Question 1

1. (1) True
   (2) False
   (3) True
   (4) True
   (5) True
   (6) True
   (7) True
   (8) False.

2. Discriminative classifier learns a model for $p(t|X)$. An example is the logistic regression.
   Generative classifier learns a model for $p(x|t)$ with a prior assumption on the marginal density $Pr(x)$. One example is the Gaussian Discriminant Analysis.

3. (a) $\sigma_{ML} = \arg\max \prod_{i=1}^{N} p(t|x_i, t^{(i)}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2\sigma^2}(t^{(i)} - \beta_0 - \beta^T x_i)^2)$ Since ε is i.i.d then t is i.i.d

$$= \arg\max \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2}(t - \beta_0 - \beta^T x_i)^2$$

(b) Define $L(\sigma^2) = \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2}(t^{(i)} - \beta_0 - \beta^T x_i)^2$

$$\frac{\partial L(\sigma^2)}{\partial \sigma^2} = \sum_{i=1}^{N} \sqrt{2\pi\sigma^2}\left(-\frac{1}{2}\frac{1}{(\sqrt{2\pi\sigma^2})^3}\right) 2\pi - \frac{(t^{(i)} - \beta_0 - \beta^T x_i)^2}{2}\left(-\frac{1}{\sigma^4}\right)$$

$$= \sum_{i=1}^{N} -\frac{\pi}{2\pi\sigma^2} + \frac{(t^{(i)} - \beta_0 - \beta^T x_i)^2}{2}\frac{1}{\sigma^4}$$

$$= -\frac{N}{2\sigma^2} + \frac{1}{\sigma^4}\frac{\sum_{i=1}^{N}(t^{(i)} - \beta_0 - \beta^T x_i)^2}{2}$$

Set $\frac{\partial L}{\partial \sigma}$ to be 0

$$\frac{N}{2\sigma_{ML}^2} = \frac{1}{2\sigma_{ML}^4}\sum_{i=1}^{N}(t^{(i)} - \beta_0 - \beta^T x_i)^2$$

$$\sigma_{ML}^2 = \frac{1}{N}\sum_{i=1}^{N}(t^{(i)} - \beta_0 - \beta^T x_i)^2$$

4. Assume that we have K class with target value 1 to K
   All data information is stored in <u>data</u> variable and target information is stored in <u>target</u> variable

```
X1min = np.min(data[:,0])
X1max = np.max(data[:,0])
X min = np.min(data[:,1])
X2max = np.max(data[:,1])

X1 = np.linspace(X1min, X1max, 50)
X2 = np.linspace(X2min, X2max, 50)
XX1, XX2 = np.meshgrid(X1, X2)
Xprediction = np.hstack((XX1.reshape(-1,1), XX2.reshape(-1,1)))

from sklearn.linear_model import LinearRegression
prediction = np.zeros((len(XX1.flatten()), K-1))
```

```
for k in np.arange(K-1):
    indices_class_k = np.where(target == k)
    target_one_vs_rest = np.zeros((len(target), 1))
    target_one_vs_rest[indices_class_k] = 1
    reg = LinearRegression()
    reg.fit(data, target_one_vs_rest)
    prediction[:,k] = np.squeeze(reg.predict(data_prediction) > 1/2)

final_prediction = np.zeros((len(xx1.flatten()), 1))
for i in np.arange(len(final_prediction)):
    if np.argwhere(prediction[i,:] ==1).size == 1 :
        final_prediction[i]= np.argwhere(prediction[i,:] ==1)
    elif np.argwhere(prediction[i,:] ==1).size >1 :
        final_prediction[i]= K
    else:
        final_prediction[i] = K-1
```

Question 2

1. (1) False
   (2) True
   (3) True
   (4) False
   (5) False
   (6) False

2. (a) The normal equation is $X^T t = X^T X \beta$ where $X=\begin{bmatrix} \bar{x}_1^{S,(1)} & \bar{x}_2^{S,(1)} \\ \cdots & \cdots \\ \bar{x}_1^{S,(N)} & \bar{x}_2^{S,(N)} \end{bmatrix}$ $t=\begin{bmatrix} \bar{t}^{(1)} \\ \bar{t}^{(N)} \end{bmatrix}$ $\beta=\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$

$$\begin{bmatrix} \bar{x}_1^{S,(1)} & \cdots & \bar{x}_1^{S,(N)} \\ \bar{x}_2^{S,(1)} & \cdots & \bar{x}_2^{S,(N)} \end{bmatrix}\begin{bmatrix} \bar{t}^{(1)} \\ \bar{t}^{(N)} \end{bmatrix} = \begin{bmatrix} \bar{x}_1^{S,(1)} & \cdots & \bar{x}_1^{S,(N)} \\ \bar{x}_2^{S,(1)} & \cdots & \bar{x}_2^{S,(N)} \end{bmatrix}\begin{bmatrix} \bar{x}_1^{S,(1)} & \bar{x}_2^{S,(1)} \\ \cdots & \cdots \\ \bar{x}_1^{S,(N)} & \bar{x}_2^{S,(N)} \end{bmatrix}\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

