

# EMS702P Statistical Thinking and Applied Machine Learning

## Week 6.2 – Linear classification

Yunpeng Zhu



## **Linear regression**

©Copyright 2022 Yunpeng Zhu. All Rights Reserved

Edition: v1.1

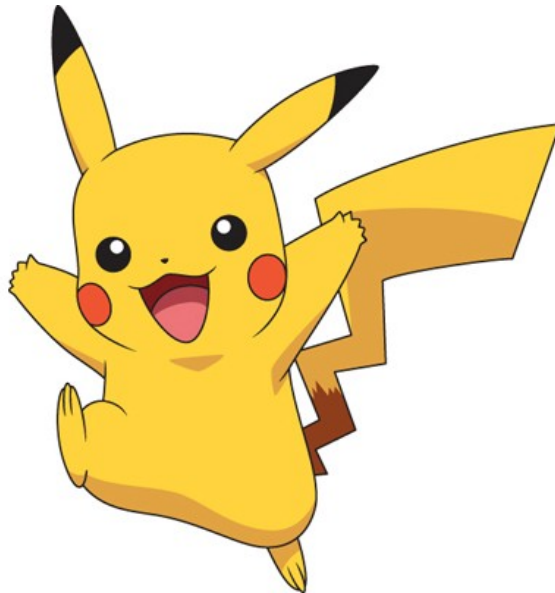
## Table of Contents

<b>1</b>	<b>Differentiation .....</b>	<b>- 1 -</b>
<b>2</b>	<b>Introduction to linear classification.....</b>	<b>- 1 -</b>
<b>3</b>	<b>The binary (2-class) classification .....</b>	<b>- 2 -</b>
3.1	Classification based on polynomial basis functions .....	- 2 -
3.2	Classification based on logistic basis functions .....	- 4 -
<b>4</b>	<b>Gradient descent based logistic model regression.....</b>	<b>- 5 -</b>
4.1	The gradient descent method .....	- 5 -
4.2	The logistic regression model.....	- 7 -
<b>5</b>	<b>The K-class classification .....</b>	<b>- 9 -</b>
5.1	One-versus-the-rest classifier .....	- 10 -
5.2	One-versus-one classifier .....	- 10 -
5.3	K-class discriminant classifier .....	- 10 -
<b>6</b>	<b>Further Readings.....</b>	<b>- 13 -</b>

# 1 Differentiation

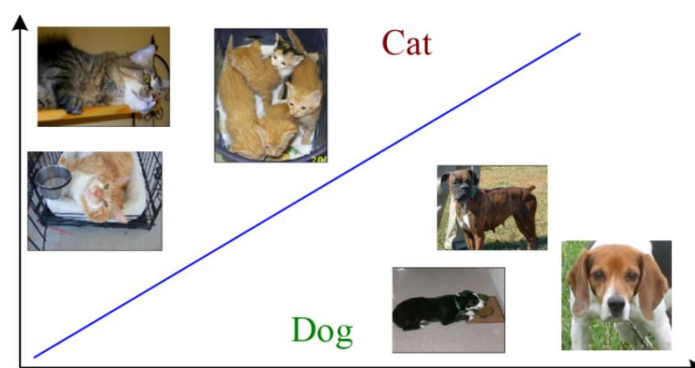
Constant Rule	$\frac{d}{dx} [C] = 0$
Power Rule	$\frac{d}{dx} x^n = nx^{n-1}$
Product Rule	$\frac{d}{dx} [f(x)g(x)] = f'(x)g(x) + f(x)g'(x)$
Quotient Rule	$\frac{d}{dx} \left[ \frac{f(x)}{g(x)} \right] = \frac{g(x)f'(x) - f(x)g'(x)}{[g(x)]^2}$
Chain Rule	$\frac{d}{dx} [f(g(x))] = f'(g(x)) g'(x)$

Quiz 1.1:



# 2 Introduction to linear classification

Classification: Separate different categories [1].



Denote Cat (True): 1, Dog (False): -1, how to use a straight line to separate the cats and dogs?

