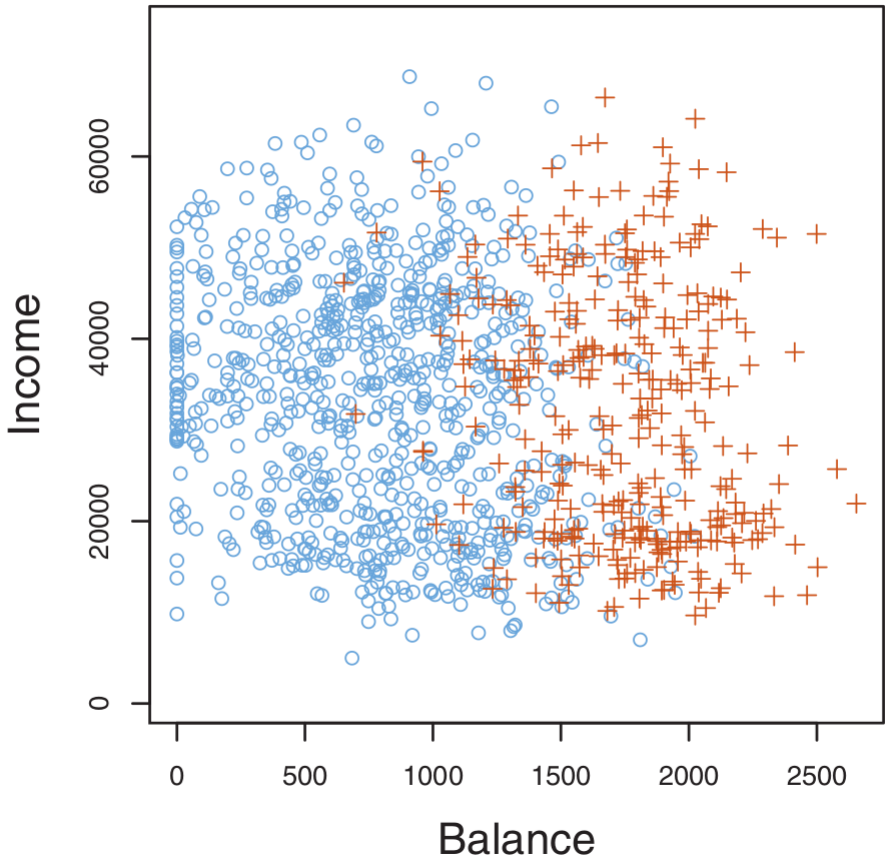
Introduction to Machine Learning

Lecture 6
Linear Models III
Classification

Goker Erdogan
26 – 30 November 2018
Pontificia Universidad Javeriana

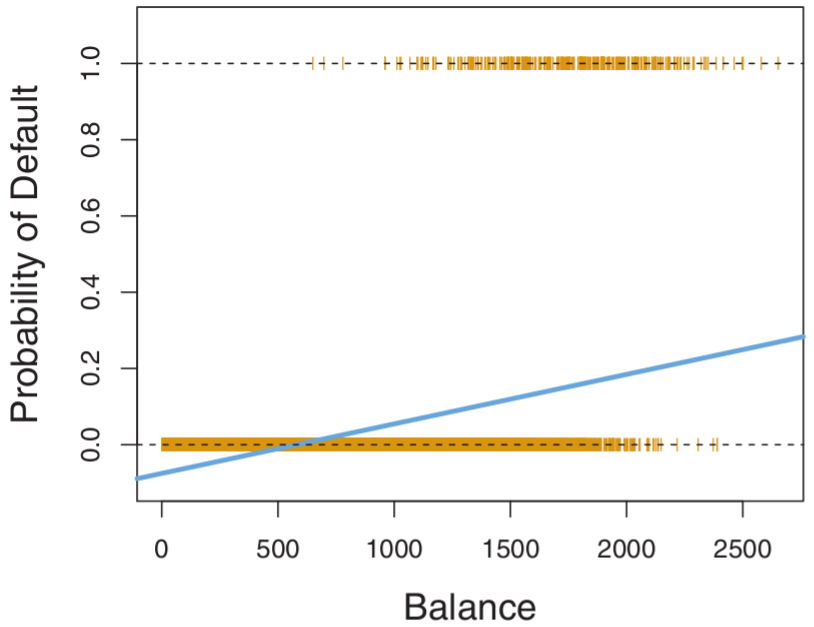
Classification

- Classification:
  - Given N samples (training set) of {x, t}, where t is one of K possible classes
  - Predict t for a new x
- Example: Predict if an individual will default on credit card payments[1]



Why not do linear regression?

