人工智能之自動化光學檢測實務

Yi-Yung Chen

適用影片單元04~單元09

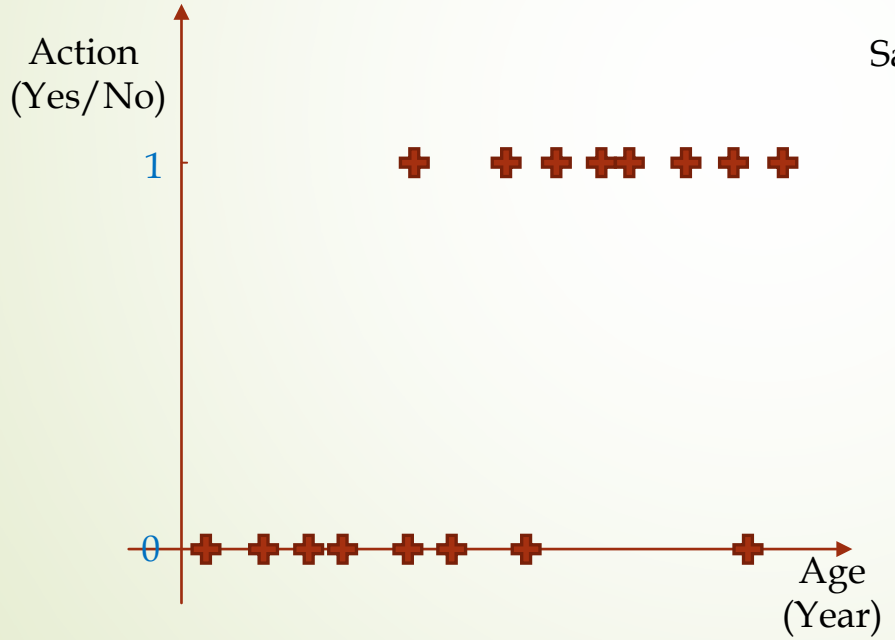
分類(Classification)

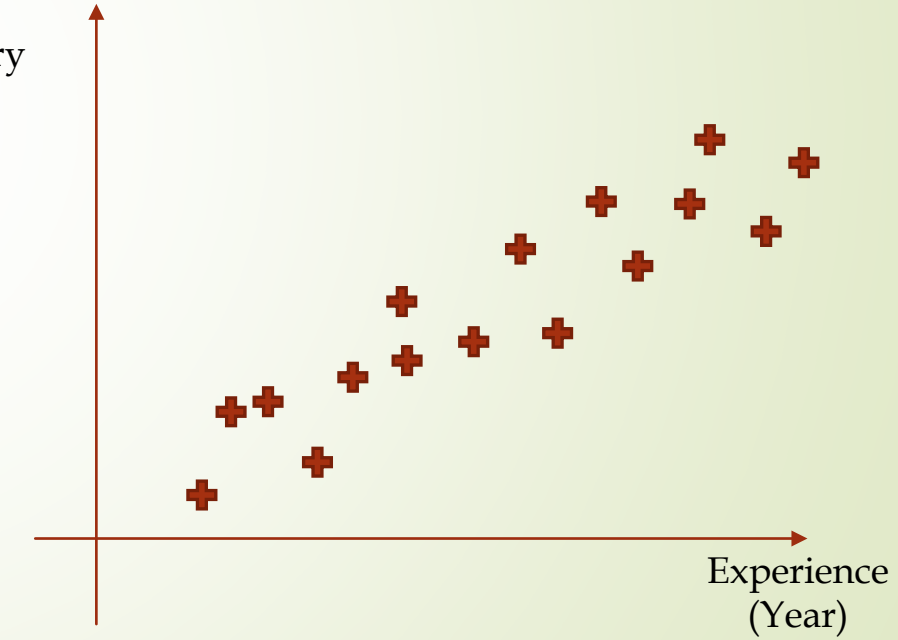
- 邏輯回歸
- 支持向量機(SVM)
- 核函數支持向量機
- 樸素貝葉斯
- 決策樹
- 隨機森林
- 分類模型性能評價與選擇

