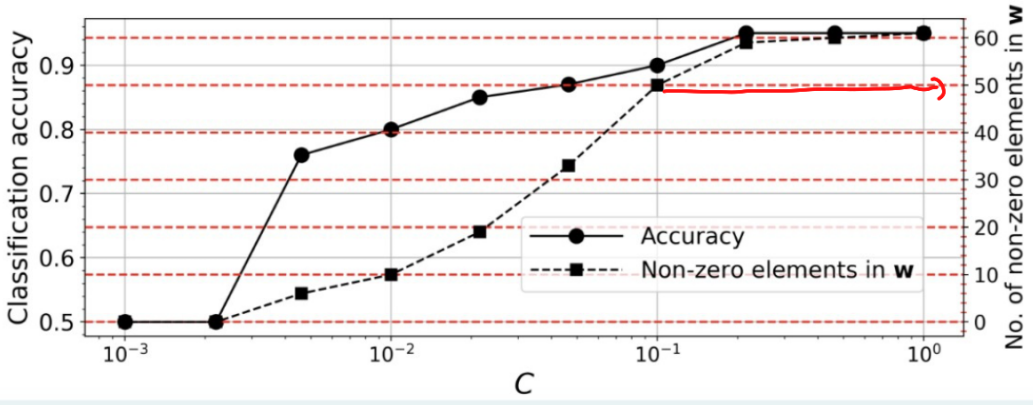
Binary class logistic regression with a regularization term, minimizes the following cost function

$$\mathscr{L}(\boldsymbol{w}) = C\left(\sum_{i=1}^{N}\left[-y_i \log(\mu_i) - (1-y_i)\log(1-\mu_i)\right]\right) + \|\boldsymbol{w}\|_1.$$

Here, $\mu_i = \text{sigm}(\boldsymbol{w}^T \boldsymbol{x}_i)$ with features $\boldsymbol{x}_i \in \mathbb{R}^{64\times 1}$, where sigm$(\cdot)$ is the sigmoid function. Figure illustrates the classification accuracy and the number of non-zero elements in the coefficient vector $(\boldsymbol{w})$ with respect to different C values.

What is the percentage increase in features $(\boldsymbol{x})$ required to change classification accuracy from 80% to 90%? **[5 marks]**



- a. 66.66%
- b. 62.50%
- c. 33.33%
- d. 75.00%

$$\frac{50-10}{64} \times 100$$

Consider a dataset consisting of three flower classes: Iris Setosa, Iris Versicolor, and Iris Virginica. In this dataset each data sample is represented by a 4-dimensional vector. After the learning process, a linear classifier is given as follows:

$$\boldsymbol{W} = \begin{bmatrix} 0.41 & 1.46 & -2.26 & -1.02 \\ 0.42 & -1.61 & 0.57 & -1.40 \\ -1.70 & -1.53 & 2.47 & 2.55 \end{bmatrix}^T, \text{ and } \boldsymbol{b} = \begin{bmatrix} 0.26 & 1.09 & -1.21 \end{bmatrix}^T.$$

Suppose that two data samples $\boldsymbol{x}_1$ and $\boldsymbol{x}_2 \in \mathbb{R}^{4\times 1}$ are fed to this linear classifier. $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are given by

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 \end{bmatrix} = \begin{bmatrix} 2 & 3 & 3 & 1 \\ 6 & 1 & 4 & 2 \end{bmatrix}^T \in \mathbb{R}^{4\times 2}.$$

Output of the linear classifier for $i-$th data sample is given by $\boldsymbol{y}_i = f\left(\boldsymbol{W}^T\boldsymbol{x}_i + \boldsymbol{b}\right) \in \mathbb{R}^{3\times 1}$. Here, $f(\cdot)$ is the sigmoid function. Suppose that probability of $i-$th data sample belongs to $j-$th class denoted by $P_{j,i}$. What are the probabilities of both data samples? **[10 marks]**

a.

Index (i)	$P_{1,i}$	$P_{2,i}$	$P_{3,i}$
1	0.4	0.2	0.4
2	0.3	0.2	0.5

b.