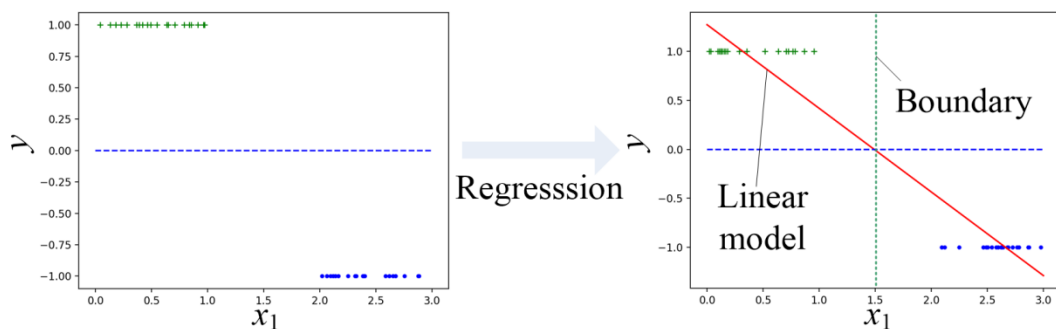
### 3 The binary (2-class) classification

#### 3.1 Classification based on polynomial basis functions

Assuming we have **1 feature** ( $x_1$ ) characterizing the cats and dogs, the samples are labeled either true (1) or false (-1):

$$y = \begin{cases} 1 & \text{True} \\ -1 & \text{False} \end{cases}$$

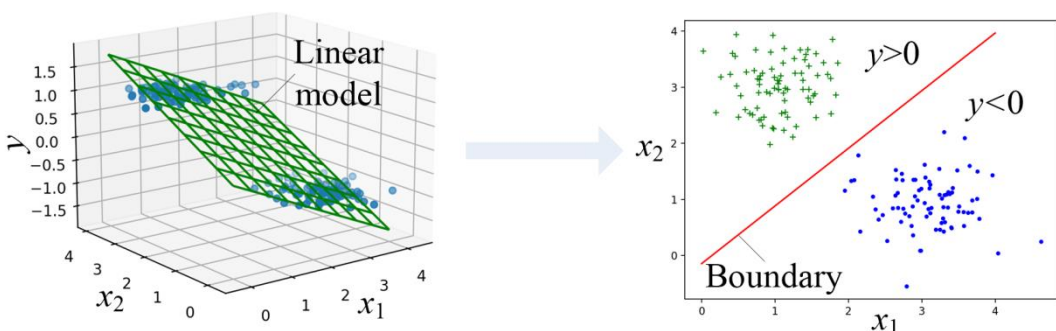
where  $y$  represents the label.



The linear classifier is  $y = w_0 + w_1x_1$ . This can be obtained by using **linear regression** approaches. The decision boundary is the vertical line across the point at  $w_0 + w_1x_1 = 0$ .

Assuming we have 2 features ( $x_1, x_2$ ) characterizing the cats and dogs, the samples are labeled either true (1) or false (-1):

$$y = \begin{cases} 1 & \text{True} \\ -1 & \text{False} \end{cases}$$



The linear classifier is  $y = w_0 + w_1x_1 + w_2x_2$ . This can be obtained by using **linear regression** approaches. The decision boundary is the line across the plane at  $w_0 + w_1x_1 + w_2x_2 = 0$ .

**The linear classifier is**

$$y = w_0 + w_1x_1 + w_2x_2 + \cdots + w_nx_n = \mathbf{x}^T \mathbf{W}$$

$1 \times M \quad M \times 1$

where  $\mathbf{x} = [1, x_1, \cdots, x_n]^T$  is a the vector of basis functions,  $\mathbf{W} = [w_0, w_1, \cdots, w_n]^T$  is the parameter vector.

