See Hastie and Tibshirani, chapters 2, 4.

7 LINEAR AND QUADRATIC CLASSIFICATION

7.1 INTRODUCTION TO CLASSIFICATION

☛ In the classification problem, we know that individuals from the population come from several, K say, different classes/groups

☛ Example when $K = 2$: healthy patients, unhealthy patients.

☛ We have at our disposal a training sample of individuals from the population for which we observe $(\mathbf{X}_1, G_1), \ldots, (\mathbf{X}_n, G_n)$, where

• $G_i = 1, 2, \ldots$ or K is the class/group label of the ith individual (an individual belong to exactly one group).

• $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T$ is vector of explanatory variables (e.g. age, blood pressure, etc).

prediction: By comparing that new individual with the sample of individuals that I given to study.

☛ A new individual from the same population comes in.

☛ For that individual we only observe the value $\mathbf{x} = (x_1, \ldots, x_p)^T$ of $\mathbf{X}$ but we do not know which group the individual comes from (we don't observe his/her G).

☛ Goal: classify the new individual in the correct group, i.e. find the group of that individual. Equivalently, find the class label $G \in \{1, \ldots, K\}$ of that new individual.

☛ How can we do that?

☛ Notation: throughout chapter we use the notation $(\mathbf{X}, G)$ for a generic individual from the population. The training data $(\mathbf{X}_1, G_1), \ldots, (\mathbf{X}_n, G_n)$ have the same distribution as a generic $(\mathbf{X}, G)$, and so do observed or unobserved data from new individuals.

☛ Main idea: compare the new value $\mathbf{x}$ with the $\mathbf{X}_i$'s from the training individuals.

☛ If $\mathbf{x}$ looks more similar to $\mathbf{X}_i$'s from group k than those from the other groups, classify new indiv in group k.

☛ How to decide if $\mathbf{x}$ is more similar to the $\mathbf{X}_i$'s from group k than from other groups?

☛ For $k = 1, \ldots, K$, use training data $(\mathbf{X}_1, G_1), \ldots, (\mathbf{X}_n, G_n)$ to estimate

$$P(G = k | \mathbf{X} = \mathbf{x}),$$

the probability that an individual comes from group k given that his/her X is equal to $\mathbf{x}$, by

$$\hat{P}(G = k | \mathbf{X} = \mathbf{x}) .$$

(We will see later how to compute the estimated probabilities).

☛ Then, classify new individual in group k if

$$\hat{P}(G = k|\mathbf{X} = \mathbf{x}) > \max_{j=1,\ldots,K, j\neq k} \hat{P}(G = j|\mathbf{X} = \mathbf{x}).$$

☛ Many different methods of classification exist. They essentially correspond to various ways of estimating the above probabilities.

☛ Two important classes of methods are based on regression estimation techniques and on so-called Bayes methods.

☛ We start by introducing methods in the case where we have only $K = 2$ classes. We will see later how to extend them to $K > 2$ classes.

- Suppose the individuals come from only $K = 2$ groups. Then we classify a new individual in group 1 if

$$\hat{P}(G = 1|\mathbf{X} = \mathbf{x}) > \hat{P}(G = 2|\mathbf{X} = \mathbf{x}),$$

where, for $k = 1, 2$, $\hat{P}(G = k|\mathbf{X} = \mathbf{x})$ is an estimator of $P(G = k|\mathbf{X} = \mathbf{x})$ computed from the training sample $(\mathbf{X}_1, G_1), \ldots, (\mathbf{X}_n, G_n)$.

- One way to obtain $\hat{P}(G = k|\mathbf{X} = \mathbf{x})$ is through regression.
- To see how, for $k = 1, 2$ define a variable from a generic G by taking

$$Y_k = I\{G = k\} = \begin{cases} 1 & \text{if } G = k \\ 0 & \text{otherwise.} \end{cases}$$

- At the sample level, for $i = 1, \ldots, n$, define

$$Y_{ik} = I\{G_i = k\}.$$

☛ Ex: suppose we observe a training sample of size $n = 7$ where the G_i's are

$$(G_1, \ldots, G_7) = (1, 1, 1, 2, 2, 2, 1).$$

Then, for $i = 1, \ldots, n$ and $k = 1, 2$, the Y_{ik}'s are given by:

$$(Y_{11}, \ldots, Y_{71}) = (1, 1, 1, 0, 0, 0, 1)$$
$$(Y_{12}, \ldots, Y_{72}) = (0, 0, 0, 1, 1, 1, 0)\,.$$