邏輯回歸

Classification

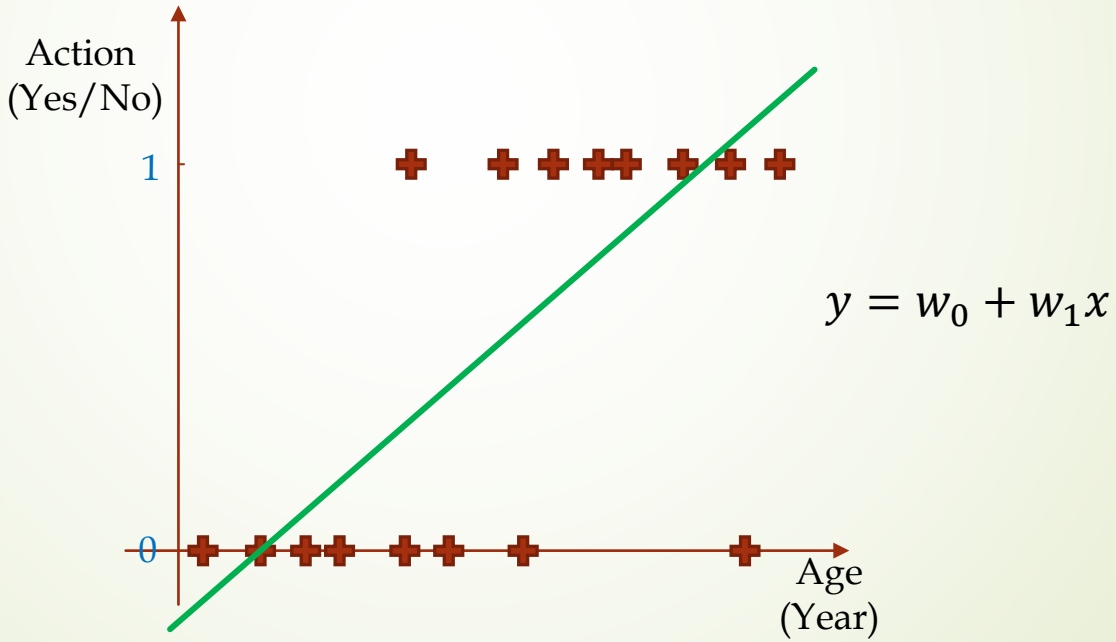
- From what we know





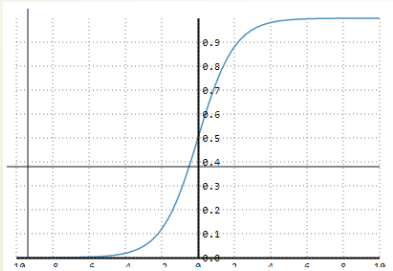
From Regression

- If we use regression to be classification

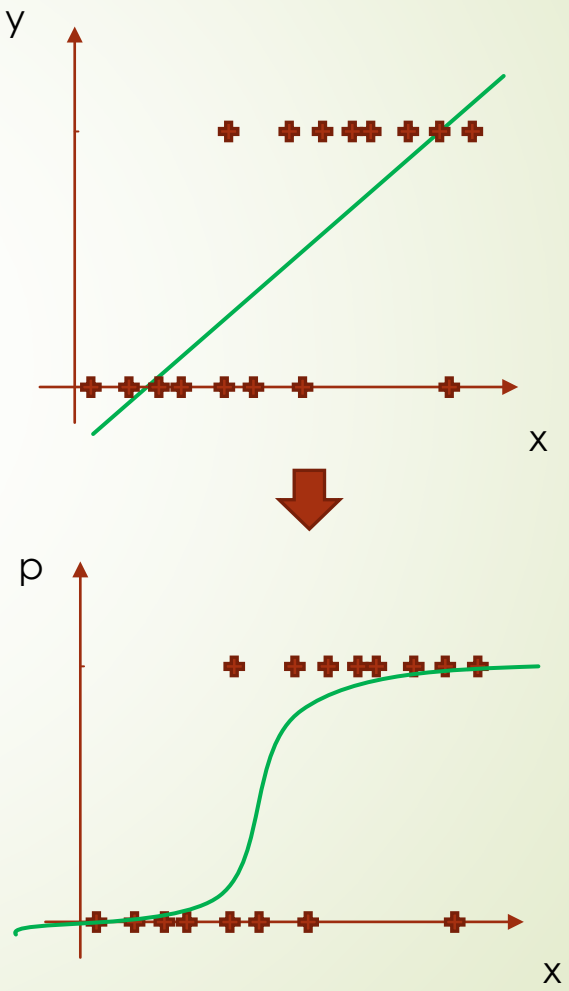


Sigmoid Function

- Regression
  - $y = w_0 + w_1 x$
- Use Sigmoid Function
  - $P = \frac{1}{1+e^{-y}}$
  - $ln\left(\frac{p}{1-p}\right) = w_0 + w_1 x$

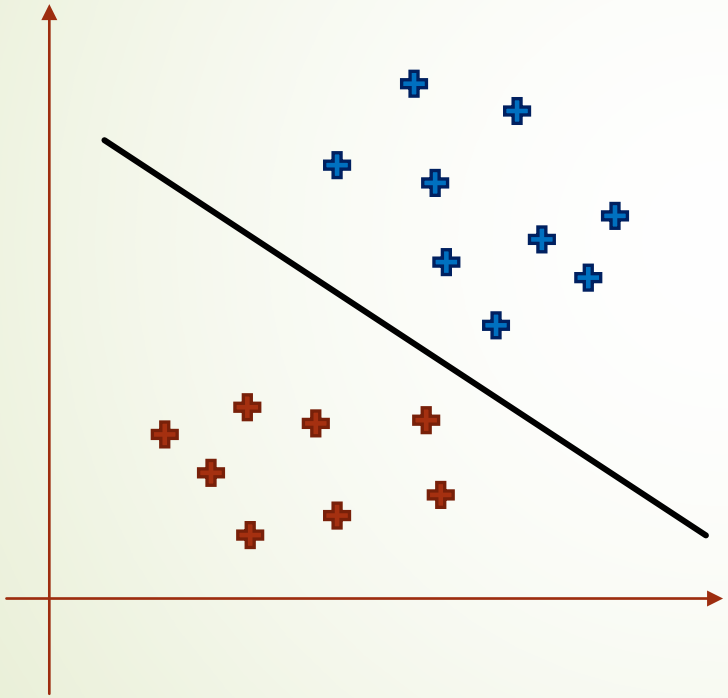


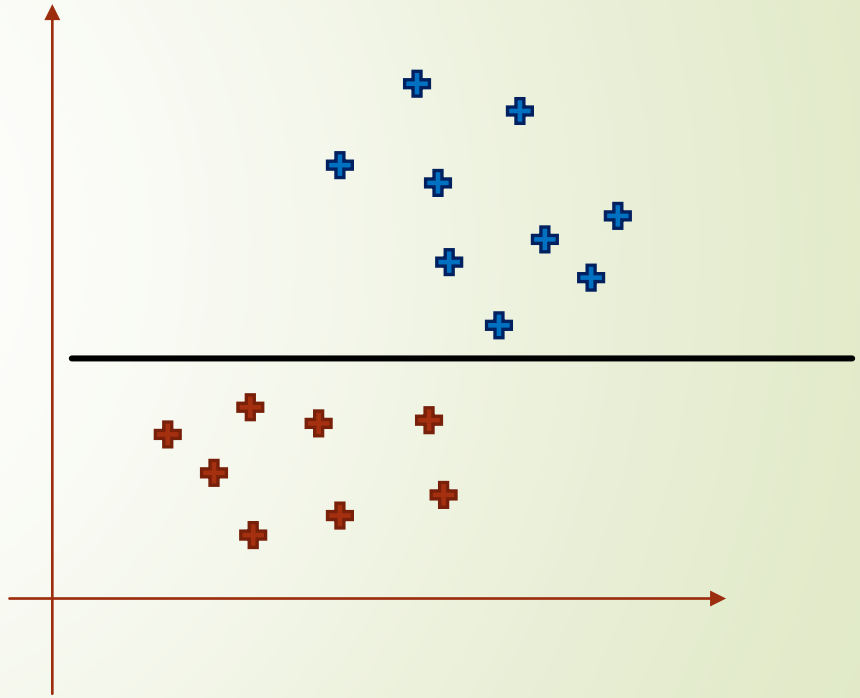
S function



支持向量機(SVM)