Index (i)	$P_{1,i}$	$P_{2,i}$	$P_{3,i}$
1	0.1047	0.0832	0.8121
2	0.0006	0.479	0.5204

c.

Index (i)	$P_{1,i}$	$P_{2,i}$	$P_{3,i}$
1	0.0223	0.2351	0.4576
2	0.0215	0.852	0.0212

d.

Index (i)	$P_{1,i}$	$P_{2,i}$	$P_{3,i}$
1	0.0879	0.0698	0.6814
2	0.001	0.8146	0.8849

$W^T x + b$

$$\begin{bmatrix} 0.41 & 1.46 & -2.26 & -1.02 \\ 0.42 & -1.61 & 0.57 & -1.40 \\ -1.70 & -1.53 & 2.47 & 2.55 \end{bmatrix} \begin{bmatrix} 2 & 6 \\ 3 & 1 \\ 3 & 4 \\ 1 & 2 \end{bmatrix} + \begin{bmatrix} 0.26 & 0.26 \\ 1.09 & 1.09 \\ -1.21 & -1.21 \end{bmatrix}$$

$$\begin{bmatrix} -2.6 & -7.16 \\ -3.68 & 0.39 \\ 1.97 & 3.25 \end{bmatrix} + \begin{bmatrix} 0.26 & 0.26 \\ 1.09 & 1.09 \\ -1.21 & -1.21 \end{bmatrix}$$

$$\begin{bmatrix} -2.34 & -6.9 \\ -2.59 & 1.48 \\ 0.76 & 2.04 \end{bmatrix} \xrightarrow{\text{sigmoid}} \begin{bmatrix} 0.087 & 0.001 \\ 0.0698 & 0.8146 \\ 0.6814 & 0.885 \end{bmatrix}$$

Consider the linear regression model of $y(\boldsymbol{x}_i) = w_0 + w_1 x_{1,i} + w_2 x_{2,i}$ and loss function with ℓ_2 regularization as

$$L(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\left(y_i - y(\boldsymbol{x_i})\right)^2 + \lambda\|\boldsymbol{w}\|_2^2.$$

Suppose that for the j-th iteration $w_0^{(j)}$, $w_1^{(j)}$ and $w_2^{(j)}$ are given by 1, 1.5, 1.5, respectively. Find the value of w_2 for (j+1)-th iteration using stochastic gradient descent (SDG) algorithm with a learning rate of 0.2 and regularization parameter $\lambda = 0.1$. Here, the data sample is used for the SGD is $y_i = 6$, $x_{1,i} = 2$, and $x_{2,i} = 2$, respectively. **[10 marks]**

- a. 1.04
- b. 1.1
- c. 1.7
- d. 1.55

Clear my choice

$$y(x_i) = 1 + 1.5x_1 + 1.5x_2$$

$$x_1 = 2 \qquad x_2 = 2$$

$$y(x_i) = 1 + 1.5 \times 2 + 1.5 \times 2 = 7$$

$$L(w) = \frac{1}{2}\left(y_i - w_0 - w_1x_1 - w_2x_2\right)^2 + \lambda\left(w_0^2 + w_1^2 + w_2^2\right)$$

$$\frac{\partial L}{\partial w_2} = \left(y_i - w_0 - w_1x_1 - w_2x_2\right)(-x_2) + 2\lambda w_2$$

$$\frac{\partial L}{\partial w_2} = \left(6 - y(x_i)\right)(-2) + 2(0.1)(1.5)$$

$$\frac{\partial L}{\partial w_2} = (6-7)(-2) + 0.3 = 2.3$$

$$w_2^{new} = w_2^{old} - \alpha\frac{\partial L}{\partial w_2}$$

$$w^{new} = 1.5 - 0.2 \times 2.3$$

$$w^{new} = 1.04$$

Index (i)	1	2	3	4	5	6	7	8	9	10
$x_{1,i}$	1	4	6	9	7	-1	-14	-15	-13	-7
$x_{2,i}$	9	0	-1	4	3	3	-8	-13	-17	-6
Class label y_i	c_1	c_1	c_1	c_1	c_1	c_1	c_2	c_2	c_2	c_2