- For binary classification (two classes),
  - Use t=0, t=1 to represent classes
  - Linear regression is OK
  - But it will give you numbers outside [0, 1]



Why not do linear regression?

- For binary classification (two classes),
  - May not give you what you expect
  - Points far away from the boundary matter!

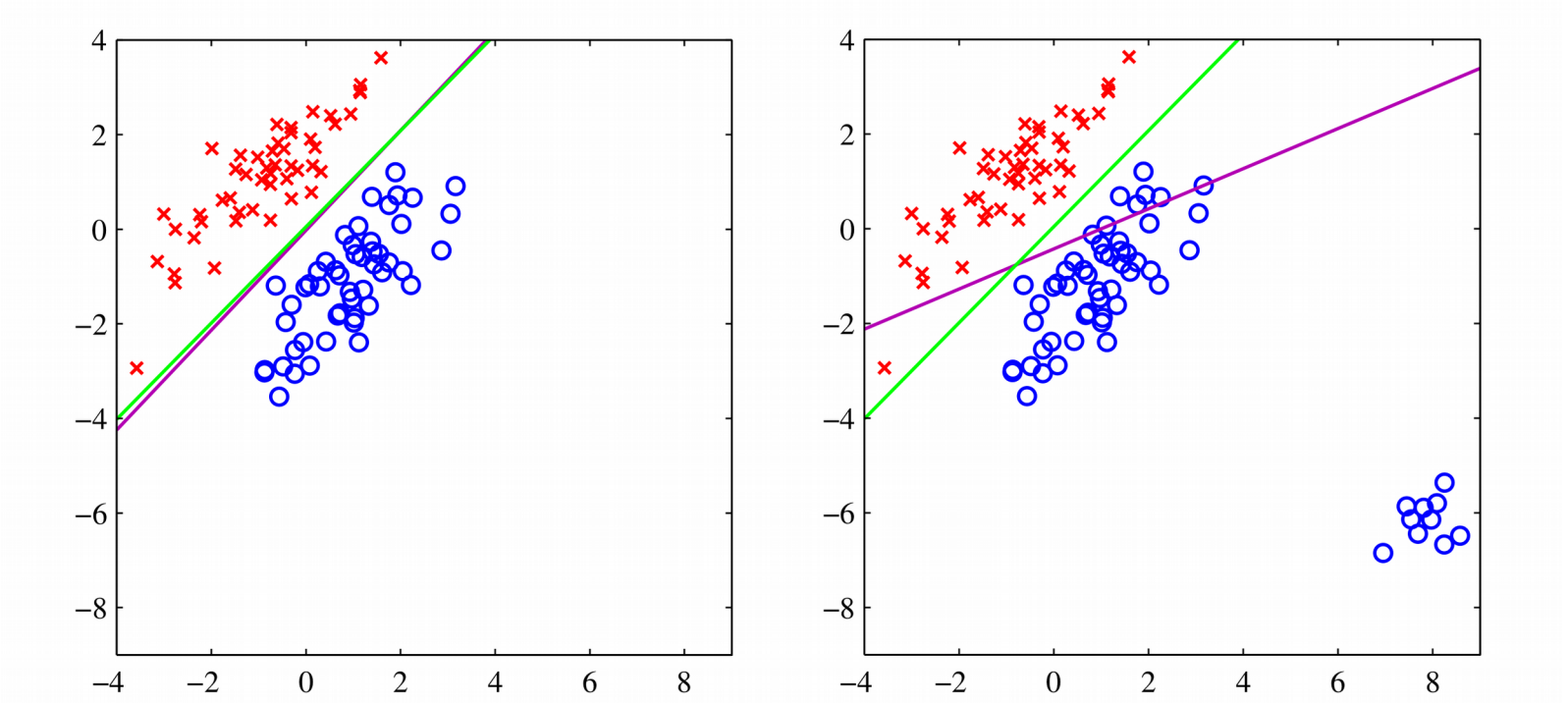


Figure 4.4 The left plot shows data from two classes, denoted by red crosses and blue circles, together with the decision boundary found by least squares (magenta curve) and also by the logistic regression model (green curve), which is discussed later in Section 4.3.2. The right-hand plot shows the corresponding results obtained when extra data points are added at the bottom left of the diagram, showing that least squares is highly sensitive to outliers, unlike logistic regression.