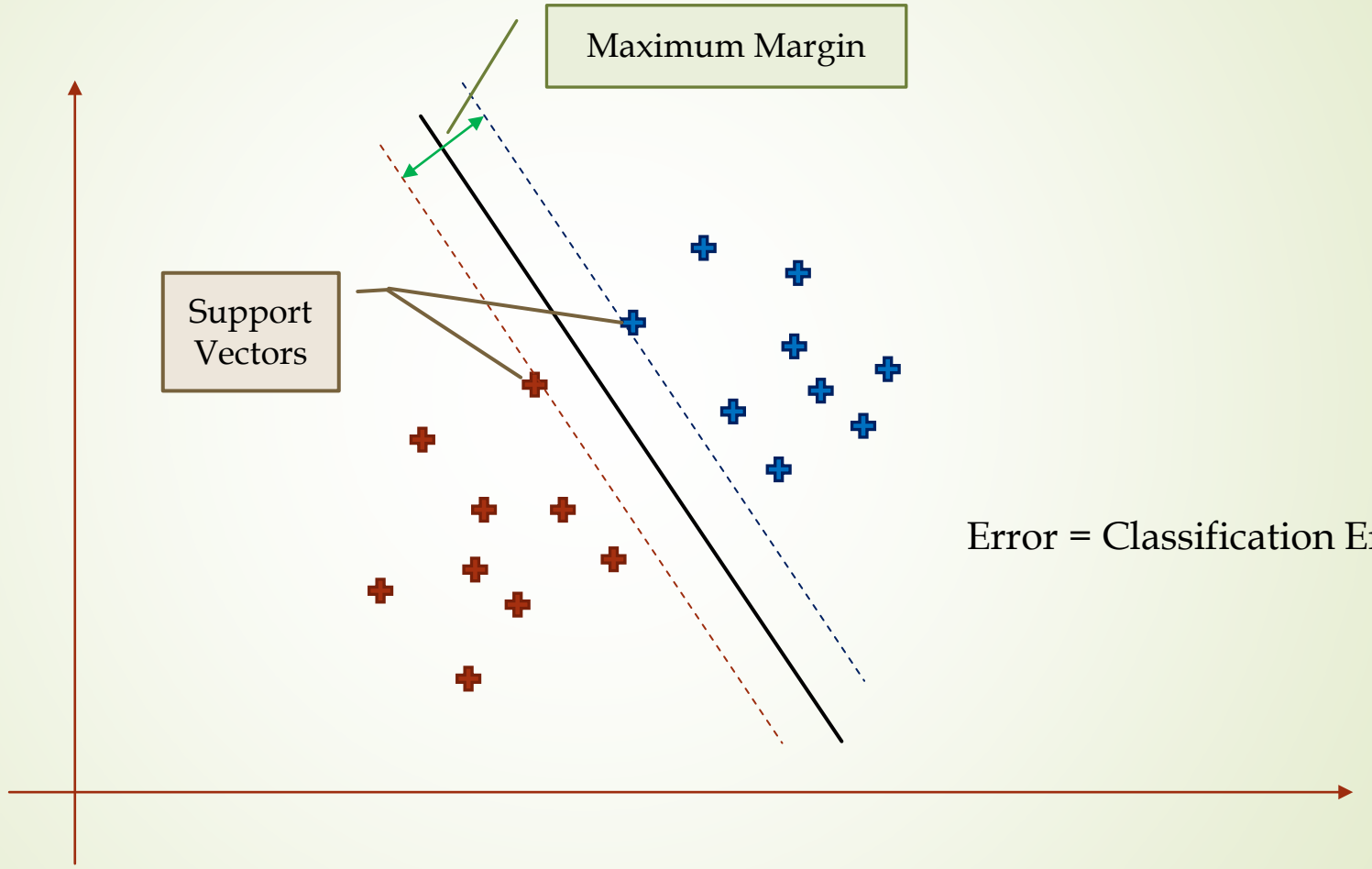
Which one is better





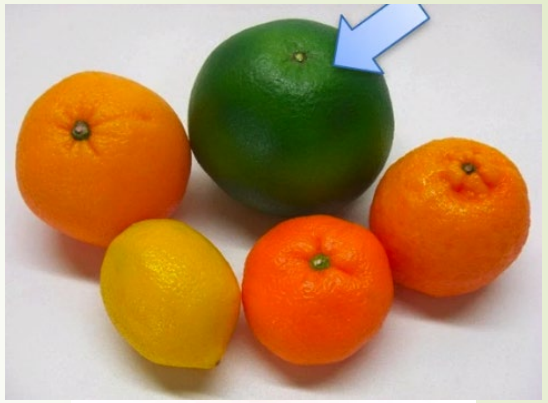
All the results are correct, which one is better?

支持向量機



Error = Classification Error + Margin Error

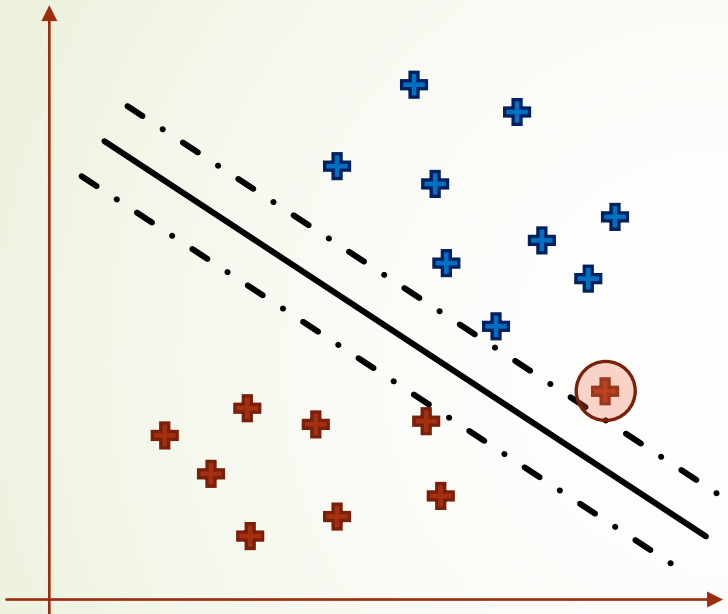
What's so Special about SVMs?

