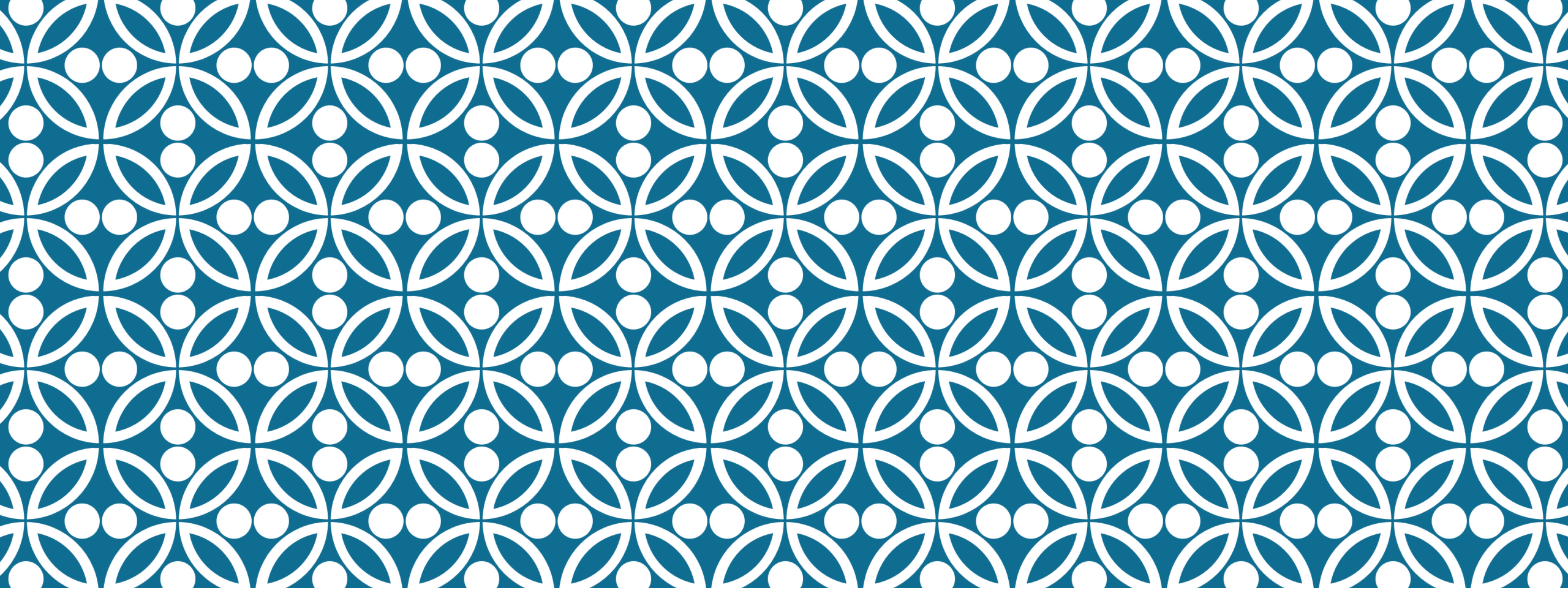


應用機器學習

Brian Chan 陳醒凡

課程目標

1. 了解基本的數據分析
2. 了解基本的機器學習(Machine Learning)方法
3. 掌握Python的基本操作和一些有用的package
4. 處理及從網上下載數據
5. 在Python上應用機器學習

今天課堂概要

Logistic regression & Support vector machine

1. Logistic regression
2. Support vector machine (SVM)
3. Case Study: Breast Cancer Detection

CLASSIFICATION

Classification, like regression, is a predictive task, but one in which the outcome takes only values across discrete categories; classification problems are very common (arguably just as or perhaps even more common than regression problems!

Observed Data be $D_i = (x_i, y_i), i = 1, \ldots, N.$

Examples:

–Predicting whether a patient will develop breast cancer or remain healthy, given genetic information

–Predicting whether or not a user will like a new product, based on user covariates and a history of his/her previous ratings

–Predicting the region of Italy in which a brand of olive oil was made, based on its chemical composition

–Predicting the next elected president, based on various social, political, and historical measurements

LOGISTIC REGRESSION

Throughout this section we will assume that the outcome has two classes, for simplicity. (We return to the general K class setup at the end.)