Why not do linear regression?

- For multiclass classification,
  - Imagine three classes: table, chair, cup
  - Which numbers should we assign to each?
    - table=1, chair=2, cup=3
    - Assumes an ordering
    - Assumes the difference (table-chair) = (cup-chair)

- A better representation is one-hot encoding

Table	1	0	0
Chair	0	1	0
Cup	0	0	1

Logistic regression

- Generalize linear regression to classification
- Predict the probability a sample belongs to a class
  - Ensures the outputs are in [0, 1]

- Binary case

$$p(\mathcal{C}_1|\phi) = \sigma(w_0 + \sum_d w_d \phi_d(x_d))$$

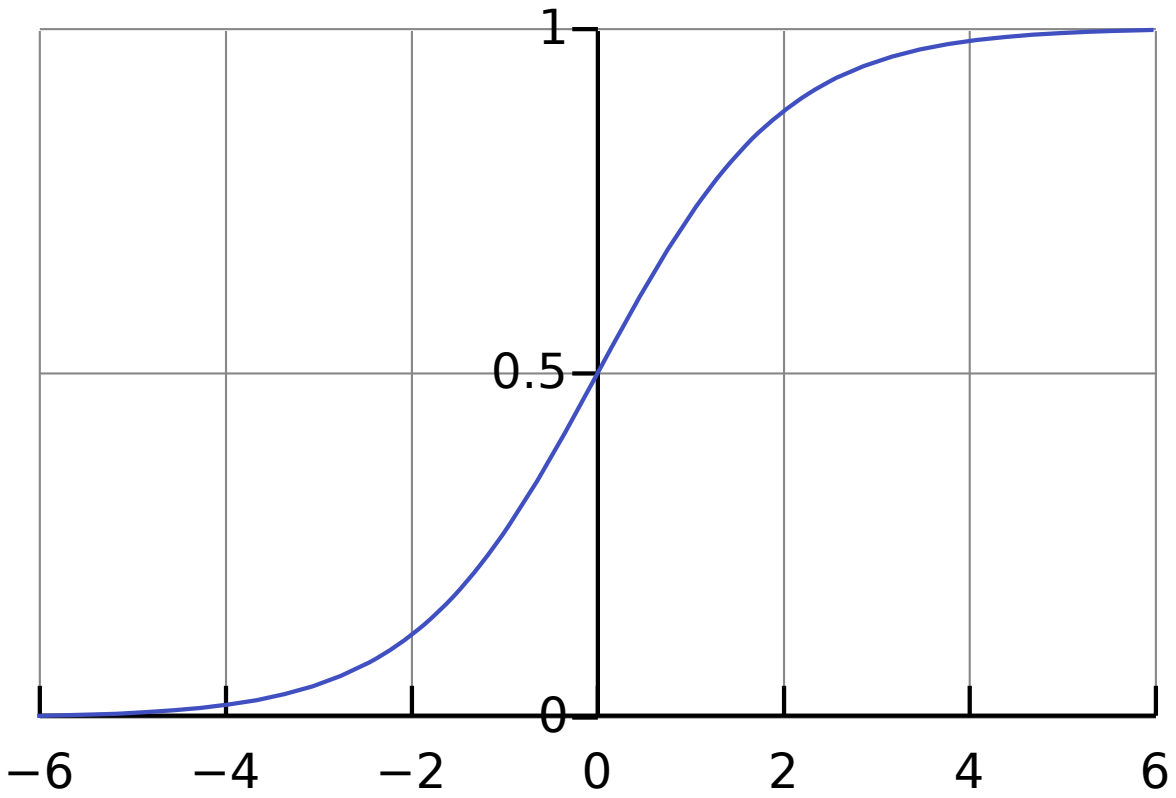
$$= \sigma\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}\right)$$

- σ maps [-∞, +∞] to [0, 1]
- There are many such functions
  - Logistic regression uses the logistic (sigmoid) function