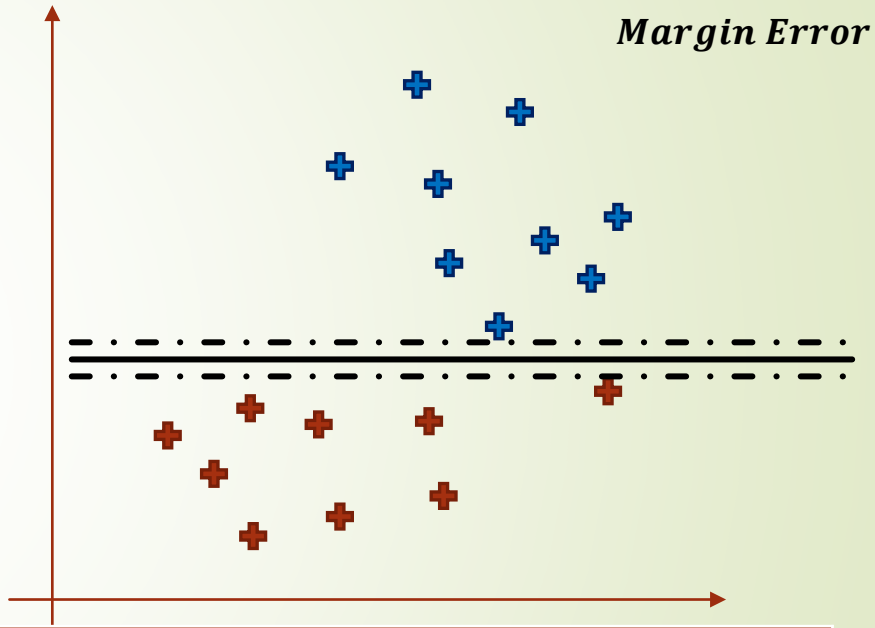


The C Parameter

Error = C × Classification Error + Margin Error

$$Margin\ Error = |W|^2$$



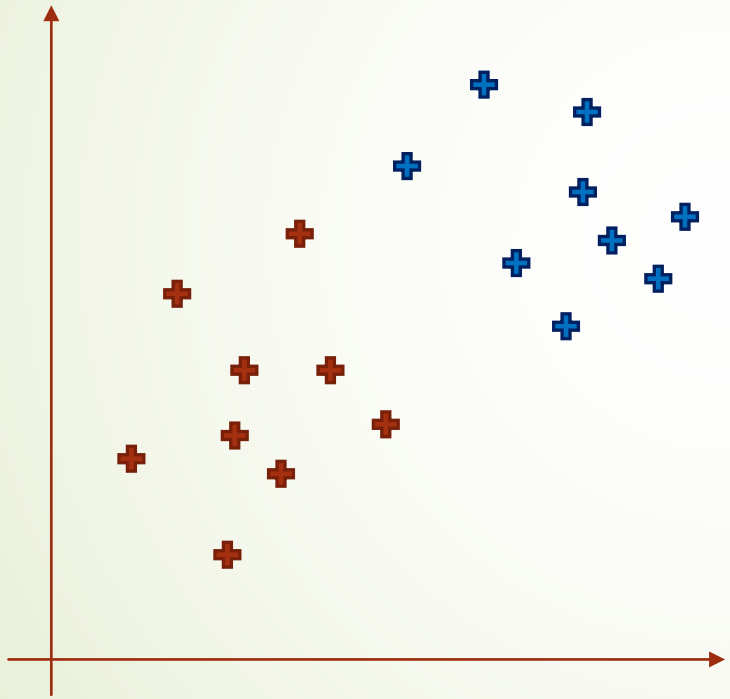


- $w \downarrow$
- $|wX| \downarrow$
- $p \downarrow$
- $Classification\ Error \uparrow$

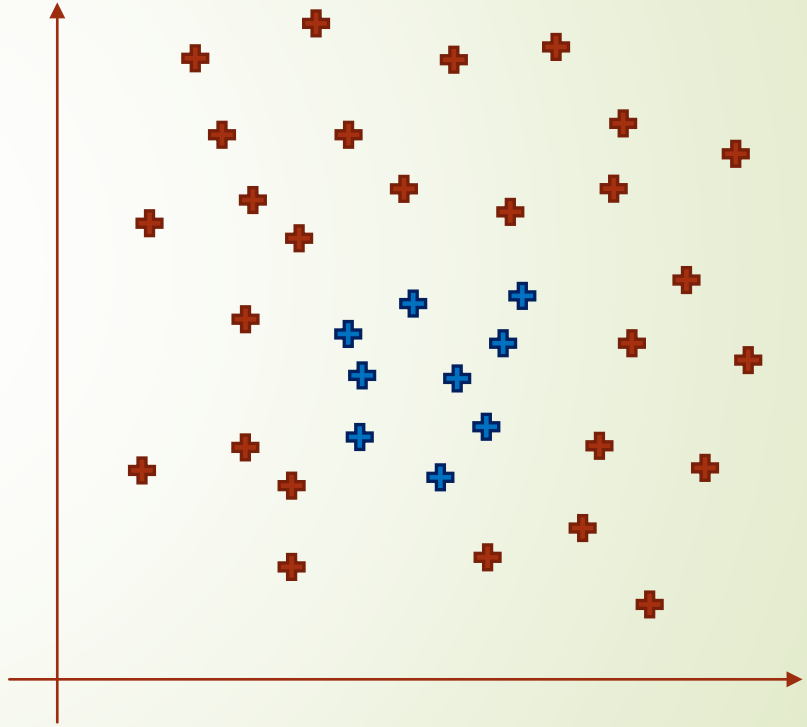
Small C	Large C
Large margin	Classifies points well
May make classification errors	May have a small margin
Reduce Overfitting	