class 1 class 2

Table displays feature values x_1 and x_2 of data samples which are belongs to two classes namely c_1 and c_2. Note that feature values are rounded to nearest integer. Here, it is assumed that class-conditional densities are Gaussian distributed, i.e., $p(\boldsymbol{x}|y = c_k, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$. Calculate mean vectors $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and prior probabilities of classes ($p(y = c_1|\boldsymbol{\theta}) = \pi_1$ and $p(y = c_2|\boldsymbol{\theta}) = \pi_2$). **[10 marks]**

a. $\boldsymbol{\mu}_1 = [1.83\ 3.83]^T, \boldsymbol{\mu}_2 = [-9.75\ -10.75]^T, \pi_1 = 0.6$ and $\pi_2 = 0.4$

b. $\boldsymbol{\mu}_1 = [4.33\ 3.0]^T, \boldsymbol{\mu}_2 = [-12.25\ -11.0]^T, \pi_1 = 0.4$ and $\pi_2 = 0.6$

c. $\boldsymbol{\mu}_1 = [4.33\ 3.0]^T, \boldsymbol{\mu}_2 = [-12.25\ -11.0]^T, \pi_1 = 0.5$ and $\pi_2 = 0.5$

d. $\boldsymbol{\mu}_1 = [4.33\ 3.0]^T, \boldsymbol{\mu}_2 = [-12.25\ -11.0]^T, \pi_1 = 0.6$ and $\pi_2 = 0.4$

Clear my choice

$$\mu_1 = \left[\frac{1+4+6+9+7-1}{6}, \frac{9+0-1+4+3+3}{6} \right]^T$$

$$\mu_1 = [4.33, 3]^T$$

$$\mu_2 = \left[\frac{-14-15-13-7}{4}, \frac{-8-13-17-6}{4} \right]^T$$

$$\mu_2 = [-12.25, -11]^T$$

$$\pi_1 = 0.6$$

$$\pi_2 = 0.4$$

A data set of four data samples are given in table. Suppose that regression model $f(\boldsymbol{x}) = w_0 + w_1\left(\frac{x_1}{x_2}\right)^2$ fits to the given data set.

Use all the data samples to find parameters of the regression model (w_0 and w_1). **[10 marks]**

Sample index (i)	$x_{1,i}$	$x_{2,i}$	y_i
1	3	3	5.5
2	4	2	11.5
3	3	1	21
4	16	4	36

- a. $w_0 : 3.30$ and $w_1 : 2.03$
- b. $w_0 : 5.54$ and $w_1 : 2.10$
- c. $w_0 : -0.35$ and $w_1 : 5.28$
- d. $w_0 : 5.11$ and $w_1 : 10.19$
- e. $w_0 : -4.16$ and $w_1 : 3.02$

Clear my choice

$$f(x) = w_0 + w_1 z_1$$

$$z = \left(\frac{x_1}{x_2}\right)^2 \qquad z = \begin{bmatrix} 1 & 1 \\ 1 & 4 \\ 1 & 9 \\ 1 & 16 \end{bmatrix}$$

$$\hat{w} = (z^T z)^{-1} z^T y$$

$$\hat{w} = \left(\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 4 & 9 & 16 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 4 \\ 1 & 9 \\ 1 & 16 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 4 & 9 & 16 \end{bmatrix} \begin{bmatrix} 5.5 \\ 11.5 \\ 21 \\ 36 \end{bmatrix}$$

$$\hat{w} = \begin{bmatrix} 3.296 \\ 2.027 \end{bmatrix}$$

Q10. A data set of three data samples are given in table Q10. Suppose that linear regression model $f(\boldsymbol{x}) = w_0 + w_1 x_1 x_2$ fits to the given data set. Use all the data samples to find parameters of the linear regression model (w_0 and w_1). **[10 marks]**

Table Q10: Data set for **Q10**.