Logistic function

- Maps [-∞, +∞] to [0, 1]

$$\sigma(a) = \frac{1}{1 + e^{-a}}$$

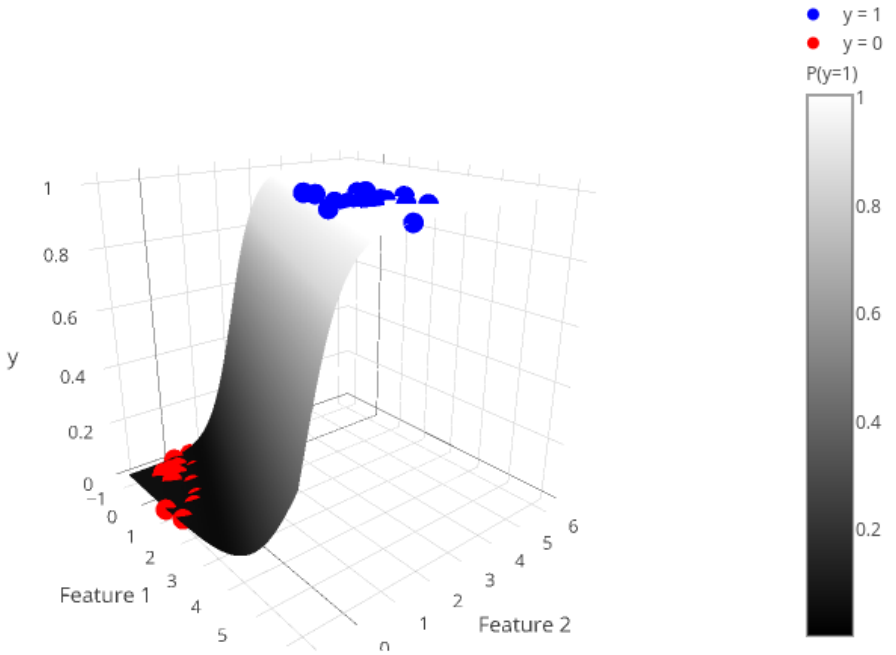


Logistic regression

- Note the boundary between C_1 and C_2 is a line (hyper-plane)
- Predict C_1 if $p(\mathcal{C}_1|\phi) > 0.5$
  - C_2 otherwise
- Boundary:

$$\sigma(w_0 + \sum_d w_d \phi_d(x_d)) = 0.5$$

$$w_0 + \sum_d w_d \phi_d(x_d) = 0$$



[4]

Solving Logistic regression

- Given a training set of N samples {x, t}
  - Find the best w (that maximizes some measure of accuracy)

- Define the following error function

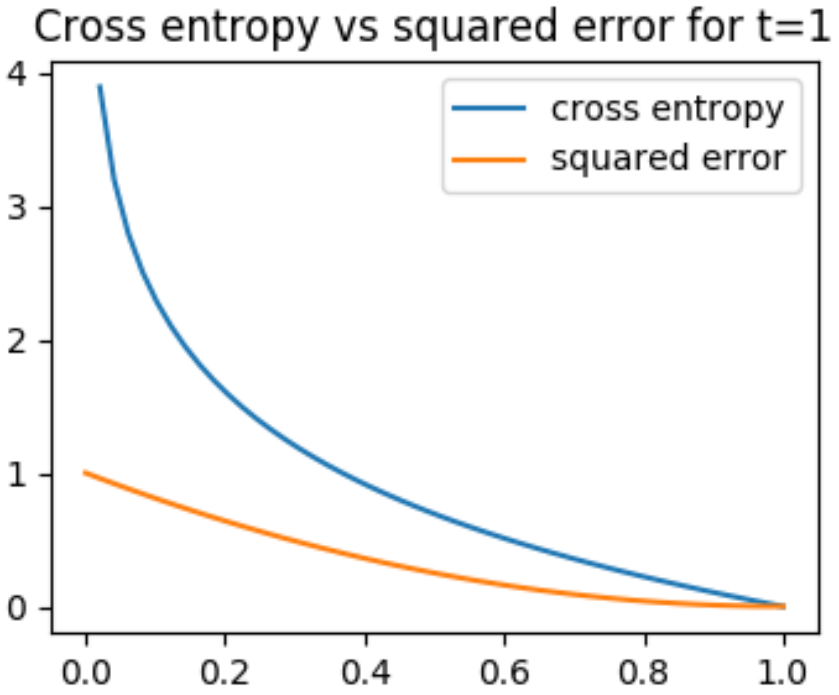
$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}) = -\sum_{n=1}^{N} \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\}$$