Logistic regression starts with different model set up than linear regression: instead of modeling Y as a function of X directly, we model the probability that Y is equal to class 1, given X. First, abbreviate

$$p(X) = P(Y = 1|X).$$

Then the logistic model is

$$p(X) = \frac{e^{\beta^T X}}{1 + e^{\beta^T X}}$$

LOGISTIC REGRESSION

Given samples $(x_i, y_i) \in \mathbb{R}^p \times \{0,1\},\ i = 1, \ldots n$, we let $p(x_i) = P(y_i = 1|x_i)$ and assume

$$\log(p(x_i)/[1 - p(x_i)]) = \beta^T x_i, \qquad i = 1, \ldots n.$$

To construct an estimate $\hat{\beta}$ of the coefficients, we will use the principle of maximum likelihood. i.e., assuming independence of the samples, the likelihood (conditional on $\mathrm{x}_i = 1, \ldots n$) is

$$L(\beta) = \prod_{i=1}^{n} p(x_i)^{y_i}\big(1 - p(x_i)\big)^{1-y_i}$$

Note that maximizing a function is the same as maximizing the log of a function (because log is monotone increasing). Therefore $\hat{\beta}$ is equivalently chosen to maximize the log likelihood

$$l(\beta) = \sum_{i=1}^{n} y_i \log p(x_i) + (1 - y_i) \log(1 - p(x_i))$$

LOGISTIC REGRESSION

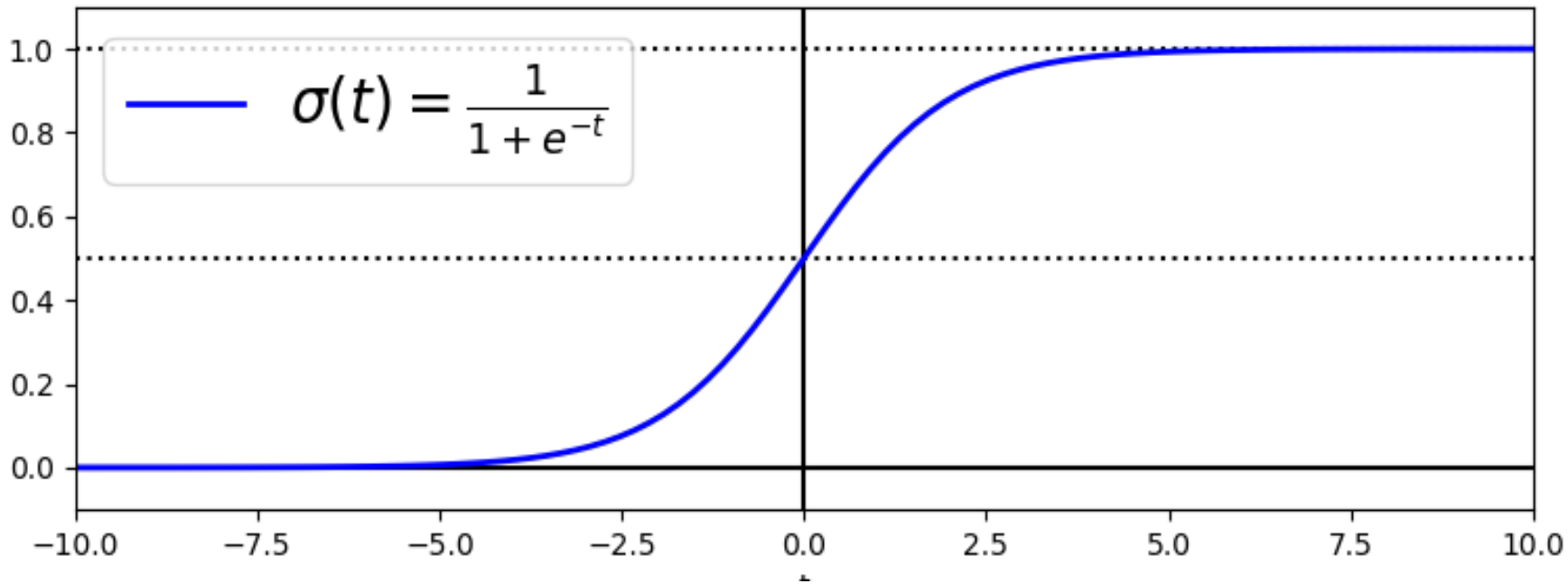
Suppose that we have formed the estimate $\hat{\beta}$ of the logistic coefficients, as discussed in the last section. To predict the outcome of a new input $x \in \mathbb{R}^p$, we for

