{-\theta-2} \lambda \theta^{-\lambda-2}$$

$$\Rightarrow \ln\{p(x_1, x_2 \ldots x_n \mid \theta)) = n \ln\theta + (-\theta-2) \sum_{i=1}^{n} \ln x_i + \ln\lambda + (-\lambda-2)\ln\theta$$
$$\underbrace{\ln P(\theta)\}}_{\ln P(\theta)'}$$

$\therefore$ we want, $\frac{\partial P(\theta)'}{\partial \theta} = 0$

$$\Rightarrow \frac{n}{\theta} + \sum_{i=1}^{n} -\ln x_i - \frac{(\lambda+2)}{\theta} = 0$$

$$\frac{\partial^2 P(\theta)'}{\partial \theta^2} = -\frac{n}{\theta^2} + \frac{(\lambda+2)}{\theta^2}$$

$$= -\frac{(n-\lambda-2)}{\theta^2} < 0$$

$$\therefore \hat{\theta}_{MAP} = \frac{n-k-2}{\sum_{i=1}^{n} \ln x_i}$$

Q3) a) The likelihood function $L(\theta)$ is defined which is the probability distribution of a set of observations (x) ; given parameters (θ)

b) Log of the likelihood is taken as it makes computation easier

c) The set of parameters need to be specified as several unknown parameters are calculated from the data

d) An optimization of the MLE is performed. Here, we differentiate to get maxima

e) Finally, the log likelihood function is optimized using the parameter values achieved from previous step.

Q3)

Q1) Naive Bayes assumes all the features are independent and have an equal contribution in predicting the class value. In this dataset there are 3 features, hence 3 independent parameters.

Q2) Estimation of the parameter values

Size	Yes	No	P(Yes)	P(No)
Small	1	3	1/4	3/6 = 1/2
large	3	3	3/4	3/6 = 1/2
Total	4	6	1	1

Shape	Yes	No	P(Yes)	P(No)
Circle	3	2	3/4	2/6 = 1/3
Irregular	1	4	1/4	4/6 = 2/3
Total	4	6	1	1

Color	Yes	No	P(Yes)	P(No)
Red	4	1	1	1/6
Green	0	5	0	5/6
Total	4	6	1	1

Good Apple	
Yes	4
No	6
Total	10

$$\therefore P(Yes) = \frac{4}{10} = \frac{2}{5}$$

$$P(No) = \frac{6}{10} = \frac{3}{5}$$

Q3) New feature $x = \{ Small, Red, Round \}$

$$P(y=No|x) = \frac{P(small|No) \cdot P(Red|No) \cdot P(Round|No) \cdot P(No)}{P(x)}$$

$$= \frac{3}{6} \times \frac{1}{6} \times \frac{3}{6} \times \frac{6}{10}$$

$$= 0.016$$

$$P(y=Yes|x) = \frac{P(small|Yes) \cdot P(Red|Yes) \cdot P(Round|Yes) \times P(Yes)}{P(x)}$$

$$= \frac{1}{4} \times \frac{4}{4} \times \frac{3}{4} \times \frac{4}{10}$$

$$= 0.075$$

In order to convert it to a probability value

$$p(\text{yes}|x) = \frac{0.075}{0.075+0.016}$$

$$= 0.824$$

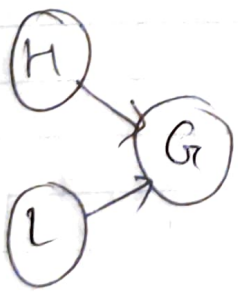
$$p(\text{No}|x) = \frac{0.016}{0.075+0.016}$$

$$= 0.176$$

$\because p(\text{yes}|x) > p(\text{No}|x)$, the Naive Bayes classifies y = yes for x

4) G = Going out ; H = Homework ; L = Lauren

a1)



$$p(H)\,p(L)\,p(G|H,L)$$

$$= p(G,H,L)$$

$\therefore$ Going out depends on having homework and if Lauren is available

Q2) Let the following hold:

$p(G, H, L) = p1$
$p(\bar{G}, H, L) = p2$
$p(G, \bar{H}, L) = p3$
$p(G, H, \bar{L}) = p4$
$p(\bar{G}, \bar{H}, L) = p5$
$p(\bar{G}, H, \bar{L}) = p6$
$p(\bar{G}, \bar{H}, \bar{L}) = p7$

$\therefore p(\bar{G}, \bar{H}, \bar{L}) = 1 - (p1 + p2 + p3 + p4 + p5 + p6 + p7)$

Let: $p(H) = P_h \quad p(\bar{H}) = 1 - P_h$

$p(L) = P_l \quad p(\bar{L}) = 1 - P_l$

$\therefore P(G, H, L) = p(H)\, p(L)\, p(G \mid H, L)$

$P1 = P_h \times P_l \times P_{g'}$

then, $P(\bar{G}, H, L) = P(H) \times P(L)\, p(\bar{G} \mid H, L)$

$P2 = P_h \times P_l \times (1 - P_{g'})$

$\Rightarrow P2 = P_h \times P_l - P_1$

$\therefore$ P1 & P2 are interrelated

$$P(G, \bar{H}, L) = P(\bar{H})\, P(L)\, P(G \mid \bar{H}, L)$$

$$P3 = (1 - P_h)\, P_l\, P_{G''}$$

$$P(\bar{G}, \bar{H}, L) = P(\bar{H})\, P(L)\, P(\bar{G} \mid \bar{H}, L)$$

$$P5 = (1 - P_h)\, P_l\, (1 - P_{G''})$$

$$P_5 = (1 - P_h) P_l - P_3$$

$\therefore$ P_5 & P_3 are also related

Similarly, P_4 & P_6 are also related.

$$\therefore P(\bar{G}, \bar{H}, \bar{L}) = 1 - (P1 + P2 + P3 + P4 + P5 + P6 + P7)$$

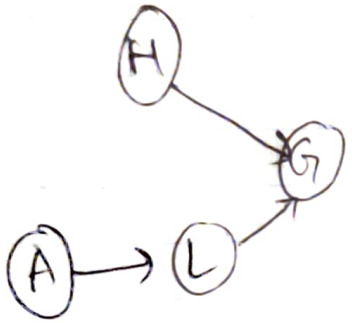
$\because$ ($P1$ & $P2$), ($P3$ & $P5$), & (P_4 & $P6$) are related

Thus, it depends on 6 parameters:

p_1/p_2; p_2/p_4, p_3/p_5; p_7, p_h and p_l

Q3) After G has been observed, H and L do not depend on each other and are thus independent.

Q4)



$$p(G, H, L, A)$$
$$= p(H)\, p(A)\, p(L|A)\, p(G|H, L)$$

Thus, if L is given then A & G are independent of each other

Q5) A → L → G

$$p(G, L, A) = p(A)\, p(L|A)\, p(G|L)$$

$p(G, L, A) = q1$ $p(\bar{G}, L, A) = q2$ $p(G, \bar{L}, A) = q3$

~~$p(G, \bar{L}, A) = q5$~~

$p(G, L, \bar{A}) = q4$ $p(\bar{G}, \bar{L}, A) = q5$ $p(\bar{G}, L, \bar{A}) = q6$

$p(G, \bar{L}, \bar{A}) = q7$

Let: $P(A) = P_a$ $P(L|\bar{A}) = P_l$
$P(L|A) = P_l$

$\therefore$

$$p(G, L, A) = p(A)\, p(L|A)\, p(G|L)$$
$$q1 = P_a P_l P_g \quad \text{(assume)}$$