☛ Note that

$$P(G = k|\mathbf{X} = \mathbf{x}) = E\big(I\{G = k\}|\mathbf{X} = \mathbf{x}\big) = m_k(\mathbf{x})\,,$$

if we define the regression curve

$$m_k(\mathbf{x}) = E\big(Y_k|\mathbf{X} = \mathbf{x}\big).$$

☛ Can use regression estimation techniques to estimate m_k from training data $(\mathbf{X}_1, Y_{1k}), \ldots, (\mathbf{X}_n, Y_{nk})$ (Y_{ik}'s observed as computed from G_i's).

☛ Now every individual must belong to one of the two groups, so that for all $\mathbf{x} \in \mathbb{R}^p$

$$P(G = 1|\mathbf{X} = \mathbf{x}) + P(G = 2|\mathbf{X} = \mathbf{x}) = 1,$$

that is

$$m_2(\mathbf{x}) = 1 - m_1(\mathbf{x})\,.$$

Thus we only need to estimate m_1 (we can deduce m_2 from it).

Discrimination/classification boundary

☛ Once we have estimated m_1 and $m_2 = 1 - m_1$ by estimators $\hat{m}_1$ and $\hat{m}_2 = 1 - \hat{m}_1$, the classification boundary is the boundary between the region where we classify the data in group 1 and the region where we classify the data in group 2. It is obtained by solving the equation

$$\widehat{m}_1(\mathbf{x}) - \widehat{m}_2(\mathbf{x}) = 0$$

which, is equivalent to

$$\widehat{m}_1(\mathbf{x}) = 1/2\,.$$

Classify in group 1 if $\widehat{m}_1(\mathbf{x}) > \widehat{m}_2(\mathbf{x}) \iff \widehat{m}_1(\mathbf{x}) > 1/2$, otherwise classify in group 2.

7.3.1 LINEAR REGRESSION CLASSIFIER

- When we use a linear regression classifier, we assume

Regression curve:

$$m_1(\mathbf{x}) = E\big(Y_1|\mathbf{X}=\mathbf{x}\big) = P(G=1|\mathbf{X}=\mathbf{x}) = \beta_0 + \beta^T\mathbf{x}.$$

The probability to be in a class can be modelled by the linear regression

- Estimate m_1 using least squares (LS) estimator of β_0 and β (or PCA or PLS estimators if we need to reduce dimension, see page 260 for how to compute CV in classification) to find

$$\hat{m}_1(\mathbf{x}) = \hat{\beta}_0 + \hat{\beta}^T\mathbf{x}\,.$$

- Deduce $\hat{m}_2 = 1 - \hat{m}_1 = 1 - \hat{\beta}_0 - \hat{\beta}^T\mathbf{x}$.

- Classify new $\mathbf{x}$ in group 1 if

based on the assumption that a reasonable model for modelling this probability.

$$\hat{m}_1(\mathbf{x}) > \hat{m}_2(\mathbf{x}) \iff \hat{\beta}_0 + \hat{\beta}^T\mathbf{x} > 1 - \hat{\beta}_0 - \hat{\beta}^T\mathbf{x} \iff \hat{\beta}_0 + \hat{\beta}^T\mathbf{x} > 1/2\,;$$

otherwise classify $\mathbf{x}$ in group 2.

- NOTE: this method relies on the validity of the linear regression model and is usually only an approximation in real life.

Example: Golub data (ESL, Section 18.4).

Golub TR et al. (1999), Molecular Classification of cancer: class Discovery and Class Prediction by gene expression monitoring, Science 286:531-7.

- Two categories of patients, corresponding to two types of leukemia ("ALL" and "AML").

- $\mathbf{X} = (X_1, X_2)^T$: gene expressions for two genes known to be connected to leukemia type.