$$\begin{bmatrix} \sum_{i=1}^{N} \bar{x}_1^{S,(i)} \bar{t}^{(i)} \\ \sum_{i=1}^{N} \bar{x}_2^{S,(i)} \bar{t}^{(i)} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N} \bar{x}_1^{S,(i)2} & \sum_{i=1}^{N} \bar{x}_1^{S,(i)} \bar{x}_2^{S,(i)} \\ \sum_{i=1}^{N} \bar{x}_2^{S,(i)} \bar{x}_1^{S,(i)} & \sum_{i=1}^{N} \bar{x}_2^{S,(i)2} \end{bmatrix}\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \Rightarrow \frac{1}{N}\begin{bmatrix} \sum_{i=1}^{N} \bar{x}_1^{S,(i)} \bar{t}^{(i)} \\ \sum_{i=1}^{N} \bar{x}_2^{S,(i)} \bar{t}^{(i)} \end{bmatrix} = \frac{1}{N}\begin{bmatrix} \sum_{i=1}^{N} \bar{x}_1^{S,(i)2} & \sum_{i=1}^{N} \bar{x}_1^{S,(i)} \bar{x}_2^{S,(i)} \\ \sum_{i=1}^{N} \bar{x}_2^{S,(i)} \bar{x}_1^{S,(i)} & \sum_{i=1}^{N} \bar{x}_2^{S,(i)2} \end{bmatrix}\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} 1 & \frac{1}{N}\sum_{i=1}^{N} \bar{x}_1^{S,(i)} \bar{x}_2^{S,(i)} \\ \frac{1}{N}\sum_{i=1}^{N} \bar{x}_1^{S,(i)} \bar{x}_2^{S,(i)} & 1 \end{bmatrix}\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N} \bar{x}_1^{S,(i)} \bar{t}^{(i)} \\ \sum_{i=1}^{N} \bar{x}_2^{S,(i)} \bar{t}^{(i)} \end{bmatrix}$$

Since $\sum_{i=1}^{N} \bar{x}_R^{S,(i)2} = \sum \frac{(x_R^{(i)} - \frac{1}{N}\sum_i x_R^{(i)})^2}{\sigma_R^2} = \frac{\sum (x_R^{(i)} - \frac{1}{N}\sum_i x_R^{(i)})^2}{\frac{1}{N}\sum (x_R^{(i)} - \frac{1}{N}\sum_i x_R^{(i)})^2} = N$

$$r_{12} = \frac{1}{N}\sum_{i=1}^{N} \bar{x}_1^{S,(i)} \bar{x}_2^{S,(i)} = \frac{1}{N}\sum_{i=1}^{N} \frac{x_1^{(i)} - \frac{1}{N}\sum_i x_1^{(i)}}{\sigma_1} \frac{x_2^{(i)} - \frac{1}{N}\sum_i x_2^{(i)}}{\sigma_2}$$

$$= \frac{1}{N\sigma_1\sigma_2}\sum_i \left(x_1^{(i)} - \frac{1}{N}\sum_i x_1^{(i)}\right)\left(x_2^{(i)} - \frac{1}{N}\sum_i x_2^{(i)}\right)$$

$$= \frac{1}{N\sigma_1\sigma_2}\sum_i \left(x_1^{(i)} x_2^{(i)} - \frac{x_1^{(i)}}{N}\sum_i x_2^{(i)} - \frac{x_2^{(i)}}{N}\sum_i x_1^{(i)} + \frac{1}{N^2}\sum_i x_1^{(i)}\sum_i x_2^{(i)}\right)$$

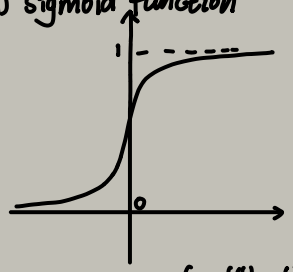
$$= \frac{1}{N\sigma_1\sigma_2}\left(\sum_i x_1^{(i)} x_2^{(i)} - \frac{1}{N}\sum_i x_2^{(i)}\sum_i x_1^{(i)}\right)$$

(b) $X^TX = \begin{bmatrix} \bar{x}_1^{s,(1)} & \cdots & \bar{x}_1^{s,(N)} \\ \bar{x}_2^{s,(1)} & \cdots & \bar{x}_2^{s,(N)} \end{bmatrix} \begin{bmatrix} \bar{x}_1^{s,(1)} & \bar{x}_2^{s,(1)} \\ \vdots & \vdots \\ \bar{x}_1^{s,(N)} & \bar{x}_2^{s,(N)} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N} \bar{x}_1^{s,(i)2} & \sum_{i=1}^{N} \bar{x}_1^{s,(i)} \bar{x}_2^{s,(i)} \\ \sum_{i=1}^{N} \bar{x}_2^{s,(i)} \bar{x}_1^{s,(i)} & \sum_{i=1}^{N} \bar{x}_2^{s,(i)2} \end{bmatrix} = N \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix}$

$(X^TX)^{-1} = \frac{N}{1-r_{12}^2} \begin{bmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{bmatrix}$