$$p(\bar{G}, L, A) = p(A)\, p(L|A)\, p(\bar{G}|L)$$
$$q2 = P_a P_l (1 - P_g)$$

$\Rightarrow q_2 = P_2 P_2 - q_1 \quad -(i)$

$\therefore q_1$ and q_2 are related

$$P(G, \bar{L}, A) = P(A)\, p(\bar{L}|A)\, p(G|\bar{L})$$

$$q_3 = P_a (1 - P_\ell)\, P_g' \text{ (assume)}$$

$$P(\bar{G}, \bar{L}, A) = P(A)\, P(\bar{L}|A)\, P(\bar{G}|\bar{L})$$

$$q_5 = P_a (1 - P_\ell)(1 - P_g')$$

$$\Rightarrow q^5 = P_a (1 - P_\ell) - q_3$$

$\therefore$ q_3 and q_5 are also related

Also,

$$P(G, L, \bar{A}) = p(\bar{A})\, p(L|\bar{A})\, P(G|L)$$

$$q_4 = (1 - P_a) P_\ell'\, P_g'' \text{ (assume)}$$

$$P(\bar{G}, L, \bar{A}) = P(\bar{A})\, P(L|\bar{A})\, P(\bar{G}|\bar{L})$$

$$q_6 = (1 - P_a) P_\ell' (1 - P_g'')$$

$$\Rightarrow q_6 = (1 - P_a) P_\ell' - q_4$$

$\therefore$ q_4 and q_6 are also related.

Thus for the joint probability $P(\bar{G}, \bar{L}, A)$ we will need q_1/q_2 ; q_3/q_5 ; q_4/q_6

; q7 ; P_a ; P_2 and P_2'

5)

Q1) Size of image : $3 \times 32 \times 32$

This is in order of $(C \times H \times W)$

originally image is : $32 \times 32 \times 3$
$(H \times W \times C)$

Conv2D : input channel $\rightarrow 3$
output channel $\rightarrow 6$
Kernel $\rightarrow 3 \times 3$

Pooling
Kernel $\rightarrow 2 \times 2$
Stride $\rightarrow 2$

Size of image after convolution :

$$H' = W' = \frac{H - K + 2p}{S} + 1$$

$$= 32 - 3 + 1$$
$$= 30$$

∵ output channel $= 6$

∴ size of image $= (30 \times 30 \times 6)$

Size after max pooling :

$$\text{Size} = \frac{H - K}{S} + 1 = \frac{30 - 2}{2} + 1$$
$$= 15$$

∴ Final image size = (15 x 15 x 6)

Q2) Max pooling outputs max values in each window position, while average pooling will give an average of the features in the window.

Size of feature map: 4x4

Max pooling:

∴ Output shape → 2x2

kernel → 2x2
stride → 2

1	1	2	4
5	6	7	8
3	2	1	0
1	2	3	4

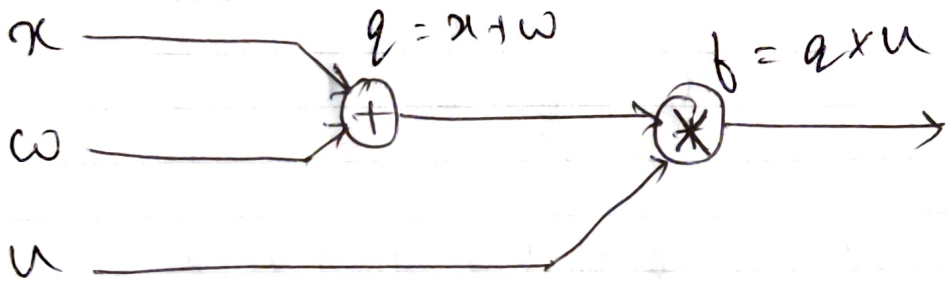
→ Max Pooling:

6	8
3	4

↓ Averg. Pooling

3	5
2	2

B3) $x = -2 \quad w = 5 \quad u = -4$ (initial values)



$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial q} \times \frac{\partial q}{\partial x} = \frac{\partial (q \times u)}{\partial q} \times \frac{\partial (x+w)}{\partial x}$$

$$= u \times 1 = -4$$

$$\frac{\partial f}{\partial w} = \frac{\partial f}{\partial q} \times \frac{\partial q}{\partial w} = \frac{\partial (q \times u)}{\partial q} \times \frac{\partial (x+w)}{\partial w}$$

$$= u \times 1$$
$$= -4$$

$$\frac{\partial f}{\partial u} = \frac{\partial (q \times u)}{\partial u} = q = (x+w) = 3$$