$$\hat{p}(X) = \frac{\exp(\hat{\beta}^T X)}{1 + \exp(\hat{\beta}^T X)}$$

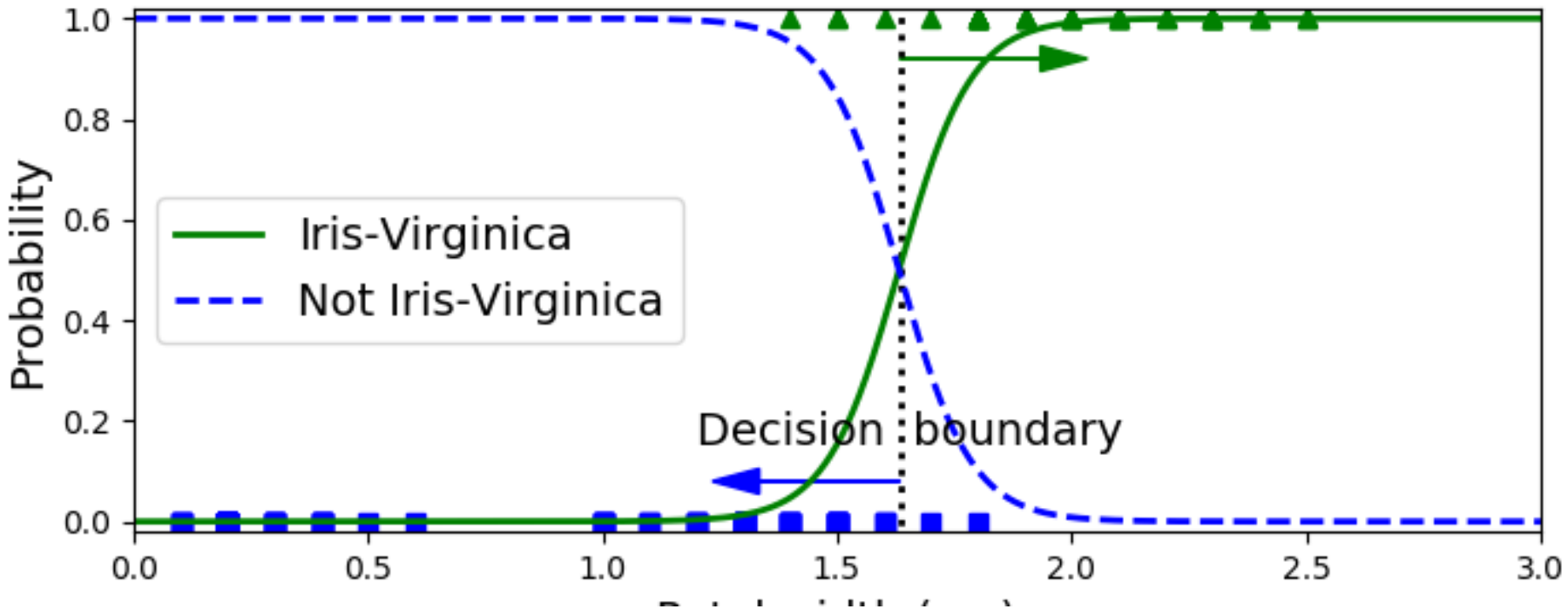
and then predict the associated class according

$$\hat{y}(x) = \begin{cases} 0, & \hat{p}(X) \leq 0.5 \\ 1, & \hat{p}(X) > 0.5 \end{cases}$$