Sample index (i)	$x_{1,i}$	$x_{2,i}$	y_i
1	$\frac{1}{5}$	5	2
2	3	$\frac{2}{3}$	4
3	3	1	6

$f(\boldsymbol{x}) = y = w_0 + w_1 x_1 x_2 = w_0 + w_1 z$, where $z = x_1 x_2$. Now, this becomes a linear regression model.

$\hat{\boldsymbol{w}} = \left(\boldsymbol{Z}^T \boldsymbol{Z}\right)^{-1} \boldsymbol{Z}^T \boldsymbol{y}$. Here, $\boldsymbol{Z} = \begin{bmatrix} 1 & z_1 \\ 1 & z_2 \\ 1 & z_3 \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1}x_{2,1} \\ 1 & x_{1,2}x_{2,2} \\ 1 & x_{1,3}x_{2,3} \end{bmatrix}$.

$$\hat{\boldsymbol{w}} = \left(\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \end{bmatrix}.$$

$w_0 = 0$ and $w_1 = 2$.

Consider the linear model given by $z_i = w_0 + w_1 x_i$. where $w_1 = 2$ and $w_0 = -5$. The output of this linear model is then mapped to a probabilistic output using the function below:

$$p\left(y_i = \text{Class A} \middle| x_i, \boldsymbol{w}\right) = \left(\frac{f(z_i)+1}{2}\right), \text{ with } f(z_i) = \left(\frac{e^{z_i} - e^{-z_i}}{e^{z_i} + e^{-z_i}}\right).$$

Here, y_i is the class label of $i-$th data sample and there are only two possible classes to which a data sample can belong: Class A and Class B.
Choose correct answer for the input value $x_i = 2.3$. **[10 marks]**

- a. The probability that x_i belongs to class B is 0.31
- b. The probability that x_i belongs to class B is 0.5987
- c. The probability that x_i belongs to class B is 0.69
- d. The probability that x_i belongs to class B is 0.7311
- e. The probability that x_i belongs to class B is 0.4013
- f. The probability that x_i belongs to class A is 0.69

Clear my choice

$$z_i = -5 + 2x_i$$

$$z_i = -5 + 2 \times 2.3$$

$$z_i = -0.4$$

$$f(z_i) = f(-0.4)$$

$$f(z_i) = \frac{e^{-0.4} - e^{0.4}}{e^{-0.4} + e^{0.4}}$$

$$f(z_i) = -0.38$$

$$P(A) = \frac{1 - 0.38}{2} = 0.31$$

$$P(B) = 1 - 0.31 = 0.69$$

	Email spam detection				Credit card fraud detection			
	True Pos.(TP)	False Pos.(FP)	True Neg.(TN)	False Neg.(FN)	TP	FP	TN	FN
Method A	970	30	930	70	850	150	900	100
Method B	960	40	940	60	900	100	850	150

You have two machine learning algorithms, namely method A and method B, which are trained and tested on two scenarios: credit card fraud detection and email spam detection. Calculate recall and precision values for each scenario based on the data given in the Table. Based on these results, which machine learning method would you choose for each scenario? **[05 marks]**

- a. Credit card fraud detection: Method B with precision=0.9 and recall=0.85, Email spam detection: Method B with precision=0.96 and recall=0.94
- b. Credit card fraud detection: Method B with precision=0.9 and recall=0.85, Email spam detection: Method A with precision=0.97 and recall=0.93
- c. Credit card fraud detection: Method A with precision=0.85 and recall=0.89, Email spam detection: Method A with precision=0.97 and recall=0.93
- d. Credit card fraud detection: Method A with precision=0.95 and recall=0.79, Email spam detection: Method A with precision=0.87 and recall=0.83
- e. Credit card fraud detection: Method A with precision=0.85 and recall=0.89, Email spam detection: Method B with precision=0.96 and recall=0.94

$$\text{Accuracy} = \frac{\text{TP} + TN}{\text{TP} + \text{TN} + \text{FP} + FN}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + FP}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$F_1 = \frac{2}{\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}\right)}$$

email spam detection

	precision	recall
A	$\frac{970}{1000}$	$\frac{970}{1040}$
	0.97	0.93
B	$\frac{960}{1000}$	$\frac{960}{1020}$
	0.96	0.94