6)

Q1) We need to choose 3 clusters from the 2D space where there are 6 points

$$\therefore {}^6C_3 = \frac{6 \times 5 \times 4}{3 \times 2} = 20$$

Q2)

	3 partition	Stable?	An ex of 3 start config
i)	{a,b}, {c,f}, {d,e}	yes	a, c, d
ii)	{a,b,e}, {c,d}, {f}	no	none
iii)	{a,d}, {b,c}, {e,f}	yes	a, b, e
iv)	{a,b,d}, {e}, {c,f}	no	none
v)	{a,d,e}, {b,c}, {f}	yes	d, c, f

Explaining using (iii) as an example:

Let the initial cluster centroids be a, b, e

1) For a(0,0) → nearest centroid a(0,0)
   For b(8,0) → nearest centroid b(8,0)
   For c(16,0) → nearest centroid b(8,0)
   For d(0,6) → nearest centroid a(0,0)
   For e(8,6) → nearest centroid e(8,6)
   For f(16,6) → nearest centroid e(8,6)

2) For centroid movement:

a = centroid for a and d → cluster 1

$$(x_{cent})_1 = \frac{(a_{cent})_x + (d_{cent})_x}{2} = 0$$

$$(y_{cent})_1 = \frac{(a_{cent})_y + (d_{cent})_y}{2} = 3$$

∴ centroid is at (0, 3), hence converged

b = centroid for b & c = cluster 2

$$(x_{cent})_2 = \frac{(b_{cent})_x + (c_{cent})_x}{2} = 12$$

$$(y_{cent})_2 = \frac{(b_{cent})_y + (c_{cent})_y}{2} = 0$$

∴ centroid is at (12, 0) and so it converged

e = centroid for e & f → cluster 3

$$(x_{cent})_3 = \frac{(e_{cent})_x + (f_{cent})_x}{2}$$

$$= 12$$

$$(y_{cent})_3 = \frac{(e_{cent})_y + (f_{cent})_y}{2}$$

$$= 6$$

∴ centroid is at (12, 6) & converged

new centroids:

{(0,3), (12,0), (12,6)}

For
a (0,0) → nearest centroid (0,3)
b (8,0) → nearest centroid (12,0)
c (16,0) → nearest centroid (12,0)
d (0,6) → nearest centroid (0,3)
e (8,6) → nearest centroid (12,6)
f (16,6) → nearest centroid (12,6)

∵ clusters remain the same, re calculating will lead to same clusters.

Thus, it has converged & is stable

7)

Q1) given,

$x, x' \in R^n$

$k: R^n \times R^n$ R is valid if there is a feature vector s.t $\phi : R^n \to R^m$

$$k(x, x') = \phi(x)^T \phi(x')$$

$k_1(x, x')$, $k_2(x, x') \to$ valid kernels

For $k(x, x')$ to be a valid kernel the gram matrix K_m must be +ve semidefinite for $\forall \{x_n\}$ where elements in K_m are $k(x_n, x_m)$

For K_m to be positive semi-definite

$$y^T K_m y \geq 0 \quad \to y \text{ is any vector}$$

$K_{m_1} \to$ gram matrix of $k_1(x, x')$

$K_{m_2} \to$ gram matrix of $k_2(x, x')$

$\because$ K_{m_1} & K_{m_2} are +ve semi-definite

$$y^T K_{m_1} y \geq 0$$

$$y^T K_{m_2} y \geq 0$$

∴ Gram matrix of $k_1(x, x')$ & $k_2(x, x')$

$= K_{m_1} + K_{m_2}$

$= y^T (K_{m_1} + K_{m_2}) y$

$= y^T K_{m_1} y + y^T K_{m_2} y \geq 0$

∴ $k(x, x') = k_1(x, x') + k_2(x, x')$ is a valid Kernel

Q2) $k_1(x, x') = f(x)^T f(x')$
$k_2(x, x') = g(x)^T g(x')$ } (assuming)