**The decision boundary is**

$$\mathbf{x}^T \mathbf{W} = 0$$

### Quiz 3.1:

Consider we have 4 sets of observed data

$$(x, y) = (0,1), (0.5,1), (1.5,0), (2,0)$$

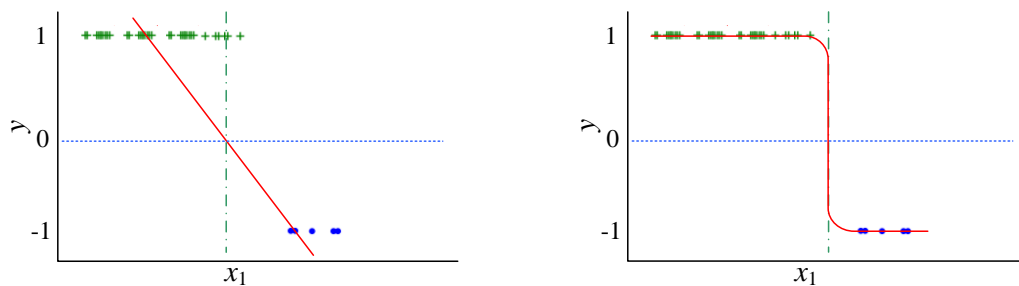
Evaluate the classifier as a linear polynomial function:

$$y = 1.5 + ax$$

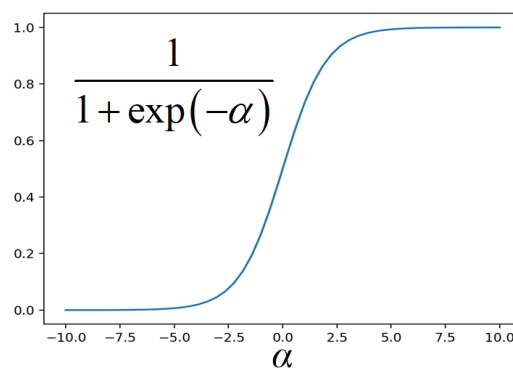


The boundary of the classification is  $1.5 - 0.846x = 0.5 \Rightarrow x = 1.18$

### 3.2 Classification based on logistic basis functions



Logistic sigmoid function:



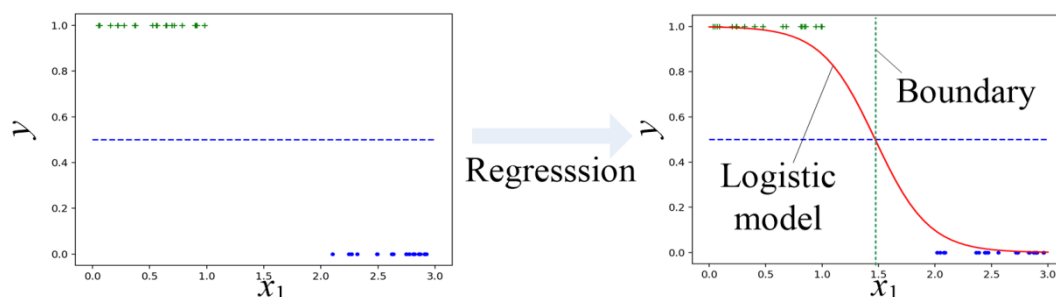
Redefine the label function as

$$y = \begin{cases} 1 & \text{True} \\ 0 & \text{False} \end{cases}$$

- 0,1 label is commonly used in classification. **Discuss** how to apply 0,1 label in the binary classification using polynomial basis functions.

Consider the case of characterizing the cats and dogs with **1 feature** ( $x_1$ ), the samples are labeled either true (1) or false (0):

$$y = \begin{cases} 1 & \text{True} \\ 0 & \text{False} \end{cases}$$



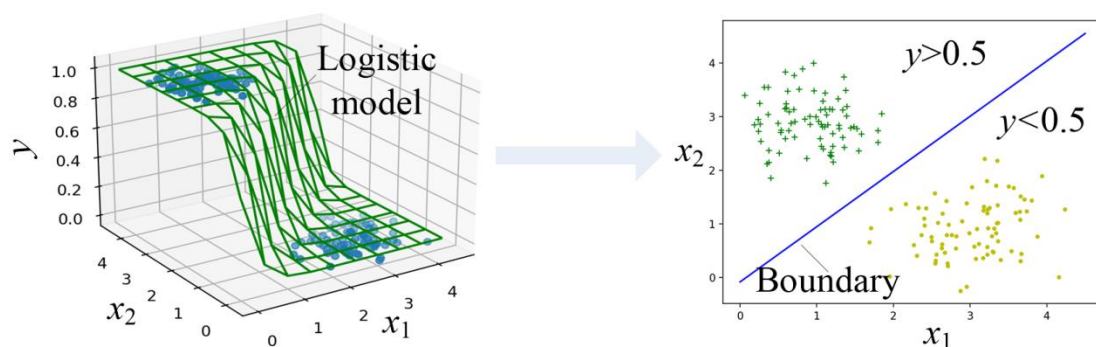
The logistic classifier is

$$y = \frac{1}{1 + \exp[-(w_0 + w_1 x_1)]}$$

and the decision boundary is the vertical line across the point at  $y = 0.5$  or  $w_0 + w_1 x_1 = 0$ .