- If we fit two linear models $m_1(\mathbf{x}) = \beta_0 + \beta^T \mathbf{x}$ and $m_2(\mathbf{x}) = \beta_{0,2} + \beta_2^T \mathbf{x}$ by LS, we find

$$\widehat{m}_1(\mathbf{x}) = 0.9231 + 0.2454 x_1 - 0.4800 x_2$$
$$\widehat{m}_2(\mathbf{x}) = 0.07691 - 0.24542 x_1 + 0.47999 x_2 \approx 1 - \hat{m}_1(\mathbf{x})$$

as expected (the $\approx$ is due to numerical errors). Illustrates we don't need to fit m_2, just take $\hat{m}_2 = 1 - \hat{m}_1$

- For the Golub data, classification boundary obtained by solving $\widehat{m}_1(\mathbf{x}) = 1/2$ is

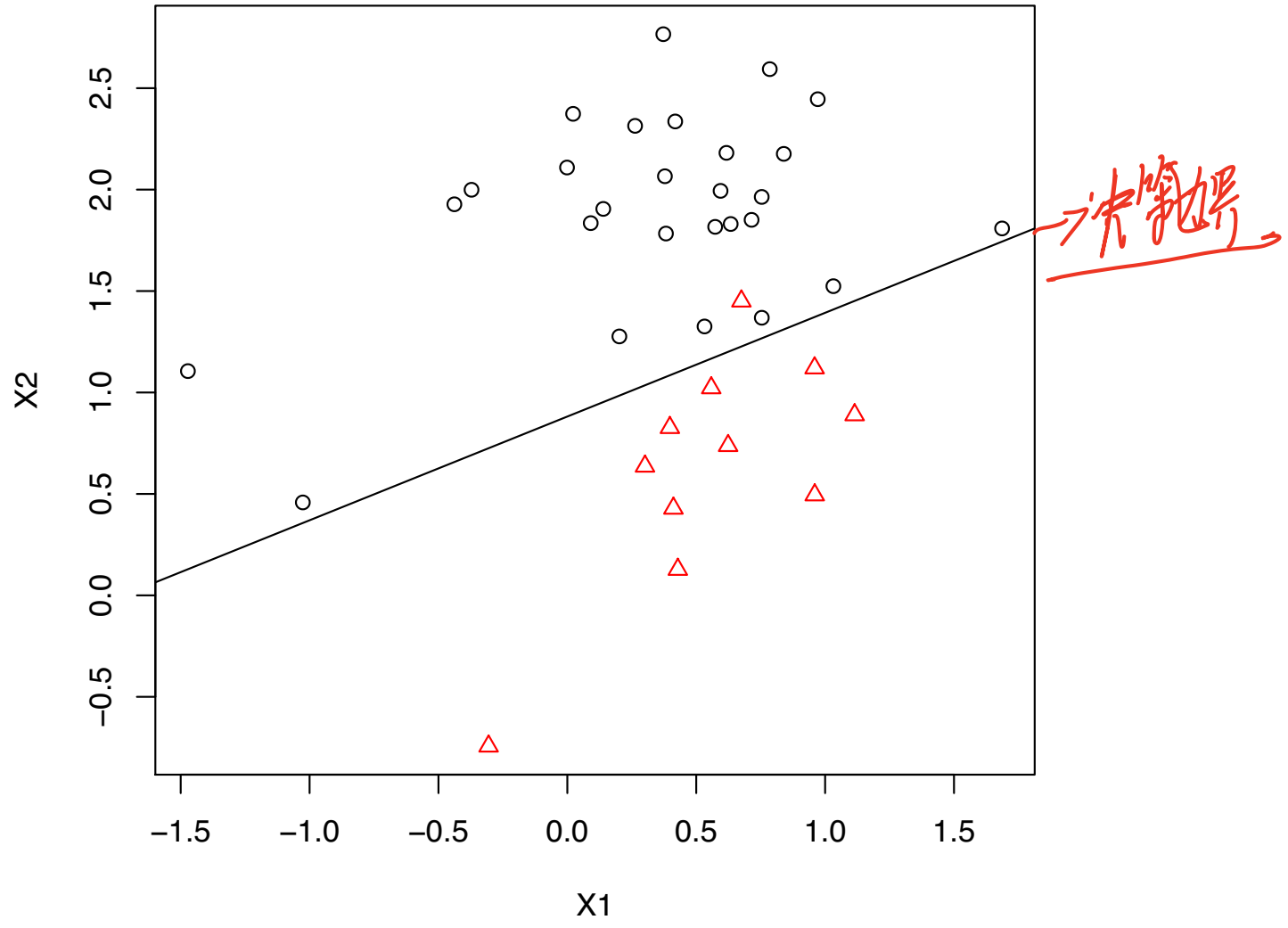
$$0.9231 + 0.2454x_1 - 0.4800x_2 = 1/2.$$