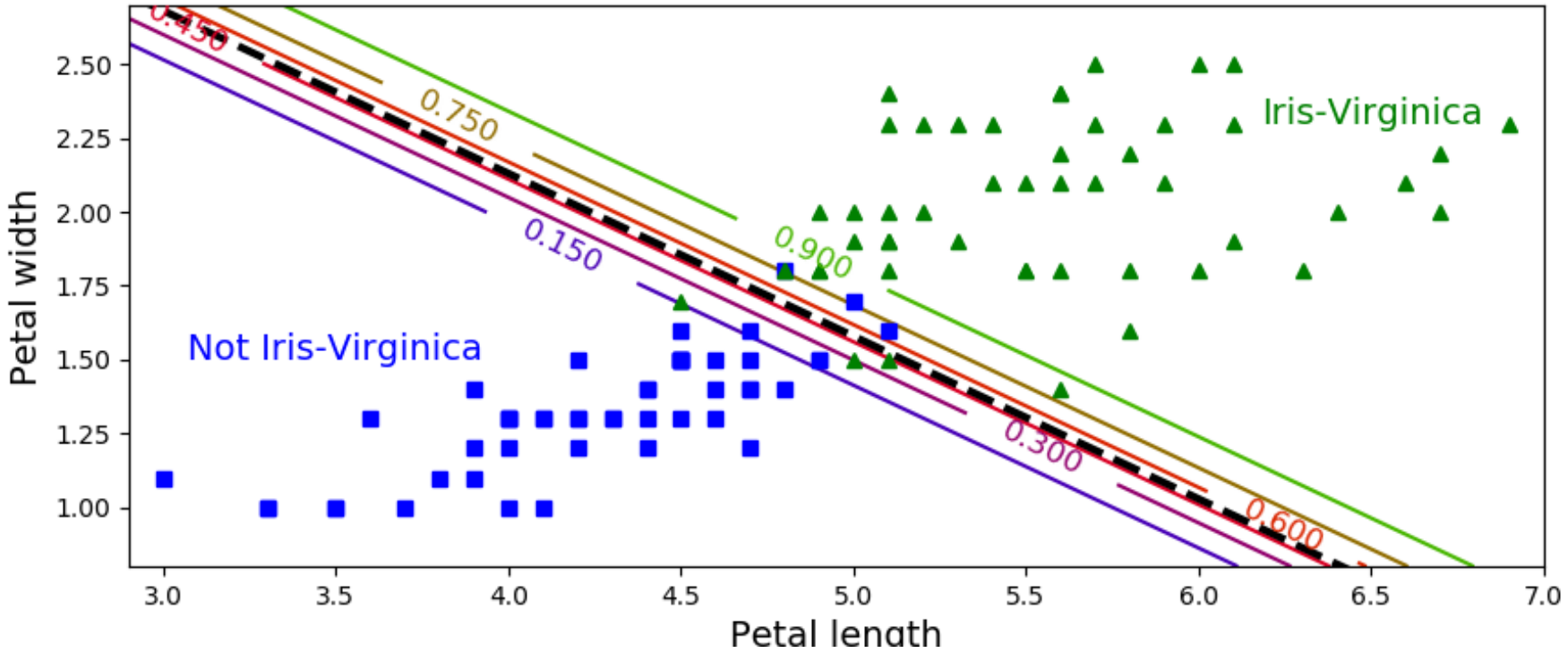
LOGISTIC REGRESSION