Consider there are **2 features** ( $x_1, x_2$ ), the samples are labeled either true (1) or false (0):

$$y = \begin{cases} 1 & \text{True} \\ 0 & \text{False} \end{cases}$$



The logistic classifier is

$$y = \frac{1}{1 + \exp[-(w_0 + w_1 x_1 + w_2 x_2)]}$$

and the decision boundary is the vertical line across the point at  $y = 0.5$  or  $w_0 + w_1 x_1 + w_2 x_2 = 0$ .

**The logistic classifier is**

$$y = \sigma(\mathbf{x}^T \mathbf{W}) = \frac{1}{1 + \exp(-\mathbf{x}^T \mathbf{W})}$$

**The decision boundary is**

$$\mathbf{x}^T \mathbf{W} = 0$$

## 4 Gradient descent based logistic model regression

### 4.1 The gradient descent method

The cost function can be defined as

$$J(\mathbf{W}) = \begin{cases} \sum_{i=1}^N (y(i) - \bar{y}(i))^2 & \text{Least squares} \\ -\ln \prod_{i=1}^N P(y(i) | \mathbf{W}) & \text{Maximum likelihood} \end{cases}$$

where  $\bar{y}(i)$  is the predicted value of the regression model.

The gradient descent method is applied to **minimize** the cost function in regression problems [2].

- Gradient: The direction of steepest ascent (vector)

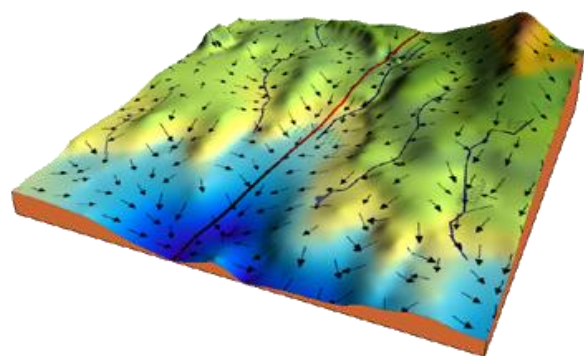
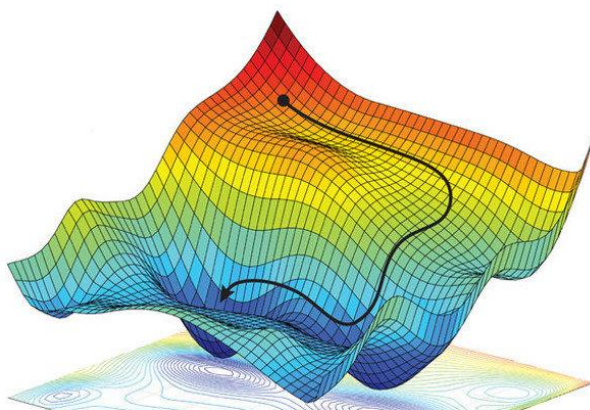
$$\nabla J(\mathbf{W}) = \left[ \frac{\partial J(\mathbf{W})}{\partial w_0}, \dots, \frac{\partial J(\mathbf{W})}{\partial w_n} \right]$$