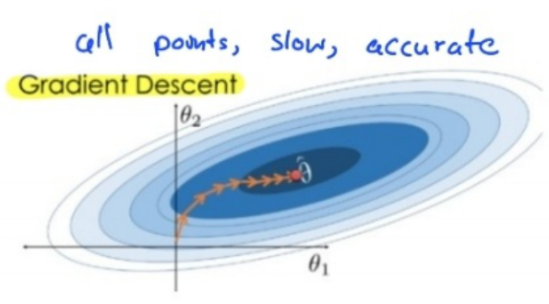
credit card fraud detection

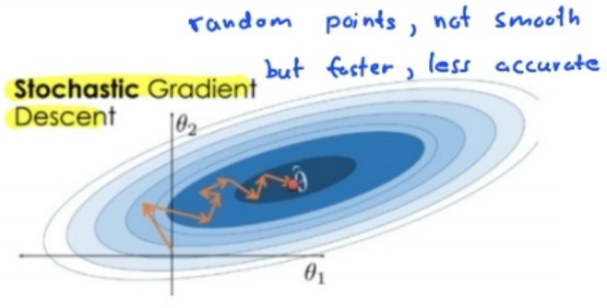
	precision	recall
A	$\frac{850}{1000}$	$\frac{850}{950}$
	0.85	0.89
B	$\frac{900}{1000}$	$\frac{900}{1050}$
	0.9	0.86

Suppose you have a huge dataset and you have time constrains. Here, which variation of the gradient descent algorithm would you select?

- ○ a. Batch Gradient Descent
- ○ b. Stochastic Gradient Descent

- **Batch Gradient Descent**
  - Calculate the gradient (derivative) of the objective function with respect to all the training data points
  - Slow but more accurate
- **Stochastic Gradient Descent (SGD)**
  - SGD uses only one randomly selected data point at a time
  - Faster convergence
  - Can be noisy and might converge to a suboptimal solution
- **Mini-batch Gradient Descent:**
  - Computes the gradient using a small randomly selected subset (mini-batch) of the dataset





In stochastic gradient descent (SGD), unlike gradient descent (GD), there's no guarantee that the objective function will decrease with each step. But gradually objective function is decreasing

Also, SGD is more sensitive to step size compared to GD In simple terms, stochastic gradient descent utilizes stochastic gradients

Consider the following data set ($\boldsymbol{X}$), which consists of five samples of ten distinct features. Which scaling methods is appropriate for preserving the structure of this data set? **[05 marks]**

$$\boldsymbol{X} = \begin{bmatrix} 0 & 0 & 0.5 & 0 & 200 & 3 & 0 & 0.1 & 0 & -100 \\ 4 & 0 & 0 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.75 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 5 \\ 0.1 & 0 & 0 & 0 & 2 & 12 & 0 & -1000 & 1000 & 0 \end{bmatrix}$$

- a. Max-abs scaling
- b. Robust scaling
- c. Standard scaling
- d. Min-max scaling

Clear my choice

for sparse graphs → max-abs scaling is prefered.

Model	Bias	Variance	Irreducible error
Model 1	0.6	0.3	0.2
Model 2	−0.5	0.5	0.2

Table provides the bias, variance, and irreducible error of two machine learning models for a given data set. Based on this information, which model would you choose? **[05 marks]**

- a. Both Model 1 and 2
- b. Model 1
- c. Cannot be determined

$MSE = bias^2 + variance + irreducible\ error$

$MSE_1 = 0.6^2 + 0.3 + 0.2 = 0.86$

$MSE_2 = (-0.5)^2 + 0.5 + 0.2 = 0.95$

$MSE_1 < MSE_2$

choose model 1. It has low error.