3.3 核函數支持向量機

線性可分 vs. 線性不可分

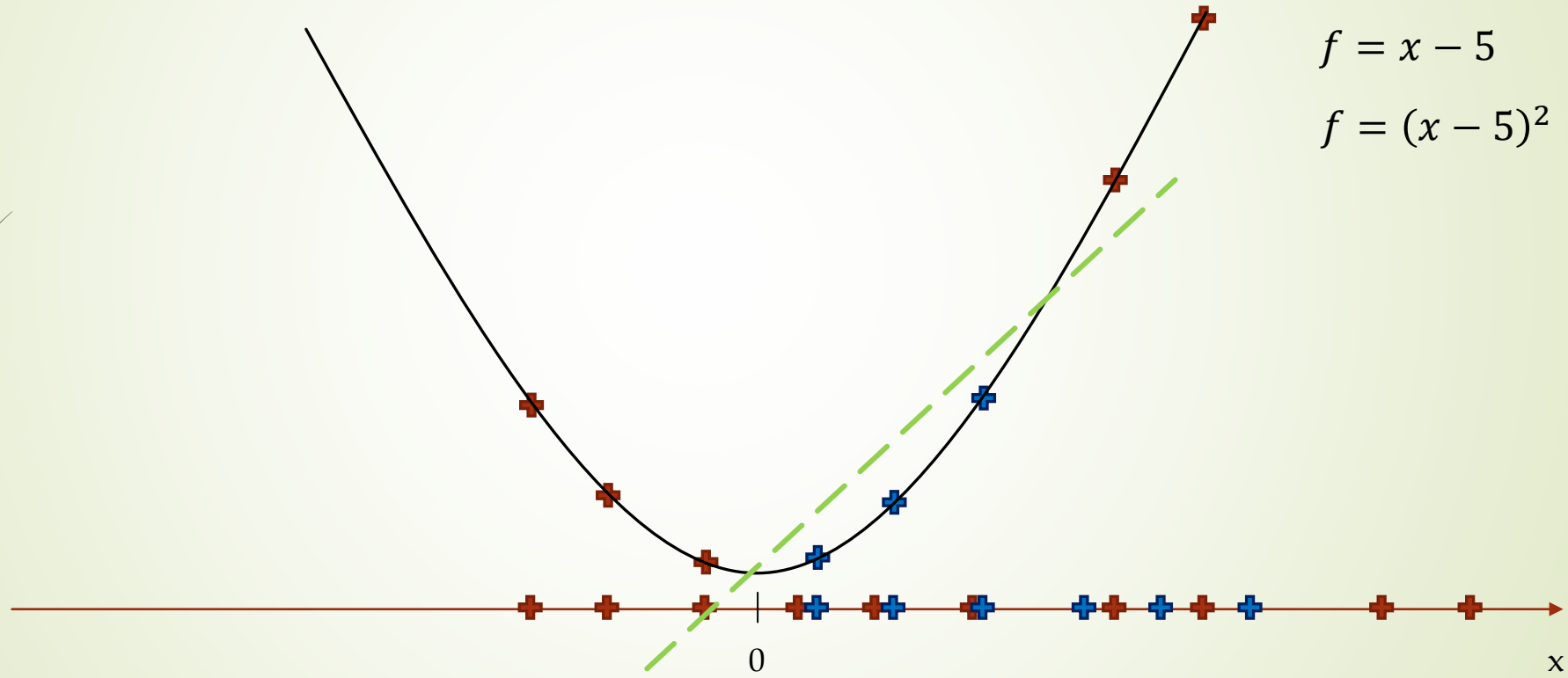


Linearly Separable



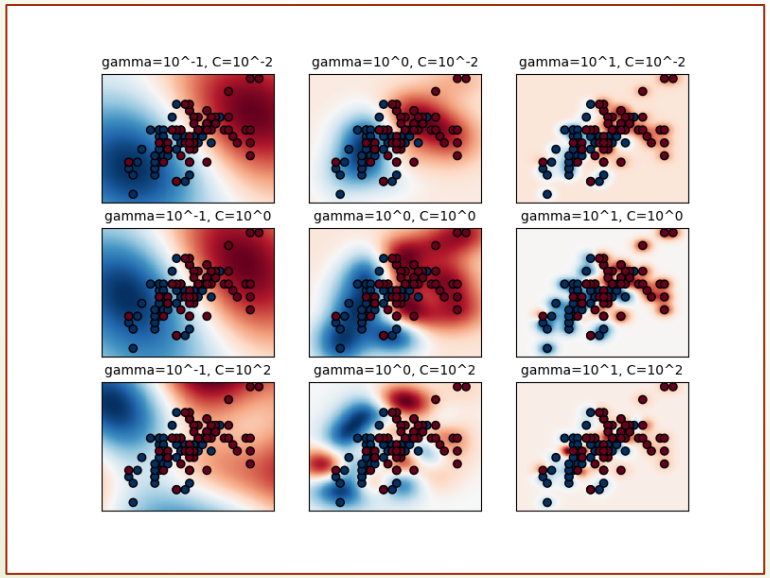
Not Linearly Separable

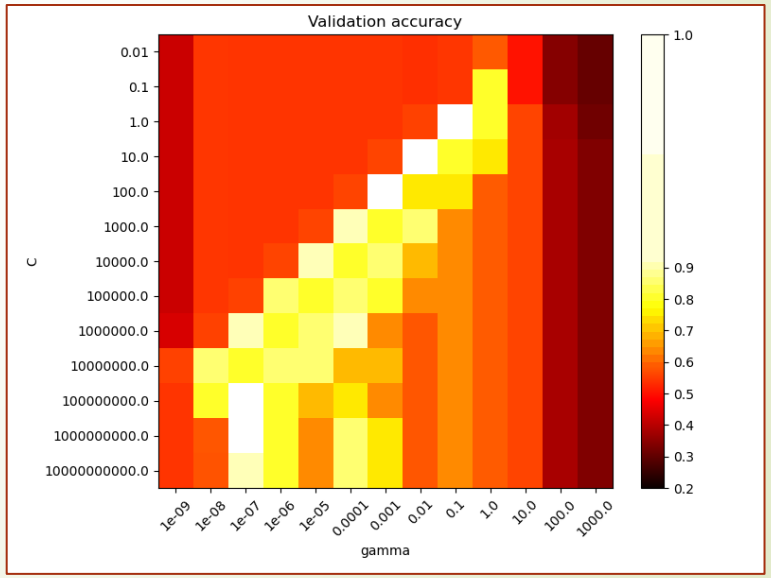
高維投射 (一維)



RBF SVM parameters

- gamma可以看作是模型選擇作為support vectors的樣本的影響半徑的倒數
  - $\gamma \propto \sigma^{-1}$
- 對於較大的 C，能夠更好地正確分類所有訓練點，但接受較小的邊距。





Hyperparameters

- C
  - The C parameter.
- Kernel
  - The kernel. The most common ones are 'linear', 'poly', and 'rbf'.
- Degree
  - f the kernel is polynomial, this is the maximum degree of the monomials in the kernel.
- Gamma
  - If the kernel is rbf, this is the gamma parameter.