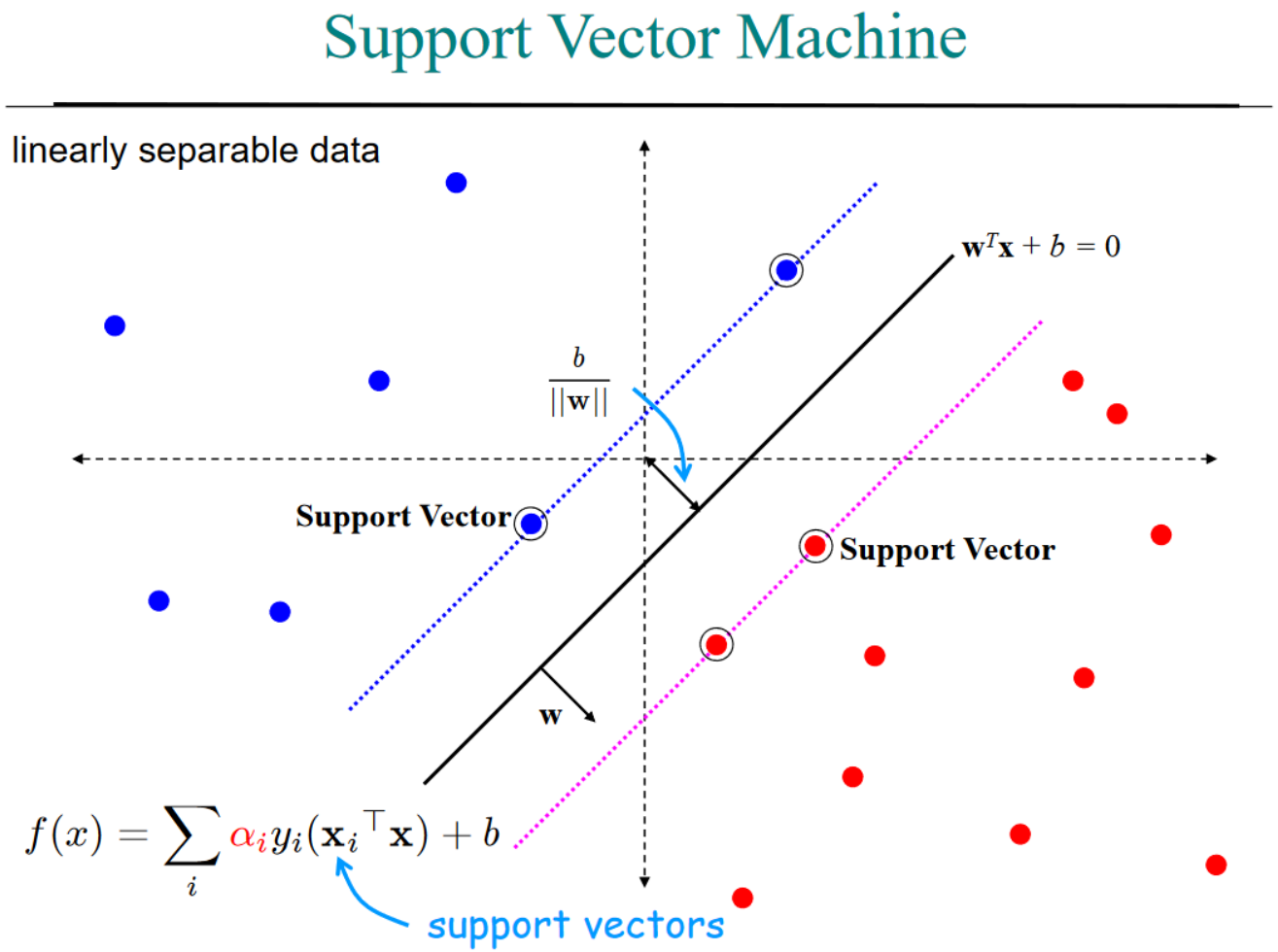
LOGISTIC REGRESSION (2D)



