1. Multiple Choice Questions.

Lasso. Question A:

True

Bagging & Bootstrap Sampling

Question B.

Option B.

Convolutional Filters

Question C.

Fig 2 (a) : K_3

Fig 2 (b) : K_1

Fig 2 (c) : K_2

Multiclass SVMs.

Question D.

$\hat{y} = 2$ is the maximizer of (1)

Feature Maps:

Question E.

Option A

Hard-Margin SVMs:

Question F.

False

Question G.

Option A.

AdaBoost:

Question H:

False.

Tensor Model Training

Question I:

True

Bias - Variance Decomposition:

Question J:

False

HMM EM Learning:

Question K:

Option A

Non-Negative Matrix Factorization:

Question L:

True

Decition Tree:

Question M:

False

Overfitting:

Question N:

True

2. Naive Bayes:

Question 1:

using maximum likelihood estimate formula:

eg: $$P(\text{Grade}=A \mid \text{Happy?}=\text{Yes}) = \frac{1+\sum_{(x,y)} 1_{[x_{Grade}=A \wedge y=\text{Yes}]}}{2+\sum_{(x,y)} 1_{[y=\text{Yes}]}}$$

$$= \frac{1+3}{2+4} = \frac{2}{3}$$

we can calculate:

$P(\text{Grade}=A \mid \text{Happy?}=\text{No}) = \frac{1}{3}$

$P(\text{Grade}=C \mid \text{Happy?}=\text{Yes}) = \frac{1}{3}$

$P(\text{Grade}=C \mid \text{Happy?}=\text{No}) = \frac{2}{3}$

therefore, final probability table for:

(table 1):

P(Grade | Happy?)	Grade = A	Grade = C
Happy? = Yes	2/3	1/3
Happy? = No	1/3	2/3

Similarly, we can calculate:

$P(\text{Year}=\text{Freshman} \mid \text{Happy?}=\text{Yes}) = \frac{1}{3}$

$P(\text{Year}=\text{Freshman} \mid \text{Happy?}=\text{No}) = \frac{1}{2}$

$P(\text{Year}=\text{Senior} \mid \text{Happy?}=\text{Yes}) = \frac{2}{3}$

$P(\text{Year}=\text{Senior} \mid \text{Happy?}=\text{No}) = \frac{1}{2}$

therefore, final probability table for:

(table 2):

P(Year | Happy?)	Year = Freshman	Year = Senior
Happy? = Yes	1/3	2/3
Happy? = No	1/2	1/2

Question 2:

$P(\text{Year}=\text{Freshman}, \text{Grade}=C, \text{Happy?}=\text{No})$

$= P(\text{Happy?}=\text{No})P(\text{Grade}=C \mid \text{Happy?}=\text{No})P(\text{Year}=\text{Freshman} \mid \text{Happy?}=\text{No})$

$= \frac{1}{2} \cdot \frac{2}{3} \cdot \frac{1}{2} = \frac{1}{6}$

Question 3:

Pseudocode for drawing a sample.

```
// first sample y (Happy?)
R_y = random()
Set Happy? = Yes if R_y < P(Happy? = Yes)
else Set Happy? = No

// then sample each x^d (Grade & Year)
 R_x1 = random()
 Set Grade = A if R_x1 < P(Grade = A | y)
 else Set Grade = C
 R_x2 = random()
 Set Year = Freshman if R_x2 < P(Year = Freshman | y)
 else Set Year = Senior
```

3. Data Transformation:

Question 1.

we know:

$$\tilde{x} = Ax$$

$$w^T x = \tilde{w}^T \tilde{x} = \tilde{w}^T A x$$

therefore,

$$w^T = \tilde{w}^T A$$

thus,

$$w = A^T \tilde{w} = A \tilde{w}$$

Question 2.

$$\text{argmin}_{\tilde{w}} \frac{\lambda}{2} \|\tilde{w}\|^2 + \sum_i (y_i - \tilde{w}^T \tilde{x}_i)^2$$

$$= \text{argmin}_{w} \frac{\lambda}{2} \| A^{-1} w \|^2 + \sum_i (y_i - w^T x_i)^2$$

Question 3:

when x is rescaling by A,

the regulation sector in loss function is changed from $\frac{\lambda}{2}\|w\|^2$ to $\frac{\lambda}{2}\|A^{-1}w\|^2$, then the optimal w would be different

4. Latent Markov Embedding

Question 1.