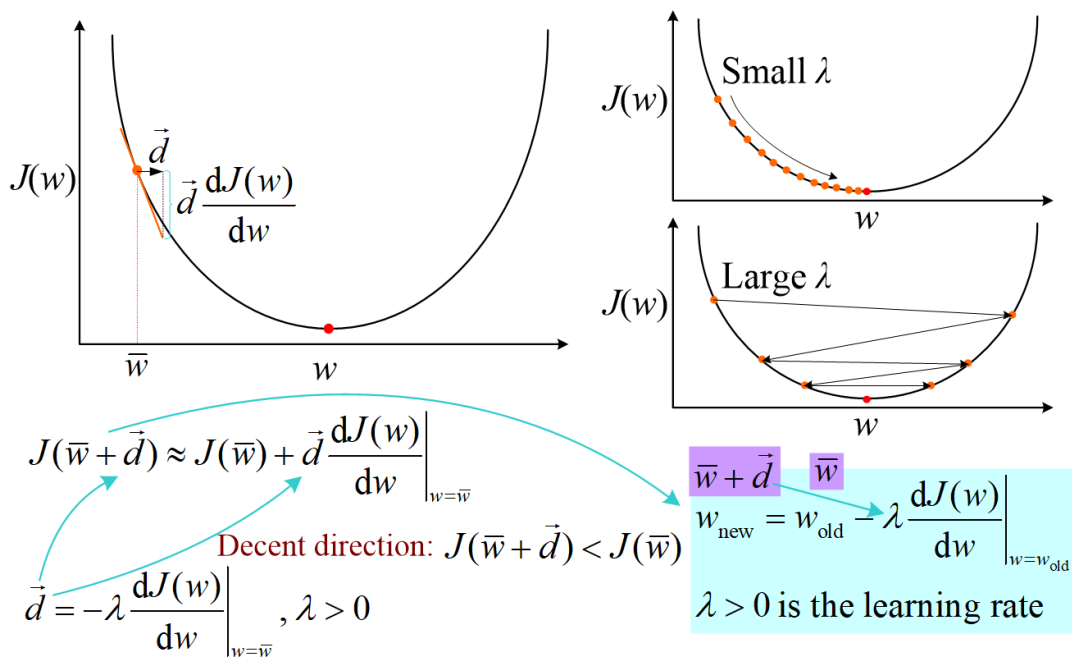
- Derivative: The rate of change (number)
- For single real function, Derivative is Gradient

### Quiz 3.2:

Considering  $y = w_0 + w_1x_1 + w_2x_2^2$ , evaluate the gradient of  $y$

Gradient of the function is





### The gradient descent method for regression

$$\mathbf{W}_{\text{new}} = \mathbf{W}_{\text{old}} - \lambda \nabla J(\mathbf{W}_{\text{old}})$$

#### Quiz 3.3:

Find the minimum of  $y = 1 + 2x + 2x^2$  by using the gradient descent method.

The learning step is  $\lambda = 0.2$ , starting from  $x = 0$ .



## 4.2 The logistic regression model

The value of the logistic sigmoid function is between 0 and 1, and the shape is similar to the **cumulative distribution function** representing the accumulation of **probabilities**.

Denote the cost function as [3]

(Consider why not using MSE as the loss function?)

Inaccurate  
Non-convex

$$J(\mathbf{W}) = -\ln \prod_{i=1}^N \bar{y}(i)^{y(i)} [1 - \bar{y}(i)]^{1-y(i)}$$

$$= -\sum_{i=1}^N \{y(i) \ln \bar{y}(i) + [1 - y(i)] \ln [1 - \bar{y}(i)]\}$$

where

$$\bar{y}(i) = \frac{1}{1 + \exp(-\mathbf{x}_i^T \mathbf{W})}; y(i) = 0, 1$$

By using the gradient descent method,

$$w_m = w_m - \lambda \frac{\partial J(\mathbf{W})}{\partial w_m}, m = 0, \dots, n$$



$$\mathbf{W} = \mathbf{W} - \lambda \mathbf{X}^T (\bar{\mathbf{Y}} - \mathbf{Y})$$

$$\mathbf{W}_{M \times 1} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{bmatrix}, \mathbf{X}_{N \times M} = \begin{bmatrix} \mathbf{1} \\ \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_N^T \end{bmatrix}, \bar{\mathbf{Y}}_{N \times 1} = \begin{bmatrix} \bar{y}(1) \\ \vdots \\ \bar{y}(N) \end{bmatrix}, \mathbf{Y}_{N \times 1} = \begin{bmatrix} y(1) \\ \vdots \\ y(N) \end{bmatrix}$$