- This can be expressed by the line

$$x_2 = \frac{0.9231 - 0.5}{0.4800} + \frac{0.2454}{0.4800}x_1 = 0.8815 + 0.5113x_1 \, .$$

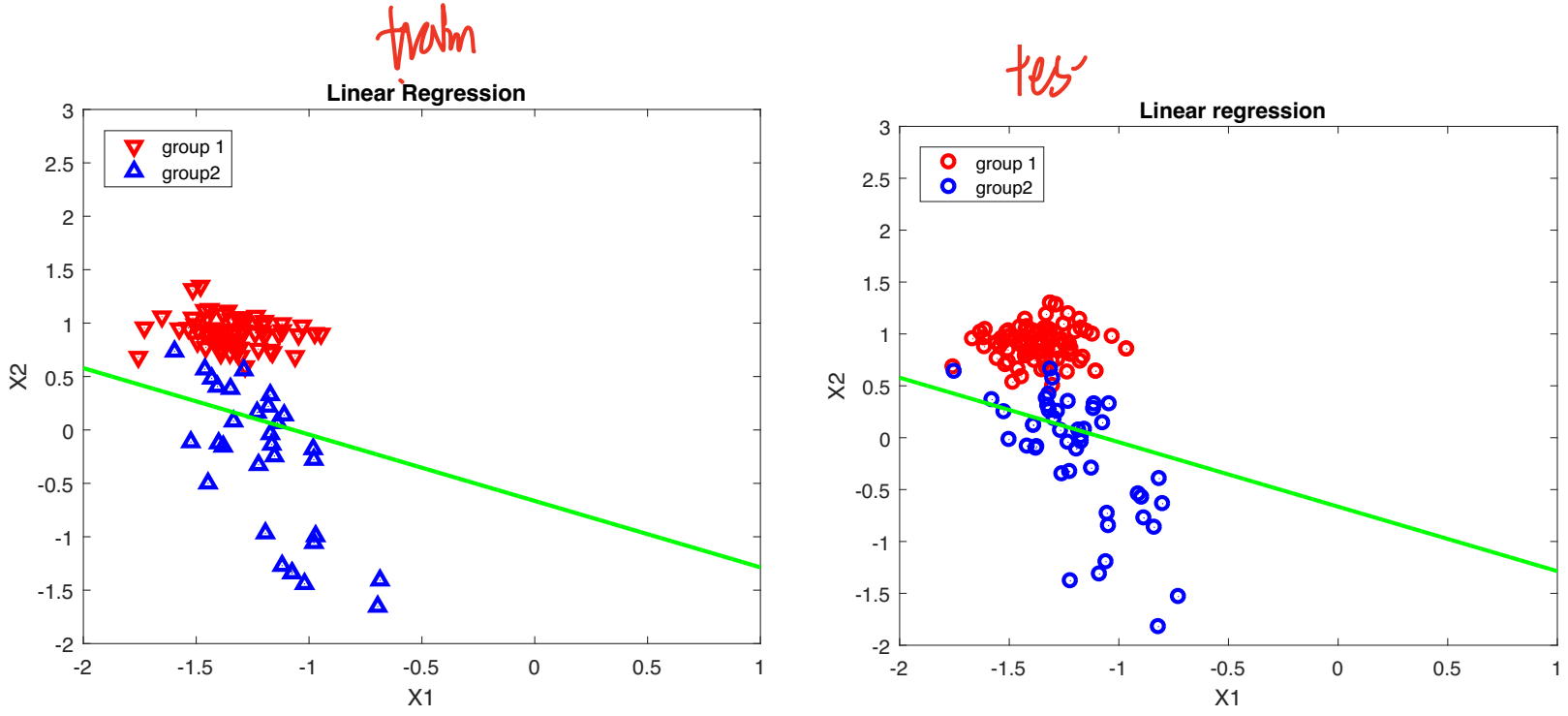
On one side of the line we classify in one class, on the other side we classify in the other class.