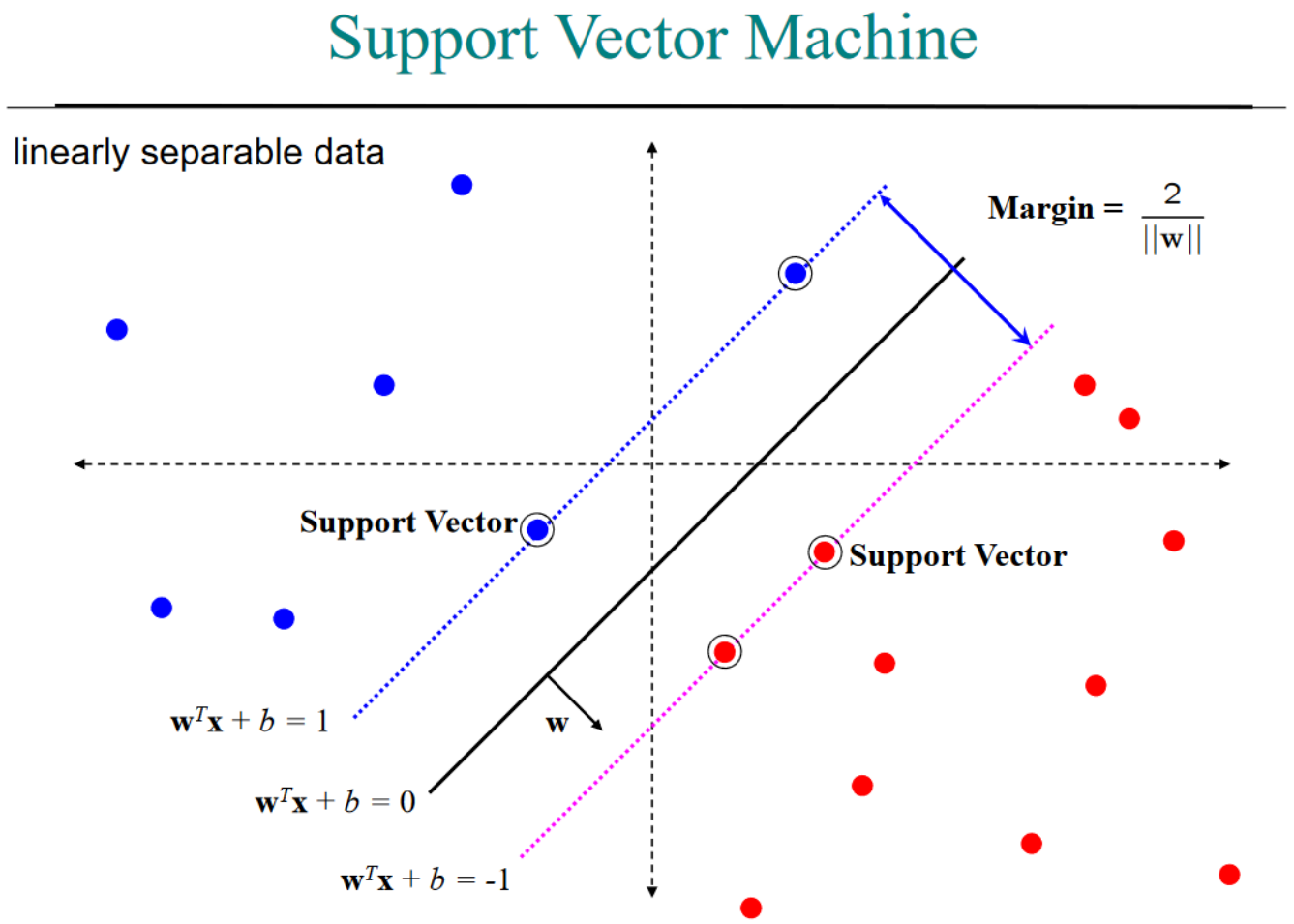
LOGISTIC REGRESSION (3D)



SUPPORT VECTOR MACHINE



SUPPORT VECTOR MACHINE



SUPPORT VECTOR MACHINE

Learning the SVM can be formulated as an optimization:

$$\max_{w,b} \frac{2}{||w||}, \quad \text{s.t.} \begin{cases} w^T x_i + b \geq 1, & if\ y_i = +1 \\ w^T x_i + b < 1, & if\ y_i = -1 \end{cases}, \ i = 1, \ldots, N.$$

Or equivalently

$$\min_{w,b} ||w||^2 , \quad \text{s.t. } y_i(w^T x_i + b) \geq 1, \ i = 1, \ldots, N.$$

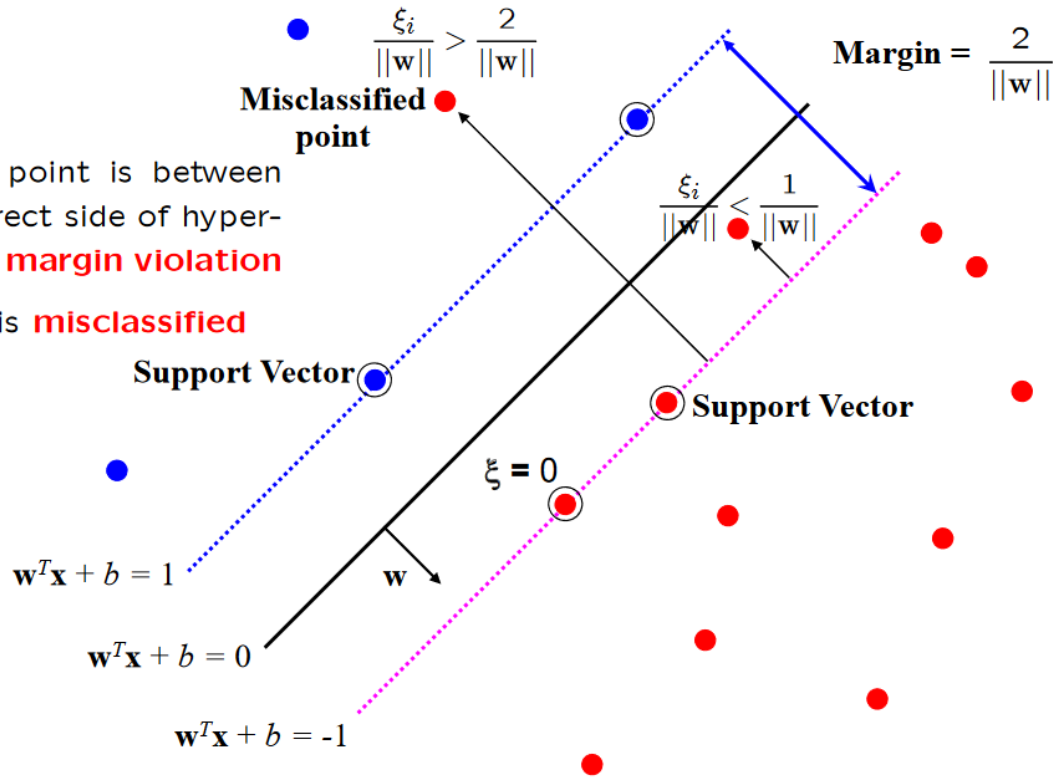
This is a quadratic optimization problem subject to linear constraints and there is a unique minimum.