Suppose that both dependent and independent variables have measurement errors. Which algorithm would you choose? **[05 marks]**

- a. Ordinary least-squares (OLS)
- b. Ridge Regression
- c. Least Absolute Shrinkage and Selection Operator (LASSO)
- d. Total least squares (TLS)
- e. Polynomial Regression

Clear my choice

Advantages of generative classifiers over discriminative classifiers?

- a. Generative classifiers always faster to train
- b. Generative classifiers always achieve better accuracy
- c. Generative classifiers can be used with with missing data/unlabeled data

Clear my choice

Advantages of generative classifiers

- Ease of fitting
- Handling missing features
- Class-specific learning
- Can handle unlabeled data
- Robustness to spurious features

Figure shows data samples of three distinct classes. Here, it is assumed that class conditional densities are Gaussian distributed. Which of the following statements are true?

[05 marks]

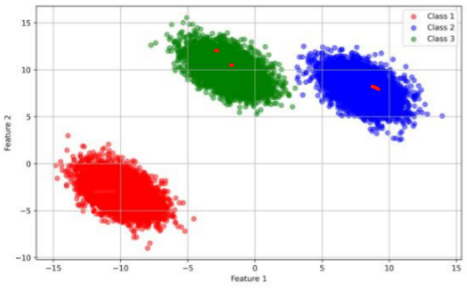


Figure A

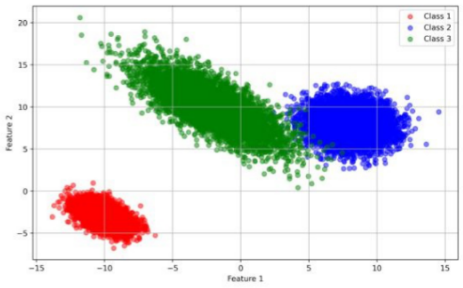


Figure B

Same covariance matrix

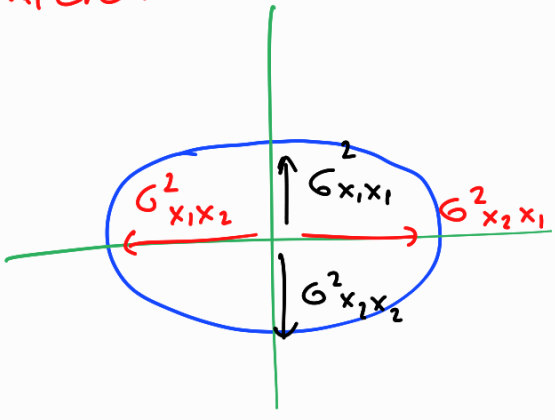
different mean vectors

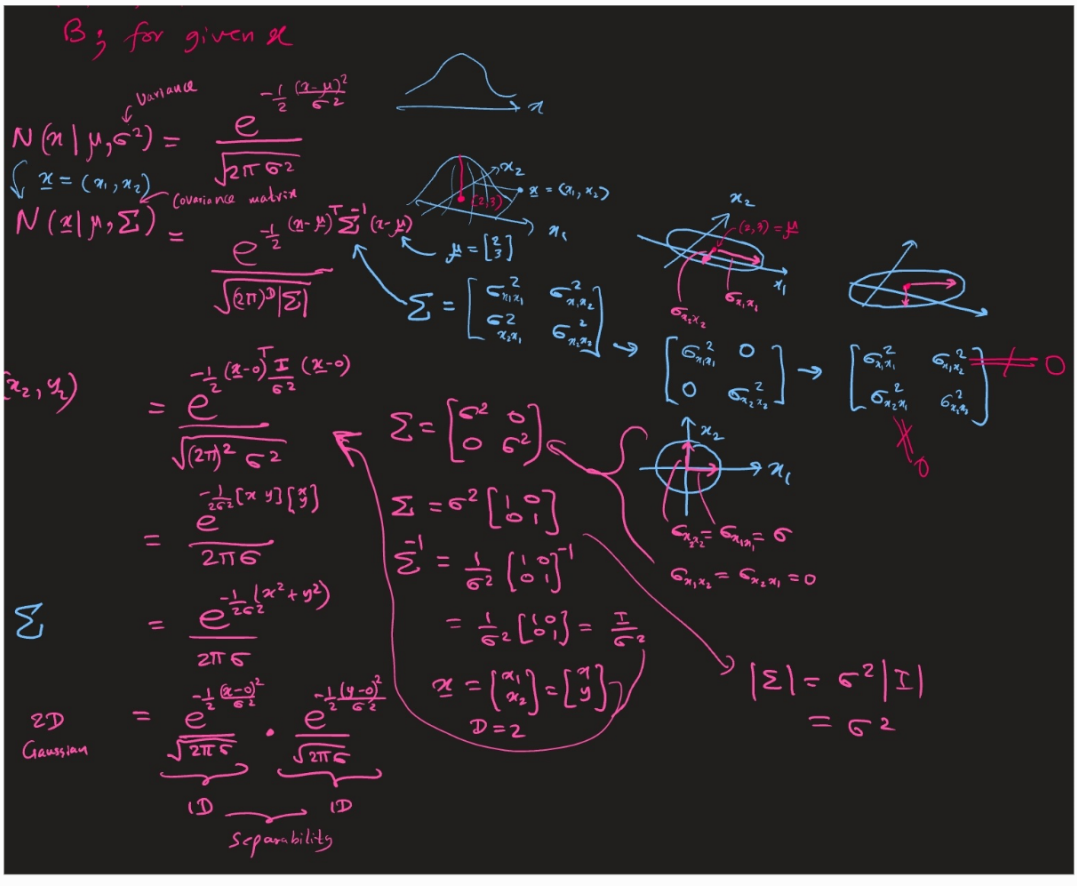
different covariance matrices

different mean vectors

$$\Sigma = \begin{bmatrix} \sigma^2_{x_1x_1} & \sigma^2_{x_1x_2} \\ \sigma^2_{x_2x_1} & \sigma^2_{x_2x_2} \end{bmatrix}$$

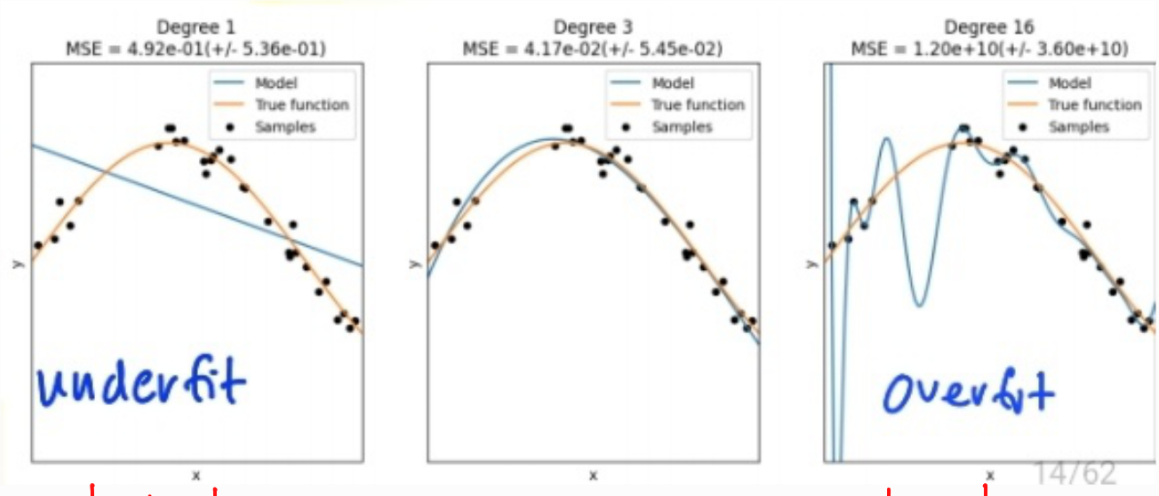
depend on shape





too simple model

highly complex model



high bias

low variance

low bias

high variance

- Overfitting: Overfitting occurs when a model performs exceptionally well on the training data but fails to generalize to new, unseen data.
- Underfitting: When a model is **too simplistic** to capture the underlying patterns in the data.