$k_1(x, x') \cdot k_2(x, x') = f(x)^T f(x') g(x)^T g(x')$

$= \sum_{p=1}^{X} f_p(x) f_p(x') \sum_{q=1}^{Y} g_q(x) g_q(x')$

$= \sum_{p=1}^{X} \sum_{q=1}^{Y} f_p(x) f_p(x') g_q(x) g_q(x')$

$= \sum_{m=1}^{M} \phi(x) \phi(x') = \phi(x)^T \phi(x')$

Where, $M = XY$

$\therefore$ $k_1(x, x')$ & $k_2(x, x')$ are valid kernels

Q3) $k_3(x, x') \rightarrow$ a valid kernel

also, $k_3(x, x') = f(x)^T f(x')$

$f(x) \rightarrow$ feature vector

$B_x \Rightarrow$ kernel const.

$B_0 k_3(x, x') = h(x)^T h(x')$ $\quad (h(x) = \sqrt{B_0}\, f(x))$

$B_0 k_3(x, x')$ is the scalar product of feature vectors. So, it is a valid kernel

Also, $p(x)\, k_3(x, x')\, p(x')$

$= p(x) f(x)^T f(x') p(x')$

$= \phi(x)^T \phi(x')$ $\quad [\phi(x) = p(x) f(x)]$

$\rightarrow$ a scalar product of coefficients of all orders in $p(x)$ & $f(x)$

$\therefore$ $k(x, x') = p(x)\, k_3(x, x')\, p(x')$ is a valid kernel

04) $e^x = 1 + x + \frac{x^2}{2!} + \dots \frac{x^n}{n!}$ (Taylor series)

$k_3(x, x') = f(x)^T f(x')$

$e^{k_3(x,x')} = 1 + f(x)^T f(x') + \frac{[f(x)^T f(x')]^2}{2!} + \dots \frac{[f(x)^T f(x')]^n}{n!}$

$[f(x)^T f(x')]^n = \underbrace{[f(x)^T f(x')] \times \dots}_{n\text{-times}}$

also, $f(x)^T f(x')$ are valid kernels

$\therefore e^{k_3(x,x')}$ is a valid kernel

05) $k_3(x, x') = \exp\{-\frac{\|x - x'\|^2}{2\sigma^2}\}$

$\Rightarrow \|x - x'\|^2 = x^T x - 2x^T x' + (x')^T x'$

we already know, $e^{k_1(x,x')}$ → valid kernel

$\therefore k_1(x, x')$ → valid kernel

$= e^{\left(\frac{1}{2\sigma^2}\left[x^T x - 2x^T x' + (x')^T (x')\right]\right)}$

$$= e^{-\frac{x^T x}{2\sigma^2}} \cdot e^{-\frac{x^T x'}{\sigma^2}} \cdot e^{-\frac{(x')^T (x')}{2\sigma^2}}$$

$$= g(z') e^{k_1(z,z')} \cdot g(z')$$

where $k_1(z,z') = B_0 x^T x$

$k_1(x,x') = \exp\left\{\frac{-\|x - x'\|^2}{2\sigma^2}\right\}$ is a valid kernel

8)