- Minimize E(w) to find the optimal w*
  - No closed form solution
  - But still a convex problem
  - Iterative reweighted least squares algorithm (IRLS)
    - Solves a sequence of least-squares problems to find w

Cross-entropy loss function

- Used in classification problems
  - Maximum likelihood for a (Bernoulli) multinomial model [next lecture]
  - Remember t=0 or t=1

$$\mathcal{L}_{CE}(t, y) = -t \log(y) - (1 - t) \log(1 - y)$$



Multiclass Logistic Regression

- Imagine we have K classes
  - Represent targets with one-hot encoding
  - Note now y for each sample is a vector of length K
- Predict the probability a sample belongs to a class
  - Need to predict K numbers y_k such that $y_1 + y_2 + \ldots + y_k = 1$
  - Learn a separate w_k for each class

$$a_k = \mathbf{w}_k^{\mathrm{T}} \boldsymbol{\phi}.$$

  - Softmax function to map them to probabilities

$$p(\mathcal{C}_k|\boldsymbol{\phi}) = y_k(\boldsymbol{\phi}) = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

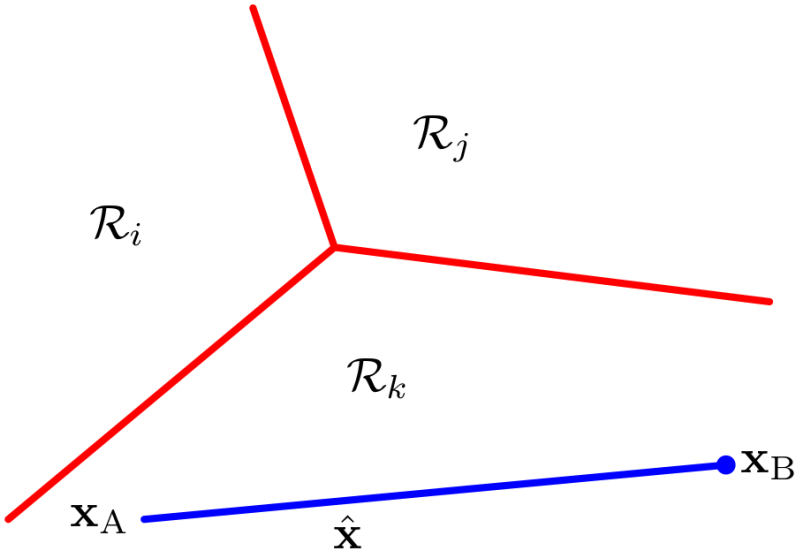
Multiclass Logistic Regression

- Error function for multiclass case

$$E(\mathbf{w}_1, \ldots, \mathbf{w}_K) = -\ln p(\mathbf{T}|\mathbf{w}_1, \ldots, \mathbf{w}_K) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_{nk}$$

- Minimize E(w) to find optimal w_1, w_2, …, w_k
- Again, boundary between class k and class m is a line (hyper-plane)

Figure 4.3 Illustration of the decision regions for a multiclass linear discriminant, with the decision boundaries shown in red. If two points $\mathbf{x}_\mathrm{A}$ and $\mathbf{x}_\mathrm{B}$ both lie inside the same decision region $\mathcal{R}_k$, then any point $\widehat{\mathbf{x}}$ that lies on the line connecting these two points must also lie in $\mathcal{R}_k$, and hence the decision region must be singly connected and convex.