**Derivation:** See supplementary material 'Determine Logistic model' on QM+.

**Example: Logistic regression by using the gradient decent method**

$$\begin{aligned} [x_1(1), y(1)] &= [0, 1] \\ [x_1(2), y(2)] &= [3, 0] \end{aligned} \quad \rightarrow \quad y = \sigma(w_0 + w_1 x_1) = \frac{1}{1 + \exp[-(w_0 + w_1 x_1)]}$$

$$J(\mathbf{W}) = - \sum_{i=1}^2 \left\{ y(i) \ln \sigma(w_0 + w_1 x_1(i)) + [1 - y(i)] \ln [1 - \sigma(w_0 + w_1 x_1(i))] \right\}$$

$$\begin{cases} \frac{\partial J(\mathbf{W})}{\partial w_0} = \sum_{i=1}^2 [\sigma(w_0 + w_1 x_1(i)) - y(i)] \\ \frac{\partial J(\mathbf{W})}{\partial w_1} = \sum_{i=1}^2 [\sigma(w_0 + w_1 x_1(i)) - y(i)] x_1(i) \end{cases}$$

$\lambda = 0.1$ , Initial  $\mathbf{W} = \mathbf{0}$

**Step 1:**

$$\begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} - \lambda \begin{bmatrix} \overset{i=1}{1} & \overset{i=2}{3} \end{bmatrix} \begin{bmatrix} \sigma(0) - 1 \\ \sigma(0) - 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 1 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 0.5 - 1 \\ 0.5 - 0 \end{bmatrix} = \begin{bmatrix} 0 \\ -0.15 \end{bmatrix}$$

**Step 2:**

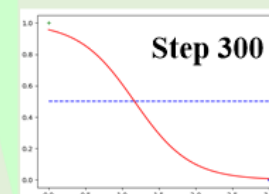
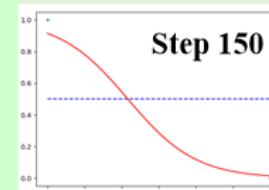
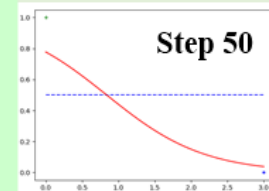
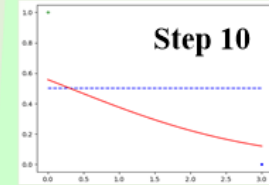
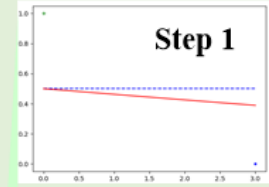
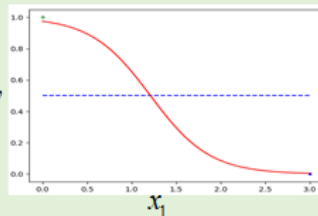
$$\begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} 0 \\ -0.15 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 1 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} \sigma(-0.15 \times 0 + 0) - 1 \\ \sigma(-0.15 \times 3 + 0) - 0 \end{bmatrix} = \begin{bmatrix} 0.01 \\ -0.27 \end{bmatrix}$$

$\vdots$

**Step 500:**

$$\begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} 3.64 \\ -3.01 \end{bmatrix}$$

$\rightarrow y$



## 5 The K-class classification

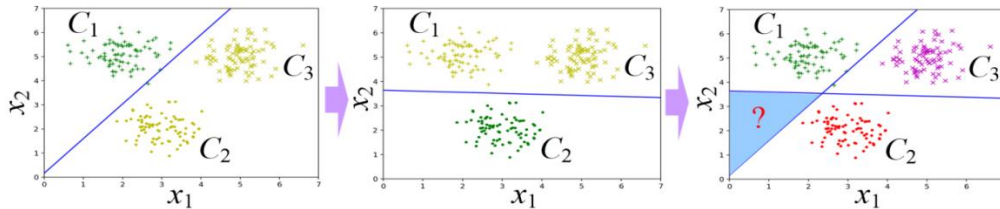
We have learned how to solve 2-classes problems by using linear regression and logistic regression approaches.

Apply the 0, 1 label for classification, how to extend the approaches to solve K-classes problems with  $K > 2$ ?

## 5.1 One-versus-the-rest classifier

Use  $K-1$  classifiers, each of which solves a 2-class problem of separating points in a particular class  $C_k$  ( $y=1$ ) from points not in that class ( $y=0$ ).