$\det(X^TX) \neq 0 \Leftrightarrow 1 - r_{12}^2 \neq 0$ to make X^TX invertible, which means $r_{12} \neq \pm 1$

3. (a) sigmoid function



(b) $y(x,w) = \sigma\left(w_{11}^{(4)} z_1^{(3)} + w_{12}^{(4)} z_2^{(3)} + w_{10}^{(4)}\right)$

where
$$\begin{cases} z_1^{(0)} = \sigma(x) \\ z_1^{(1)} = \sigma\left(w_{11}^{(1)} \sigma(x) + w_{10}^{(1)}\right) \\ z_2^{(1)} = \sigma\left(w_{21}^{(1)} \sigma(x) + w_{20}^{(1)}\right) \\ z_1^{(2)} = \sigma\left(w_{11}^{(2)} \sigma\left(w_{11}^{(1)} \sigma(x) + w_{10}^{(1)}\right) + w_{12}^{(2)} \sigma\left(w_{21}^{(1)} \sigma(x) + w_{20}^{(1)}\right) + w_{10}^{(2)}\right) \\ z_1^{(3)} = \sigma\left(w_{11}^{(3)} z_1^{(2)} + w_{10}^{(3)}\right) \\ z_2^{(3)} = \sigma\left(w_{21}^{(3)} z_1^{(2)} + w_{20}^{(3)}\right) \end{cases}$$

(c) $\delta_{out} = \sigma(a_{out}) - (1-t)$

$= \sigma\left(w_{11}^{(4)} z_1^{(3)} + w_{12}^{(4)} z_2^{(3)} + w_{10}^{(4)}\right) - (1-t)$

$\delta_{out} = \delta_1^{(4)}$

$$\delta_1^{(3)} = \sigma'(a_1^{(3)}) \sum_{j=1}^{N_4} \delta_j^{(4)} w_{j1}^{(4)} = \sigma'(a_1^{(3)}) \delta_1^{(4)} w_{11}^{(4)}$$
$$= \sigma(a_1^{(3)})\left(1 - \sigma(a_1^{(3)})\right) \delta_1^{(4)} w_{11}^{(4)}$$
$$= \sigma(w_{11}^{(3)} z_1^{(2)} + w_{10}^{(3)})\left(1 - \sigma(w_{11}^{(3)} z_1^{(2)} + w_{10}^{(3)})\right) \delta_1^{(4)} w_{11}^{(4)}$$

$$\delta_1^{(2)} = \sigma'(a_1^{(2)}) \sum_{j=1}^{N_4} \delta_j^{(3)} w_{j1}^{(3)} = \sigma'(a_1^{(2)})\left(\delta_1^{(3)} w_{11}^{(3)} + \delta_2^{(3)} w_{21}^{(3)}\right)$$
$$= \sigma(a_1^{(2)})\left(1 - \sigma(a_1^{(2)})\right)\left(\delta_1^{(3)} w_{11}^{(3)} + \delta_2^{(3)} w_{21}^{(3)}\right)$$
$$= \sigma\left(w_{11}^{(2)} \sigma\left(w_{11}^{(1)} \sigma(x) + w_{10}^{(1)}\right) + w_{12}^{(2)} \sigma\left(w_{21}^{(1)} \sigma(x) + w_{20}^{(1)}\right) + w_{10}^{(2)}\right) \cdot$$
$$\left(1 - \sigma\left(w_{11}^{(2)} \sigma\left(w_{11}^{(1)} \sigma(x) + w_{10}^{(1)}\right) + w_{12}^{(2)} \sigma\left(w_{21}^{(1)} \sigma(x) + w_{20}^{(1)}\right) + w_{10}^{(2)}\right)\right) \cdot$$
$$\left(\delta_1^{(3)} w_{11}^{(3)} + \delta_2^{(3)} w_{21}^{(3)}\right)$$

$$\delta_1^{(1)} = \sigma'(a_1^{(1)}) \sum_{j=1}^{N_2} \delta_j^{(2)} w_{j1}^{(2)} = \sigma'(a_1^{(1)}) \delta_1^{(2)} w_{11}^{(2)}$$
$$= \sigma(a_1^{(1)})\left(1 - \sigma(a_1^{(1)})\right) \delta_1^{(2)} w_{11}^{(2)}$$
$$= \sigma(w_{11}^{(1)} \sigma(x) + w_{10}^{(1)})\left(1 - \sigma(w_{11}^{(1)} \sigma(x) + w_{10}^{(1)})\right) \delta_1^{(2)} w_{11}^{(2)}$$

$$\frac{\partial L}{\partial w_{11}^{(1)}} = \delta_1^{(1)} z_1^{(0)} = \delta_1^{(1)} \sigma(x)$$

4. (a) $\beta_0 = 3$, $\beta_1 = 0.1$, $\beta_2 = 1$

In this case, all blue points have positive value and all red points have negative value

(b) The large coeffcients will be shrinked. Specifically, $\sum_{j=1}^{p} |\beta_j|^2$ will be restricted within a range. As a result, new β will increase the bias and decrese the variance.