樸素貝葉斯

Concepts

- Probabilities (機率)
  - $P(woman) = {}^{50}/_{100} = 0.5, P(man) = {}^{50}/_{100} = 0.5$
- Conditional Probabilities (條件機率)
  - $P(long\ hair \mid woman) = {}^{25}/_{50} = 0.5, P(long\ hair \mid man) = {}^{2}/_{50} = 0.04$
  - $P(A \mid B) \neq P(B \mid A)$
- Joint Probabilities (聯合機率)
  - $P(woman\ with\ long\ hair) = P(woman) \times P(long\ hair \mid woman) = 0.5 \times 0.5 = 0.25$
  - $P(A\ and\ B) = P(\mathrm{B}\ and\ A)$
- Marginal Probabilities (邊際機率)
  - $P(long\ hair) = P(woman\ with\ long\ hair) + P(man\ with\ long\ hair) = 0.25 + 0.02 = 0.27$

Bayesian Inference

- Joint Probabilities
  - $P(man\ with\ long\ hair) = P(man) \times P(long\ hair \mid man)$
  - $P(long\ hair\ and\ man) = P(long\ hair) \times P(man \mid long\ hair)$
  - $P(man\ with\ long\ hair) = P(long\ hair\ and\ man)$
- Bayesian Inference
  - $P(man \mid long\ hair) = \frac{P(man) \times P(long\ hair \mid man)}{P(long\ hair)} = \frac{P(man\ with\ long\ hair)}{P(woman\ with\ long\ hair) + P(man\ with\ long\ hair)}$

Bayes Theorem

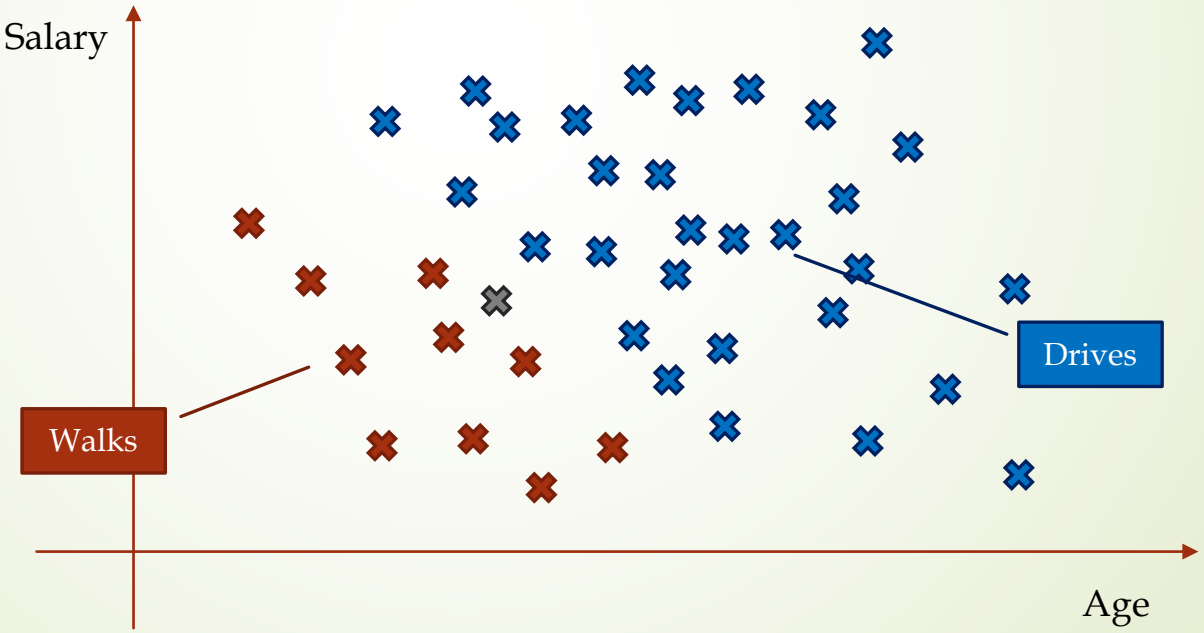
- Mach1: 30 pcs/hr
  - P(Mach1)=0.6
- Mach2: 20 pcs/hr
  - P(Mach2)=0.4
- Out of all produced parts: 1% are defective
  - P(Defect)=0.01
- Out of all defective parts: 50% came from mach1
  - P(Mach1 | Defect)=0.5, P(Mach2 | Defect)=0.5
- What is the probability that a part produced by mach2 is defective?
  - P(Defect | Mach2)=?