Discrimination boundary shown by the line. Training $\mathbf{X}_i$'s from two groups displayed in different colours/symbols. New observations falling below line will be classified in red group; others will be classified in black group.



Other example

- Left: training $\mathbf{X}_i$'s, different groups shown by blue/red.
- Right: set of new data to be classified
- Here it's just an illustration. Unlike real life we know the truth for the new data (blue/red on right fig): can check how classifier performs.
- Green line shows decision boundary $\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 = 1/2$ constructed from training $\mathbf{X}_i$'s.
- Data classified in group 2 if below line, in group 1 if above line (classifier makes mistakes: some blues are above line).



7.3.2 LOGISTIC REGRESSION CLASSIFIER

- A problem with the linear regression classifier is that $\hat{\beta}_0 + \hat{\beta}^T \mathbf{x}$, our estimator of $m_1(\mathbf{x}) = P(G = 1|\mathbf{X} = \mathbf{x})$, is not guaranteed to be between 0 and 1, whereas it estimates a probability.

- One way to ensure to be in $[0, 1]$ is use logistic regression, i.e. assume

$$m_1(\mathbf{x}) = P(G = 1|\mathbf{X} = \mathbf{x}) = E(Y_1|\mathbf{X} = \mathbf{x}) = \frac{\exp(\beta_0 + \beta^T \mathbf{x})}{1 + \exp(\beta_0 + \beta^T \mathbf{x})}.$$

- Estimate β_0, β using maximum likelihood (ML) estimators $\hat{\beta}_0, \hat{\beta}$ computed from the training data

$$(\mathbf{X}_1, Y_{11}), \ldots, (\mathbf{X}_n, Y_{n1}).$$

(or replace the data by their PCA or PLS components if needed, see page 260 for how to compute CV in classification).

- NOTE: here too, the method relies on the logistic model assumption which is usually only an approximation in real life.

- As noted at slide 234, $m_2 = 1 - m_1$, i.e. in this model:

$$m_1(\mathbf{x}) = P(G = 1|\mathbf{X} = \mathbf{x}) = \frac{\exp(\beta_0 + \beta^T \mathbf{x})}{1 + \exp(\beta_0 + \beta^T \mathbf{x})}$$
$$m_2(\mathbf{x}) = P(G = 2|\mathbf{X} = \mathbf{x}) = 1 - m_1(\mathbf{x}) = \frac{1}{1 + \exp(\beta_0 + \beta^T \mathbf{x})}.$$

- Classify new $\mathbf{x}$ in group 1 if

$$\hat{P}(G = 1|\mathbf{X} = \mathbf{x}) > \hat{P}(G = 2|\mathbf{X} = \mathbf{x})$$
$$\iff \exp(\hat{\beta}_0 + \hat{\beta}^T \mathbf{x}) > 1 \iff \hat{\beta}_0 + \hat{\beta}^T \mathbf{x} > 0$$

and otherwise classify $\mathbf{x}$ in group 2.

- As before, the classification boundary is obtained by solving

$$\widehat{m}_1(\mathbf{x}) = 1/2 \iff \exp(\hat{\beta}_0 + \hat{\beta}^T \mathbf{x}) = 1 \iff \hat{\beta}_0 + \hat{\beta}^T \mathbf{x} = 0.$$

Classify in group 1 if $\widehat{m}_1(\mathbf{x}) > \widehat{m}_2(\mathbf{x}) \iff \widehat{m}_1(\mathbf{x}) > 1/2 \iff \hat{\beta}_0 + \hat{\beta}^T \mathbf{x} > 0$, otherwise classify in group 2.

Example: Golub data

- Two categories of patients, corresponding to two types of leukemia ("ALL" and "AML"); $\mathbf{X} = (X_1, X_2)^T$: expressions of two genes.