SUPPORT VECTOR MACHINE

Introduce "slack" variables

$\xi_i \geq 0$

- for $0 < \xi \leq 1$ point is between margin and correct side of hyper-plane. This is a **margin violation**
- for $\xi > 1$ point is **misclassified**



SUPPORT VECTOR MACHINE

The optimization problem becomes

$$\min_{w,b,\xi_i} ||w||^2 + C\sum_{i=1}^{N} \xi_i$$

s.t. $y_i(w^T x_i + b) \geq 1 - \xi_i, \ \ i = 1, \ldots, N.$

Every constraint can be satisfied if ξ_i is sufficiently large

C is are regularization parameter

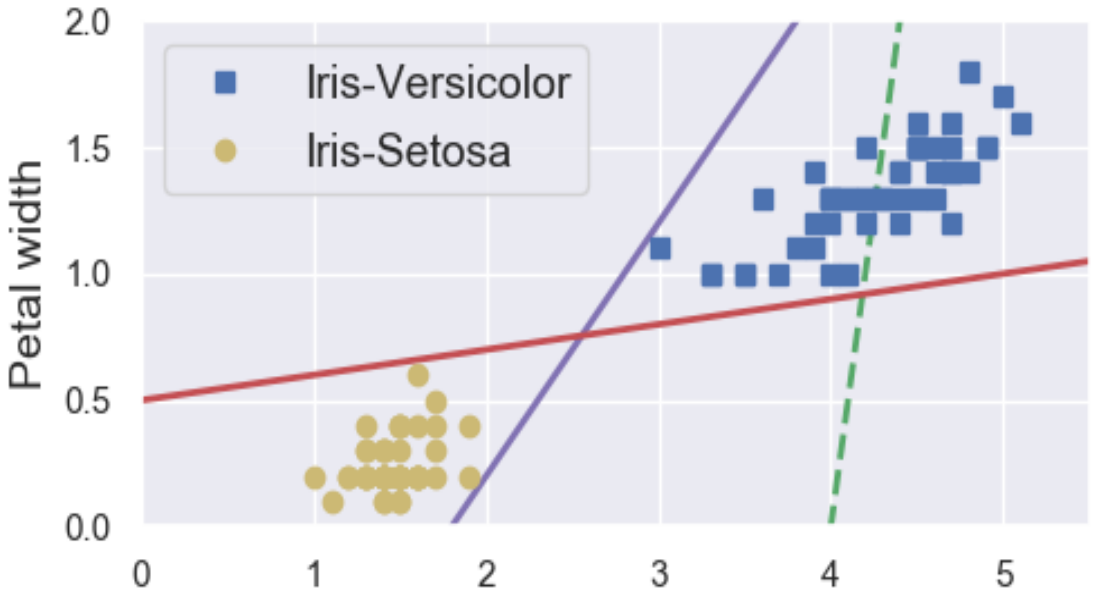
- small C allows constraints to be easily ignored → large margin