For example, consider  $K=3$ , the classification results with 2 features are:

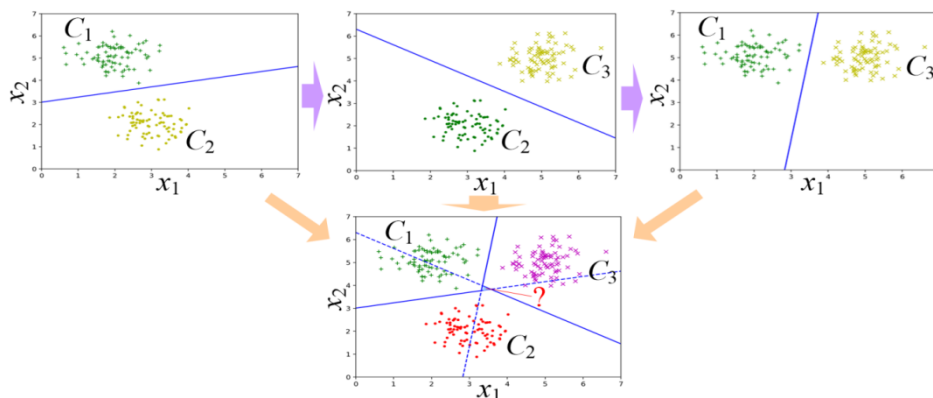


It can be seen that in one-versus-the-rest classification,  $K-1$  classifiers are applied to classify  $K$  groups. There will be a region of feature space that is ambiguously classified.

## 5.2 One-versus-one classifier

Use  $K(K-1)/2$  classifiers, each of which solves a 2-class problem for every possible pair of classes.

For example, consider  $K=3$ , the classification results with 2 features are:



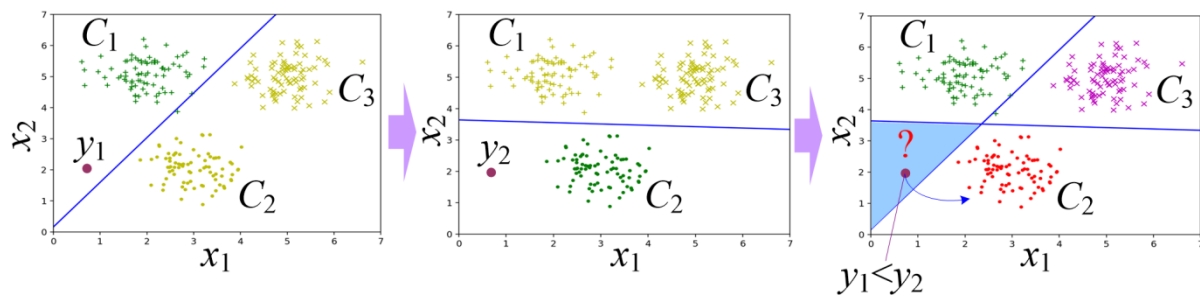
It can be seen that in one-versus-one classification. There is still a region of feature space that is ambiguously classified.

## 5.3 K-class discriminant classifier

The  $k$ th linear classifier (one-versus-the-rest):

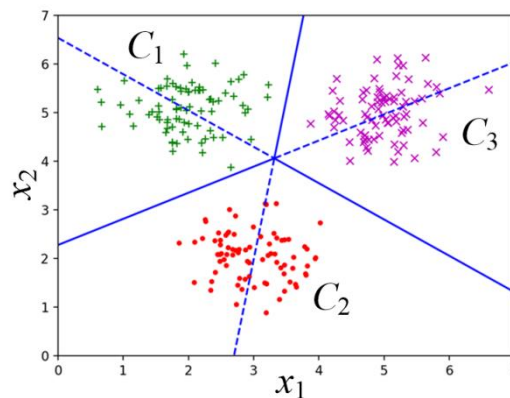
$$y_k = \mathbf{x}^T \mathbf{W}_k$$

For any  $\bar{\mathbf{x}} = [x_1, x_2]^T$ , if  $y_k > y_j$  for all  $j \neq k$ , assigning a point  $\bar{\mathbf{x}}$  to class  $C_k$ .