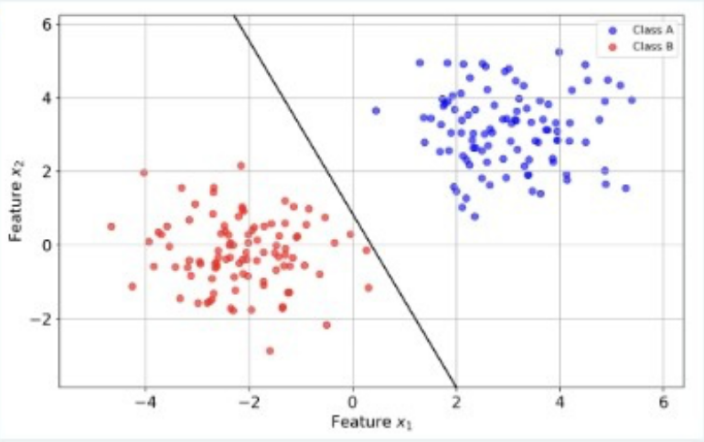
Binary class logistic regression with a regularization term, minimizes the following custom cost function

$$\mathscr{L}(\boldsymbol{w}) = \sum_{i=1}^{N} \left[-y_i \log(\mu_i) - (1 - y_i) \log(1 - \mu_i) \right] + \lambda w_0^2.$$

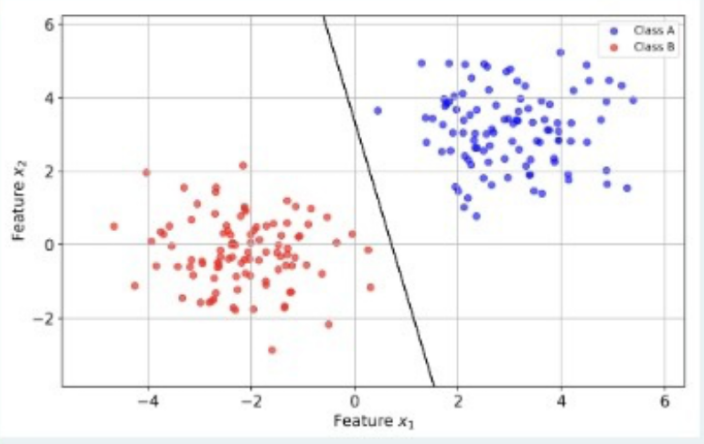
Here, $\mu_i = \text{sigm}(w_0 + w_1 x_{1,i} + w_2 x_{2,i})$, where $\text{sigm}(\cdot)$ is the sigmoid function and λ is the regularization parameter. Suppose λ is a very large number, i.e., $\lambda \to \infty$. Choose a possible decision boundary represented by the black line in the figure, which results from minimizing the loss function $\mathscr{L}(\boldsymbol{w})$.

[05 marks]

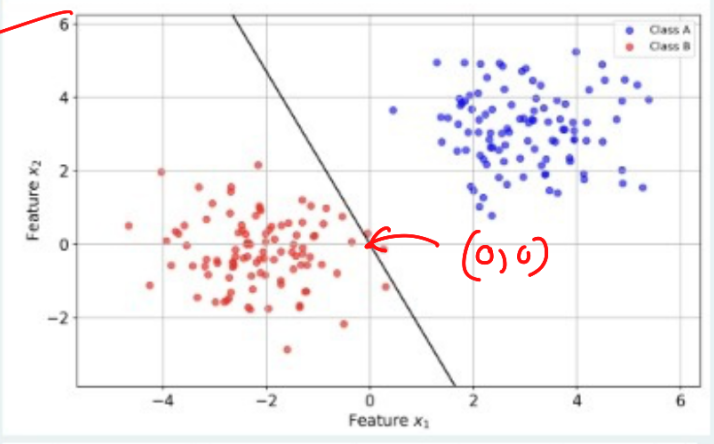
a.



b.



c.



d.