- From the estimated logistic model we have

$$\widehat{m}_1(\mathbf{x}) = \frac{\exp\left(0.9432 + 0.5487x_1 - 10.714x_2\right)}{1 + \exp\left(0.9432 + 0.5487x_1 - 10.714x_2\right)}.$$

- In this example the classification boundary is obtained by solving

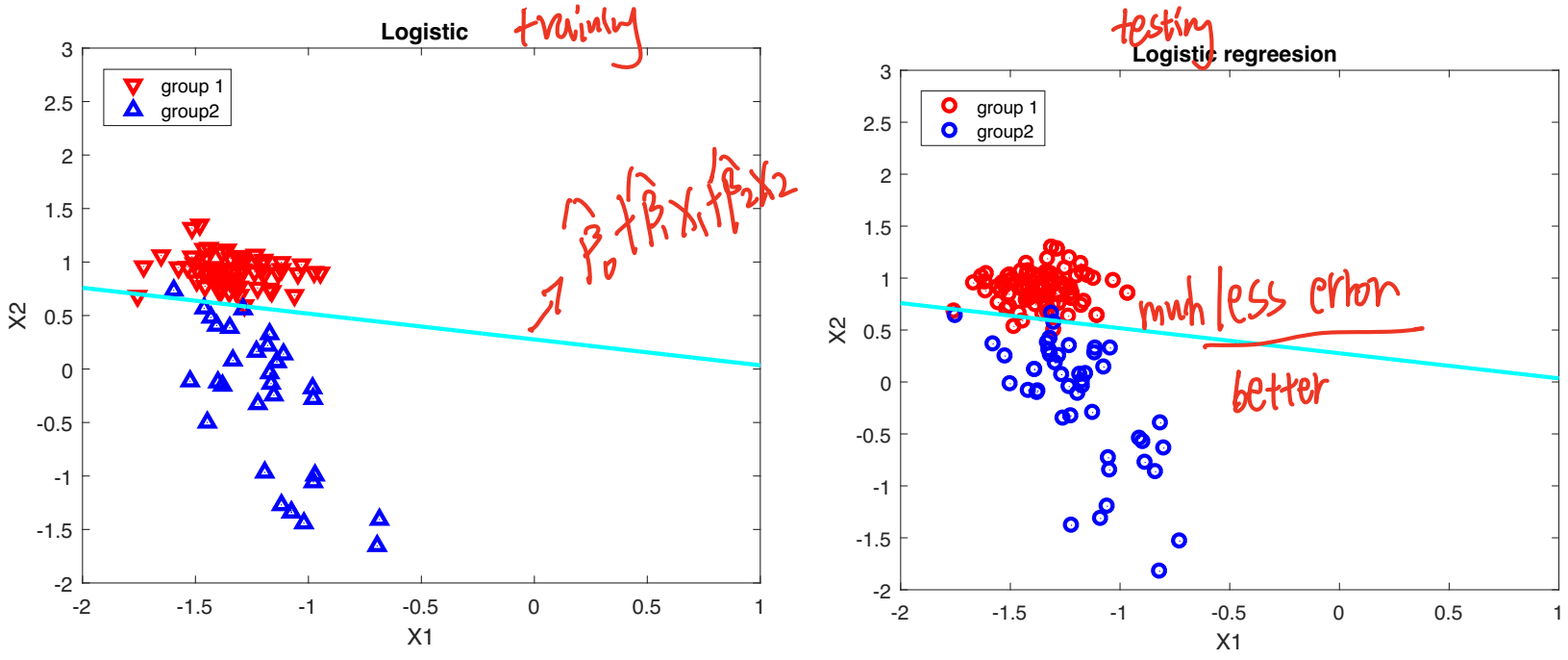
$$0.9432 + 0.5487x_1 - 10.714x_2 = 0,$$

which we can express by the line

$$x_2 = \frac{0.9432}{10.714} + \frac{0.5487}{10.714}x_1 = 0.08803435 + 0.05121337x_1$$

Other example (same as page 239)

- Left: training $\mathbf{X}_i$'s, different groups shown by blue/red.
- Right: set of new data to be classified
- Here unlike real life we know the truth for the new data (blue/red on right fig): can check how classifier performs.
- Cyan line shows decision boundary $\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 = 0$ constructed from training $\mathbf{X}_i$'s.
- Data classified in group 2 if below line, in group 1 if above line (classifier makes mistakes: some blues are above line).



Comparison linear and logistic regression