The decision boundaries are given by  $y_k = y_j$  as

$$\mathbf{x}^T (\mathbf{W}_k - \mathbf{W}_j) = 0$$



### Quiz 3.4:

Separate the three points:  $(x_1, x_2) = (1, 1), (0, -1), (-1, 0)$  by using the K-class discriminant classifier:

$$\begin{cases} y_1 = w_1^{(1)} x_1 + w_2^{(1)} x_2 \\ y_2 = w_1^{(2)} x_1 + w_2^{(2)} x_2 \\ y_3 = w_1^{(3)} x_1 + w_2^{(3)} x_2 \end{cases}$$



$$\begin{bmatrix} w_1^{(1)} \\ w_2^{(1)} \end{bmatrix} = \left( \begin{bmatrix} 1 & 1 \\ 0 & -1 \\ -1 & 0 \end{bmatrix}^T \times \begin{bmatrix} 1 & 1 \\ 0 & -1 \\ -1 & 0 \end{bmatrix} \right)^{-1} \times \begin{bmatrix} 1 & 1 \\ 0 & -1 \\ -1 & 0 \end{bmatrix}^T \times \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

$$= \left( \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \right)^{-1} \times \begin{bmatrix} 1 & 0 & -1 \\ 1 & -1 & 0 \end{bmatrix} \times \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \frac{1}{4-1} \times \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \times \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \frac{1}{3} \times \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$



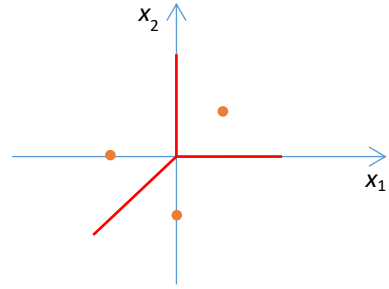
$$y_3 : \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} w_1^{(3)} \\ w_2^{(3)} \end{bmatrix} \Rightarrow$$

$$\begin{bmatrix} w_1^{(3)} \\ w_2^{(3)} \end{bmatrix} = \left( \begin{bmatrix} 1 & 1 \\ 0 & -1 \\ -1 & 0 \end{bmatrix}^T \times \begin{bmatrix} 1 & 1 \\ 0 & -1 \\ -1 & 0 \end{bmatrix} \right)^{-1} \times \begin{bmatrix} 1 & 1 \\ 0 & -1 \\ -1 & 0 \end{bmatrix}^T \times \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

$$= \left( \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \right)^{-1} \times \begin{bmatrix} 1 & 0 & -1 \\ 1 & -1 & 0 \end{bmatrix} \times \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \frac{1}{4-1} \times \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \times \begin{bmatrix} -1 \\ 0 \end{bmatrix} = \frac{1}{3} \times \begin{bmatrix} -2 \\ 1 \end{bmatrix}$$

$$y_3 = -\frac{2}{3}x_1 + \frac{1}{3}x_2$$

Boundaries:

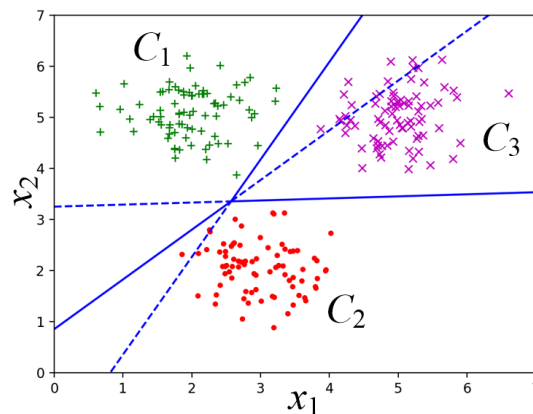


By using the logistic classifier:

$$y_k = \frac{1}{1 + \exp(-\mathbf{x}^T \mathbf{W}_k)}$$

The decision boundaries are given by  $y_k = y_j$  as

$$\mathbf{x}^T (\mathbf{W}_k - \mathbf{W}_j) = 0$$



## 6 Further Readings

[1] Linear classification.

<https://towardsdatascience.com/a-look-at-the-maths-behind-linear-classification-166e99a9e5fb>

[2] Gradient descent method.

<https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote07.html>

[3] Why not MSE as a loss function for logistic regression

<https://towardsdatascience.com/why-not-mse-as-a-loss-function-for-logistic-regression-589816b5e03c>