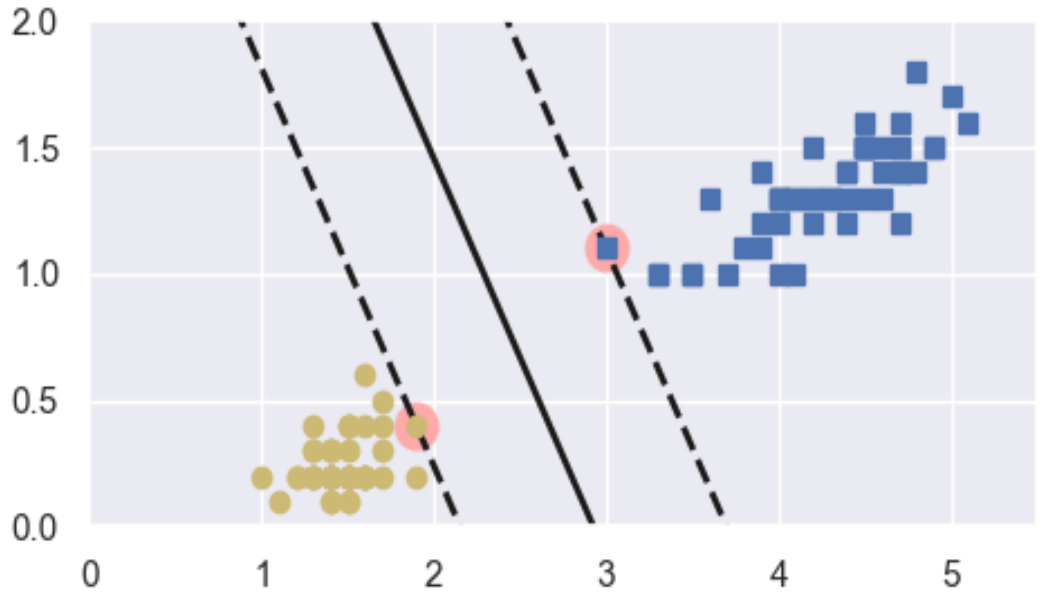
- large C makes constraints hard to ignore → narrow margin

- C = ∞ enforces all constraints: hard margin

This is still a quadratic optimization problem and there is an unique minimum.
Note, there is only one parameter, C.

SUPPORT VECTOR MACHINE





CASE STUDY: BREAST CANCER DETECTION

Background:

Breast cancer is the most common cancer amongst women in the world. It accounts for 25% of all cancer cases, and affected over 2.1 Million people in 2015 alone. It starts when cells in the breast begin to grow out of control. These cells usually form tumors that can be seen via X-ray or felt as lumps in the breast area.

Early diagnosis significantly increases the chances of survival. The key challenges against it's detection is how to classify tumors into malignant (cancerous) or benign(non cancerous). A tumor is considered malignant if the cells can grow into surrounding tissues or spread to distant areas of the body. A benign tumor does not invade nearby tissue nor spread to other parts of the body the way cancerous tumors can. But benign tumors can be serious if they press on vital structures such as blood vessels or nerves.

CASE STUDY: BREAST CANCER DETECTION

Background:

Machine Learning technique can dramatically improve the level of diagnosis in breast cancer. Research shows that experienced physicians can detect cancer by 79% accuracy, while a 91 %(sometimes up to 97%) accuracy can be achieved using Machine Learning techniques.

Project Task

In this study, my task is to classify tumors into malignant (cancerous) or benign (non-cancerous) using features obtained from several cell images.

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

DISCUSSIONS

Imbalance data

Standardization

Multi-class classification $\hat{y}_k = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}} f_k(x)$

One-vs-Rest (OvR)One-vs-One (OvO)

8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset:
https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/

REFERENCE

Reference for support vector machine:

http://cs229.stanford.edu/notes/cs229-notes3.pdf

http://www.robots.ox.ac.uk/~az/lectures/ml/lect2.pdf

Reference for logistic regression:

https://www.stat.cmu.edu/~ryantibs/advmethods/notes/logreg.pdf

Case Study: Breast Cancer

https://arxiv.org/pdf/1711.07831.pdf

今天課堂概要

Logistic regression & Support vector machine

1. Logistic regression
2. Support vector machine
3. Case Study: Breast Cancer Detection

下一課...

分群法