as U, V, X are optimal choices which maximize P(S),

single-data model is more restricted to dual-point model because it requires that "U" and "V" equals to each other,

$$P(S) = \prod_{p \in S} \prod_{i=1}^{M_p} \frac{e^{-\|U(p^{[i]}) - V(p^{[i-1]})\|_2^2}}{Z(p^{[i-1]})} \qquad (7)$$

$$P(s) = \prod_{p \in s} \prod_{i=1}^{M_p} \frac{e^{-\|X(p^{[i]}) - X(p^{[i-1]})\|_2^2}}{Z(p^{[i-1]})} \qquad (9)$$

therefore, $P(s)_{dual} \geq P(s)_{single}$

Question 2.

if (6) is equal to (8) for every pair of songs s and s'.

this case is satisfied when the optimal dual point model returns a solution U, V where both are equal to each other

$U = V$.

if this is the case, then the dual-point model changes into single-point model, as U, V, X are all optimal choices already, what we can interpret is that.

$U = V = X$.

5. Neural Net Backprop Gradient Derivation

Question 1:

$$\frac{\partial}{\partial \omega_{11}} L(y, f(x)) = \frac{\partial L}{\partial f} \cdot \frac{\partial f}{\partial h_1} \frac{\partial h_1}{\partial \omega_{11}}$$

we expand each , $\frac{\partial L}{\partial f} = -2(y-f)$

$$\frac{\partial f}{\partial h_1} = \frac{\partial \sigma(u_1 h_1 + u_2 h_2)}{\partial h_1} = \sigma(u_1 h_1 + u_2 h_2)(1-\sigma(u_1 h_1 + u_2 h_2))\, u_1$$

$$\frac{\partial h_1}{\partial \omega_{11}} = \frac{\partial \sigma(\omega_{11} x_1 + \omega_{21} x_2)}{\partial \omega_{11}} = \sigma(\omega_{11} x_1 + \omega_{21} x_2)(1-\sigma(\omega_{11} x_1 + \omega_{21} x_2))\, x_1$$

Question 2:

using formula in Question 1 and data provided.
we have:

$$\frac{\partial}{\partial \omega_{11}} L(y-f(x)) = \frac{\partial}{\partial \omega_{11}} (y-f(x))^2$$

$$= \frac{\partial L}{\partial f} \cdot \frac{\partial f}{\partial h_1} \cdot \frac{\partial h_1}{\partial \omega_{11}}$$

$\omega_{11} x_1 + \omega_{21} x_2 = 0.25 \times 0.1 + 0.05 \times 0.5 = 0.05$;

$$\frac{\partial h_1}{\partial \omega_{11}} = \sigma(0.05)(1-\sigma(0.05)) \times 0.1 = 0.025$$

$$u_1 h_1 + u_2 h_2 = u_1(\sigma(\omega_{11} x_1 + \omega_{21} x_2)) + u_2(\sigma(\omega_{12} x_1 + \omega_{22} x_2))$$

$$= 0.5 \times 0.5125 + (-0.1) \times 0.4713 = 0.2091$$

$$\frac{\partial f}{\partial h_1} = \sigma(0.2091)(1-\sigma(0.2091)) \times 0.5 = 0.1236$$

$$f = \sigma(u_1 h_1 + u_2 h_2) = \sigma(0.2091) = 0.5521$$

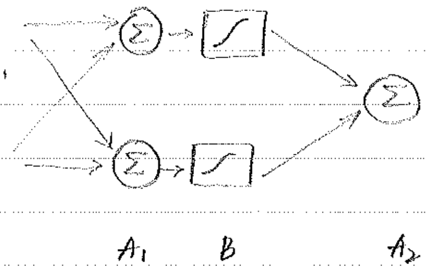
$$\frac{\partial L}{\partial f} = -2(y-f) = -2(0.75 - 0.5521) = -0.3958$$

therefore, $\frac{\partial}{\partial \omega_{11}} L(y-f(x)) = \frac{\partial L}{\partial f} \cdot \frac{\partial f}{\partial h_1} \cdot \frac{\partial h_1}{\partial \omega_{11}} = (-0.3958) \cdot 0.1236 \cdot 0.025$

$= -0.0012$

Question 3:

$\frac{\partial h_1}{\partial w_{11}}$ that is the ...



$\underbrace{\partial A_1 \cdot B}$ is always between 0 and 1, and when neural networks with more layers, they multiply together and become really small and gradient descent process becomes very ineffecient and results in vanishing gradient problem. More Layers means more value $\in (0,1)$ to be multiplied, so the problem is exacerbated exponentially with layer number.

we can solve the problem by changing the sigmoid function to Rectilinear function.