$$p(x) = \sum_{k=1}^{k} \pi_k p(x|k)$$

mean $\{k_1, \ldots k_n\} = \{\mu_1, \ldots \mu_n\}$

covariance $= \{\Sigma_1, \ldots \Sigma_n\}$

Q1) $$E(x) = \int_x x\, p(x)\, dx$$

$$= \int_x x \sum_{k=1}^{k} \pi_k p(x|k)\, dx$$

$$= \int_x x \{\pi_1 p(x|k=1) + \ldots \pi_k p(x|k=k)\}\, dx$$

$$= \pi_1 \mu_1 + \pi_2 \mu_2 + \ldots \pi_k \mu_k$$

$$\therefore E(x) = \sum_{k=1}^{K} \pi_k \mu_k$$

Q2) $E[x|k=K] = \mu_K$

$Cov[x|k=K] = \Sigma_K$

$\therefore Cov[X_i, Y_j | k=K]$

$= E(X_i Y_j | k=K) - E(X|k=K] - E(Y|k=K)$

Also, $E(x) = \sum_{k=1}^{K} \pi_k \mu_k$

$\therefore E(X_i Y_j) = \sum_{k=1}^{K} E(X_i Y_j | k=K)\ \pi_k$

$= \sum_{k=1}^{K} \pi_k \left[(\Sigma_k)_{ij} + E(X_i|k=K) + E(Y_j|k=K) \right]$

$= \sum_{k=1}^{K} \pi_k \left[(\Sigma_k)_{ij} + (\mu_k)_i + (\mu_k)_j \right]$

$- \sum_{k=1}^{K} (\mu_k)_i \pi_k \sum (\mu_k')_j \pi_{k'}$

$E(x)$ $\quad$ $E^T(x)$

$\therefore Cov(X_i, Y_j) = E[(Cov(X_i, Y_j | z))$

$+ Cov(E(X_i|z), E(Y_j|z))$

Thus, from law of co-variance

$$Cov[x] = \sum_{k=1}^{K} \pi_k (\Sigma_k + \mu_k \mu_k^T) - E(x) E(x)^T$$

9)

a) $$L(q) = \int q(z) \frac{\ln P(x,z)}{q(z)} dz$$

$$KL(q \| p) = -\int q(z) \frac{\ln p(z|x)}{q(z)} dz$$

Product rule:

$$L(q) = \int q(z) \ln \left\{ \frac{P(x|z) \cdot P(z)}{q(z)} \right\} dz$$

$$P(x|z) = \frac{P(z|x) \cdot P(x)}{P(z)}$$

$$\therefore L(q) = \int q(x) \ln \left\{ \frac{P(z|x) \cdot P(x)}{P(z) q(z)} P(z) \right\} dz$$

$$= \int q(z) \left\{ \ln \left(\frac{P(z|x)}{q(z)} \right) + \ln P(x) \right\} dz$$

$$= \int q(z) \frac{\ln p(z|x)}{q(z)} dz + \int \ln (P(x)) dz$$

$$\mathcal{L}(q) = -KL(q\|p) + \ln P(x)$$

$$\therefore \ln P(x) = KL(q\|p) + \mathcal{L}(q)$$

Hence proved

8.2) $$q(z) = \prod_{i=1}^{M} q_i(z_i)$$

$$KL(p\|q) = -\int p(z) \left[\sum_{i=1}^{M} \ln q_i(z_i)\right] dz + c$$

∴ optimising wrt $q_k(z_k)$

$$KL(p\|q) = -\int p(z) \ln q_k(z_k) + p(z) \sum_{i \neq k} \ln q_i(z_i)\, dz + c$$

$$= -\int p(z) \ln q_k(z_k)\, dz + c$$

$$= -\int \ln q_k(z_k) \left[\int p(x) \prod_{i \neq k} dz_k\right] dz$$

$$= -\int B_k(z_k) \ln q_k(z_k)\, dz_k + C$$

now, $$B_k(z_k) = \int p(z) \prod_{i \neq k} dz_k$$

Now using Lagrange's multiplier: so the $q_k(z_k)$ is:

$$-\int B_k(z_k) \ln q_k(z_k)\, dz_k + \lambda \int q_k(z_k)\, dz_k - 1$$

Taking the derivative and equating to 0

$$-\frac{B_k(z_k)}{q_k(z_k)} + \lambda = 0$$

$$\Rightarrow \lambda q_k(z_k) = B_k(z_k)$$

$\therefore$ Integrating $\rightarrow$

$$\lambda = \int B_k(z_k)\, dz_k$$

$$= \int \left[\int p(z) \prod_{i \neq k} dz_i \right] dz_k = 1$$

$$\therefore q_k(z_k) = \int p(z) \prod_{i \neq k} dz$$

10)

Q1) The likelihood of a dataset $P(X|\theta)$ can be computed ~~as~~: using the EM algorithm.