$$P(Defect|Mach2) = \frac{P(Mach2|Defect) \times P(Defect)}{P(Mach2)}$$

https://brohrer.mcknote.com/zh-Hant/statistics/how_bayesian_inference_works.html

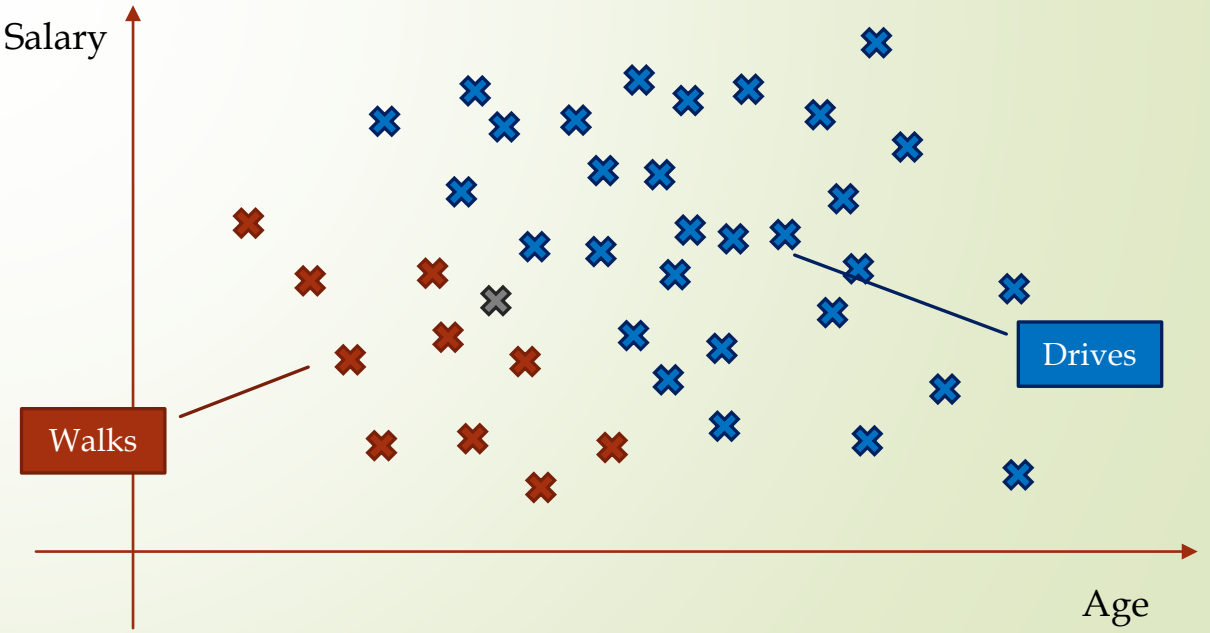
Naïve Bayes Classifier

- $P(Walks \mid X) = \frac{P(X \mid Walks) \times P(Walks)}{P(X)}$



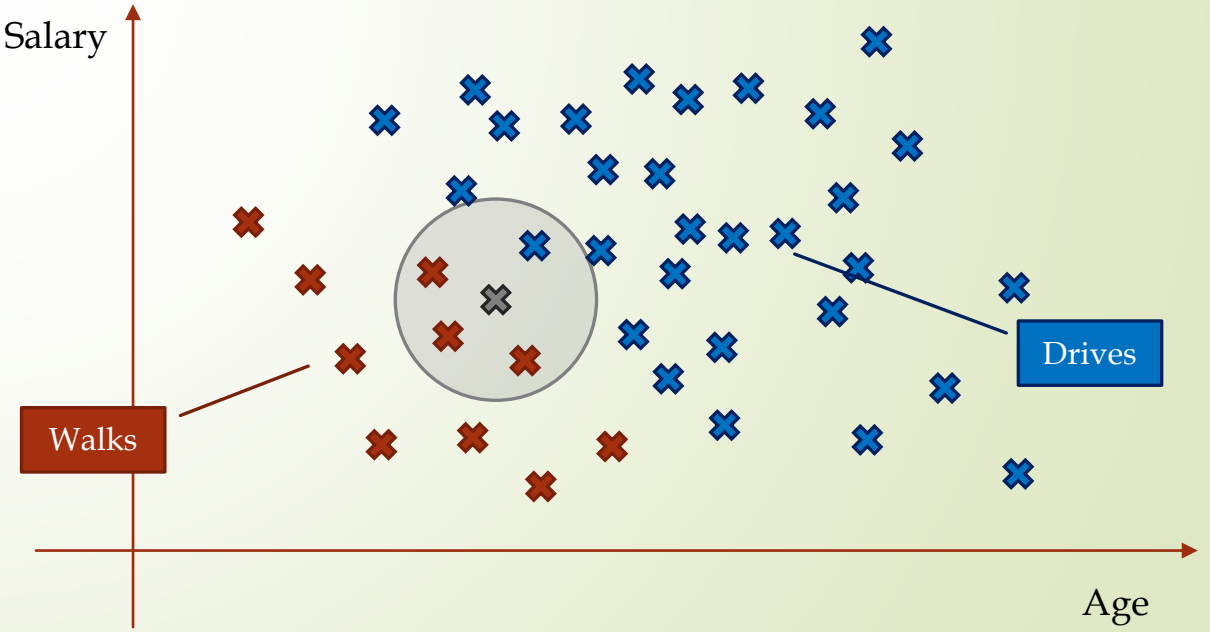
The First Step

- $P(Walks \mid X) = \frac{P(X \mid Walks) \times P(Walks)}{P(X)}$
  - $P(Walks) = \frac{10}{20+10} = \frac{10}{30}$



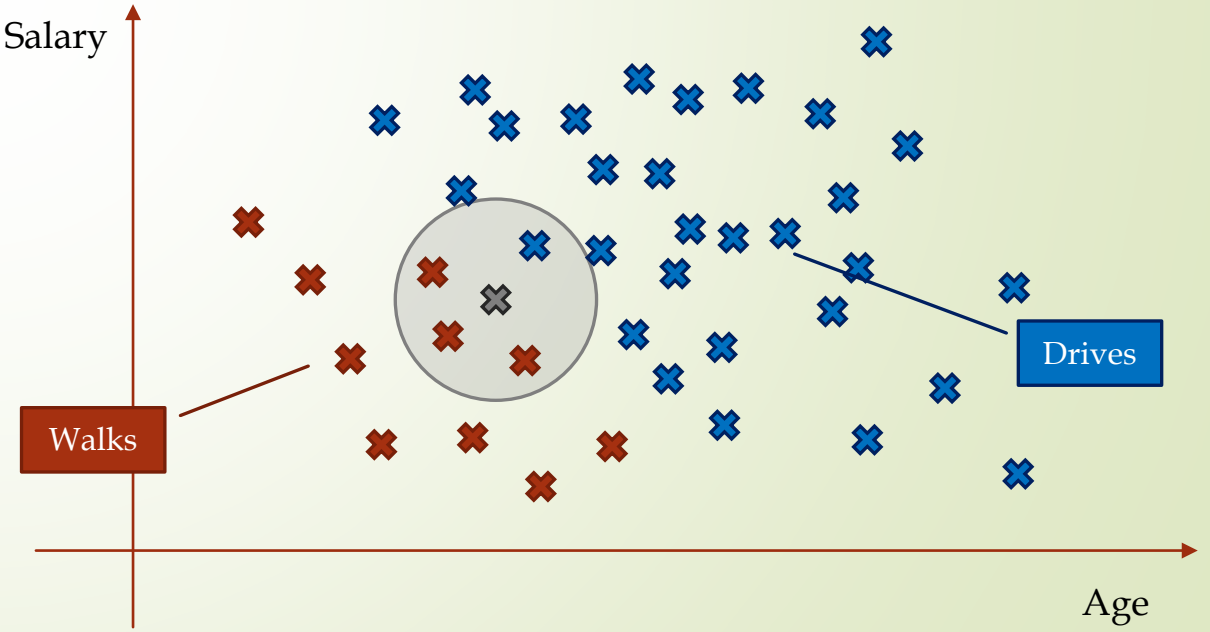
The Second Step

- $P(Walks \mid X) = \frac{P(X \mid Walks) \times P(Walks)}{P(X)}$
  - $P(\mathrm{X}) = \frac{4}{30}$