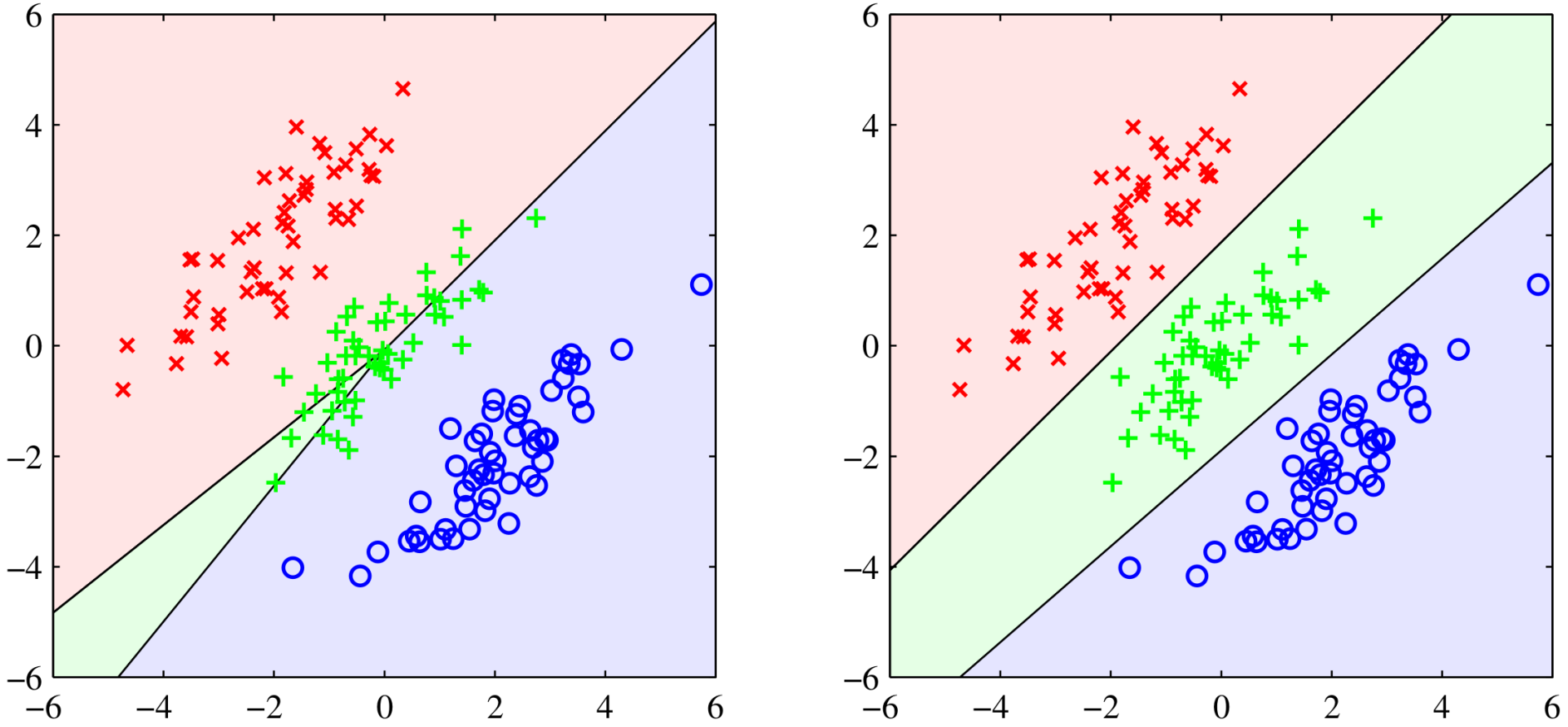


Figure 4.5 Example of a synthetic data set comprising three classes, with training data points denoted in red $(\times)$, green $(+)$, and blue $(\circ)$. Lines denote the decision boundaries, and the background colours denote the respective classes of the decision regions. On the left is the result of using a least-squares discriminant. We see that the region of input space assigned to the green class is too small and so most of the points from this class are misclassified. On the right is the result of using logistic regressions as described in Section 4.3.2 showing correct classification of the training data.

Summary

- Classification
  - Why not linear regression?
- Logistic Regression
  - Logistic function
  - Cross-entropy loss function
- Multiclass case
  - Softmax function

- Exercises
  - Do the lab in Section 4.6 (up to 4.6.3) of ISLR

References

[1] James, Witten, Hastie, and Tibshirani. An Introduction to Statistical Learning with Applications in R. Chapter 4.

[2] Hastie, Tibshirani, and Friedman. The Elements of Statistical Learning. Chapter 4.

[3] Bishop. Pattern Recognition and Machine Learning. Chapter 4.

[4] https://florianhartl.com/logistic-regression-geometric-intuition.html