Question 1

a)
Good training performance with poor validation performance suggests the model is overfitting. For SVM, hyperparameters C and gamma can be changed to ease the issue of overfitting. Specifically, we can 1) lower the value of C to increase the margin and penalty for misclassification, or 2) lower the value of gamma for stricter classification.

b)
AdaGrad converges slower than RMSProp. The calculation of AdaGrad involves dividing the square root of the accumulative sum of the squared gradient, as the accumulative sum increases, the update may become very slow. On the other hand, RMSProp focuses only on recent squred gradients and thus converges faster.

c)
The reparameterisarion trick is used in VAE to make it capable of applying gradient descent through the samples of latent z to the encoder. The reparameterization trick involves sampling from a standard Gaussian distribution, then applying transformation $\mu + \sigma^2 z_0$
This makes the sampling processing equivalent to sample from $N(\mu, \sigma^2 I)$

d)
Because RNN takes sequence that involves time-states as input.

The output in RNN = $\text{activation}(W_0 \cdot \text{input\_t} + U_0 \cdot \text{state\_t} + b_0)$

where the stats_t are the time-states. As a result, when performing backpropagation to update the weights, we need to unroll all input time-states, therefore the name backpropagation through time.

e)
L ⊂ M ⊂ E

Question 1 continue

f)
For the first square-loss risk equation, we are decomposing the square-loss between the actual θ and the estimated $\hat{\theta}$ of an observed sample, while for the second, we are decomposing the square-loss risk between the population Y and the estimated $\hat{Y}$ from the sample.

g)

$$L = \sum_{i=1}^{n} \log p(x_1, \dots x_n) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{k} p(c_j)\, p(x_i \mid c_j) \right)$$

$$= \sum_{i=1}^{n} \log \left(\sum_{j=1}^{k} w_j\, \mathcal{N}(x \mid \mu_j, \Sigma_j) \right)$$

h)
Using Newton-Raphson to train linear regression is equivalent to solving the weights analytically. That's why only one iteration is needed.
More specifically:

Normal equation : $\hat{\theta} = (X^T X)^{-1} X^T y$

Newton-Raphson : $\theta^{t+1} = \theta^t - (\nabla_2 L(\theta))^{-1} \nabla L(\theta)$

with linear regression : $L(\theta) = \frac{1}{2} \|X\theta - y\|_2^2$

Newton-Raphson becomes: $\theta^t - (X^TX)^{-1}(X^TX\theta^t - X^Ty)$

$= (X^TX)^{-1}[X^TX\theta^t - (X^TX\theta^t - X^Ty)]$

$= (X^TX)^{-1}X^Ty$

which is the same as the normal equation (analytical solution)

Question 2

a) $Z = \sum_{a \in A, b \in B, c \in C, d \in D, e \in E} f(A,B,C) \cdot G(C,D,E)$

A	B	C	D	E	result (f(A,B,C) · G(C,D,E))
0	0	0	0	0	3 x 1 = 3
0	0	0	0	1	3 x 4 = 12
0	0	0	1	0	3 x 4 = 12
0	0	0	1	1	3 x 2 = 6
0	0	1	0	0	2 x 0 = 0
0	0	1	0	1	2 x 0 = 0
0	0	1	1	0	2 x 0 = 0
0	0	1	1	1	2 x 0 = 0
0	1	0	0	0	3 x 1 = 3
0	1	0	0	1	3 x 4 = 12
0	1	0	1	0	3 x 4 = 12
0	1	0	1	1	3 x 2 = 6
0	1	1	0	0	5 x 0 = 0
0	1	1	0	1	5 x 0 = 0
0	1	1	1	0	5 x 0 = 0
0	1	1	1	1	5 x 0 = 0