In the E step the current parameters are used to compute posterior distribution of latent variables $P(Z|X,\theta^{old})$

~~In the M step the current parameters are used to compute the posterior distribution of latent variables~~

In the M step - the parameters are optimized

$$\theta^{new} = \arg\max Q(\theta, \theta^{old})$$

$$Q(\theta, \theta^{old}) = \sum_{Z} P(Z|X,\theta^{old}) \ln P(X,Z|\theta)$$

$$\alpha(z_t) = P(x_1, \dots x_t, z_t)$$

$$\beta(z_t) = P(x_{t+1}, \dots x_T | z_t)$$

$$\xi(z_{t-1}, z_t) = P(z_{t-1}, z_t | X)$$

$$= \frac{\alpha(z_{t-1}) P(x_t|z_t) P(z_t|z_{t-1}) \beta(z_t)}{\sum_{z_t} \alpha(z_T)}$$

Also,

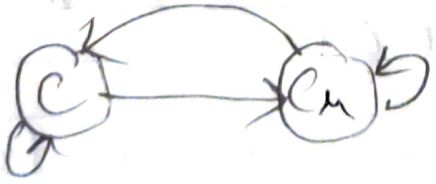
$$\pi_k = \frac{\gamma(z,k)}{\sum_{j=1}^{K} \gamma(z,j)}$$

$$A_{jk} = \frac{\sum_{t=1}^{T} \xi_t(z_{t-1}, j z_{tk})}{\sum_{m=1}^{K} \sum_{t=2}^{T} \xi_t(z_{t-1}, j z_{tk})}$$

∴ From $\xi(z_{t-1}, j z_{tk})$, if $A_{jk} = 0$ then $\xi(z_{t-1}, j z_{tk}) = 0$

Thus, no more updates in EM algorithm

Q2) Chocolate lever = C
Caramel lever = C_R



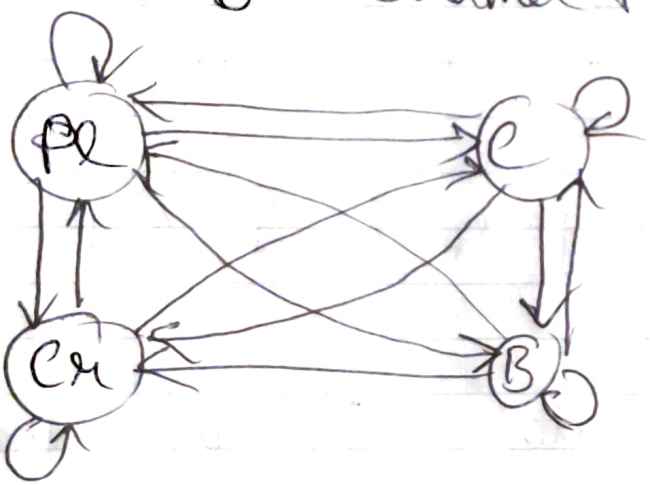
Prior probability ds^n

$$P(C) = P(C_R) = 0.5$$

Transition probability table :

T →	C	Cr
C	0.7	0.3
Cr	0.3	0.7

Let : Pl = Plain
B = Caramel + Chocolate (Both)



∵ levers are ind.
∴ $P(Pl) = P(C) = P(Cr) = P(B) = 0.25$

State transition probability of both being flipped = $(0.3)2 = 0.09$

State transition probability of one being flipped
$= 0.3 \times 0.7 = 0.21$

State transition probability of neither being flipped

$= 1 - 0.09 - 0.21 - 0.21$

$= 0.49$

$\therefore T =$

	Pl	C	Ca	B
Pl	0.49	0.21	0.21	0.09
C	0.21	0.49	0.09	0.21
Ca	0.21	0.09	0.49	0.21
B	0.09	0.21	0.21	0.49

03) Probability ~~of~~ the order:

P{Plain, Chocolate, Chocolate, Chocolate + Caramel}

$= P(Pl, C, C, B)$

$= P(P) \cdot P(C|Pl) \cdot P(C|C) \cdot P(B|C)$

$= 0.25 \times 0.21 \times 0.49 \times 0.21$

$= 0.0054$