A	B	C	D	E	result (f(A,B,C) · G(C,D,E))
1	0	0	0	0	3 x 1 = 3
1	0	0	0	1	3 x 4 = 12
1	0	0	1	0	3 x 4 = 12
1	0	0	1	1	3 x 2 = 6
1	0	1	0	0	4 x 0 = 0
1	0	1	0	1	4 x 0 = 0
1	0	1	1	0	4 x 0 = 0
1	0	1	1	1	4 x 0 = 0
1	1	0	0	0	3 x 1 = 3
1	1	0	0	1	3 x 4 = 12
1	1	0	1	0	3 x 4 = 12
1	1	0	1	1	3 x 2 = 6
1	1	1	0	0	9 x 0 = 0
1	1	1	0	1	9 x 0 = 0
1	1	1	1	0	9 x 0 = 0
1	1	1	1	1	9 x 0 = 0

$$\therefore Z = 3+12+12+6 \;+3+12+12+6+3+12+12+6+$$

$$3+12+12+6 = 132$$

b) $P(A=F, B=F, C=F) = \frac{1}{Z} \sum_{D=d, E=e} f(A=F, B=F, C=F)$

$$\cdot G(C=F, D=d, E=e)$$

$$= \frac{1}{132} \cdot \Big(3 \times G(C=F, D=T, E=T) + 3 \times G(C=F, D=T, E=F)$$

$$+ 3 \times G(C=F, D=F, E=T) + 3 \times G(C=F, D=F, E=F)\Big)$$

$$= \frac{1}{132} \cdot (3\times 2 + 3\times 4 + 3\times 4 + 3\times 1)$$

$$= \frac{1}{132} \cdot 33 = \frac{33}{132}$$

Question 3

a) $E_\theta[\hat{\theta}(X_1, \ldots X_n)] - \theta \to 0$

$\therefore \hat{\theta}_n = \hat{\theta}(X_1, \ldots X_n)$ is an unbiased estimator of θ in general

b) No

c) consistency is asymptotic, but $\hat{\theta}_n$ does not involve multiple samples

Question 4

a) VC(F) = 3 + 3 = 6

b)

Question 5

a) B and D

b) A, C, E, F and G

c) No. Because point A is not a support vector (does not line on the margin), while the decision boundary is affected by the support vectors.

Question 6

a)
11 x 11 x 3 x 32 = 11616 (filter size x input channels x output channels)

b)
1. stacking 5 3 x 3 convolutional layers: 3 x 3 x 5 < 11 x 11
2. stacking 2 5 x 5 and 1 3 x 3 convolutional layers: 5 x 5 x 2 + 3 x 3 < 11 x 11
3. stacking 1 7 x 7 and 1 5 x 5 convolutional layers: 7 x 7 + 5 x 5 < 11 x 11
4. stacking 1 7 x 7 and 2 3 x 3 convolutional layers: 7 x 7 + 3 x 3 x 2 < 11 x 11
5. stacking 1 5 x 5 and 3 3 x 3 convolutional layers: 5 x 5 + 3 x 3 x 3 < 11 x 11

Question 7

a)
arms: the choices of restaurants displayed on the app front page for users to choose from;
rounds: 1 round corresponds to the app is opened for 1 time;
rewards: whether the user clicks on one of the restaurants pushed to the app front page and makes an order - when an order's made, get rewards, otherwise no rewards;
context: the user's demographic info (e.g. gender, age, address), the user's order history, and the user's feedback for previous orders etc

b)
Information relating to the user's context as suggested above;
the restaurants that have been pushed to the app front page for the user to choose from each time the app is opened;
if the user places an order in one of the restaurants pushed to the app front page;
if not, in which restaurant the user actually places an order

c)
With MlinUCB, a reward amount would be estimated for missing rewards using its associated contexts and past rewards with similar contexts.
Personally, I believe it is more sensible to use only past rewards with similar contexts specific to the restaurant in question, given the restaurant itself is a factor that affects whether the user makes on order, on top of the contexts.
The limitation might be that if the previous experience with the restaurant was unpleasant, it may not be appropriate to use previous rewards to estimate current reward.

d)
Might be more effective if coupons aim more at the users who tend to browse more without placing an order.