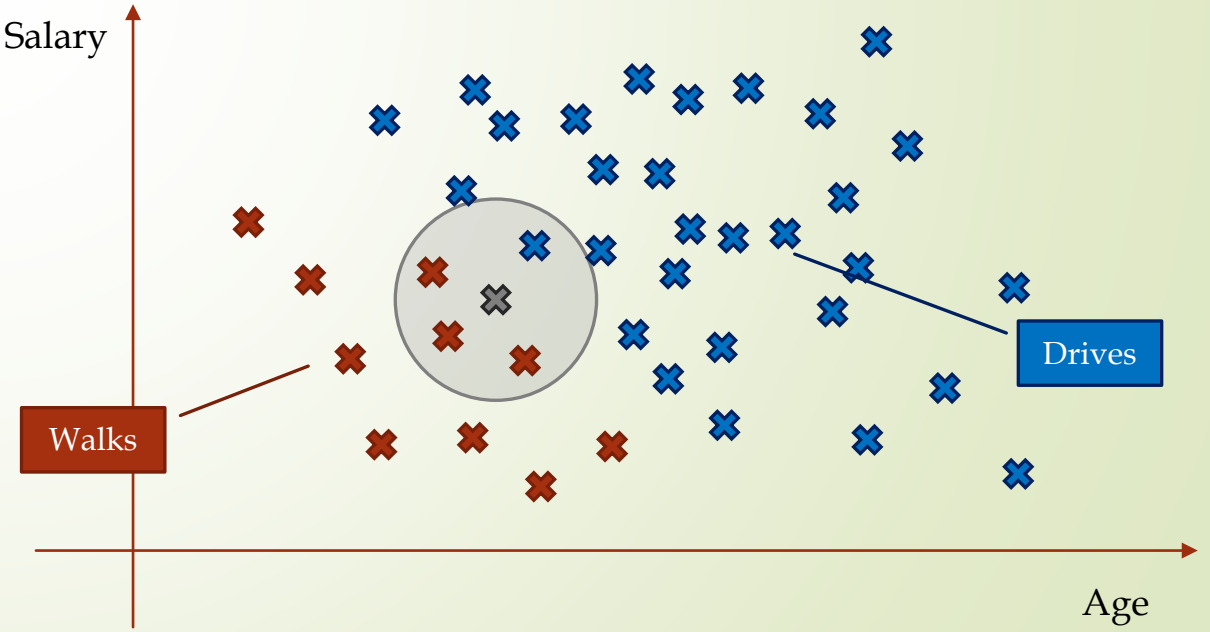
The Third Step

- $P(Walks \mid X) = \frac{P(X \mid Walks) \times P(Walks)}{P(X)}$
  - $P(\mathrm{X} \mid Walks) = \frac{3}{10}$

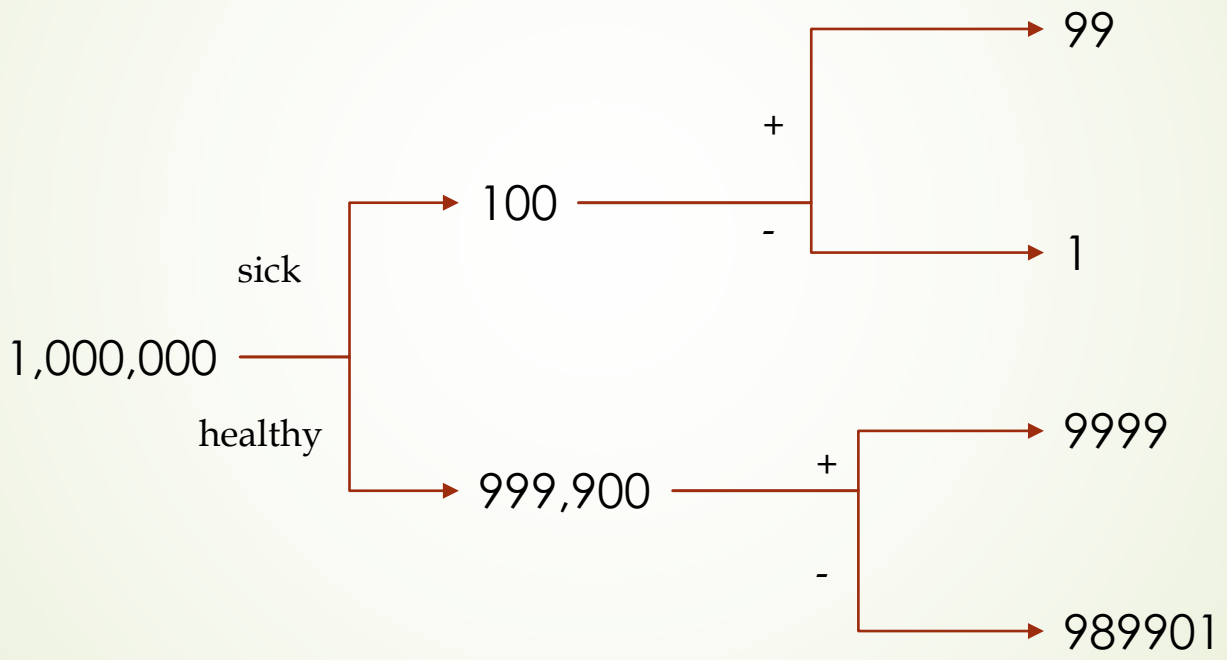


Posterior Probability

- $P(Walks \mid X) = \frac{P(X \mid Walks) \times P(Walks)}{P(X)}$
  - $P(Walks \mid X) = \frac{\frac{3}{10} \times \frac{10}{30}}{\frac{4}{30}} = \frac{3}{4}$



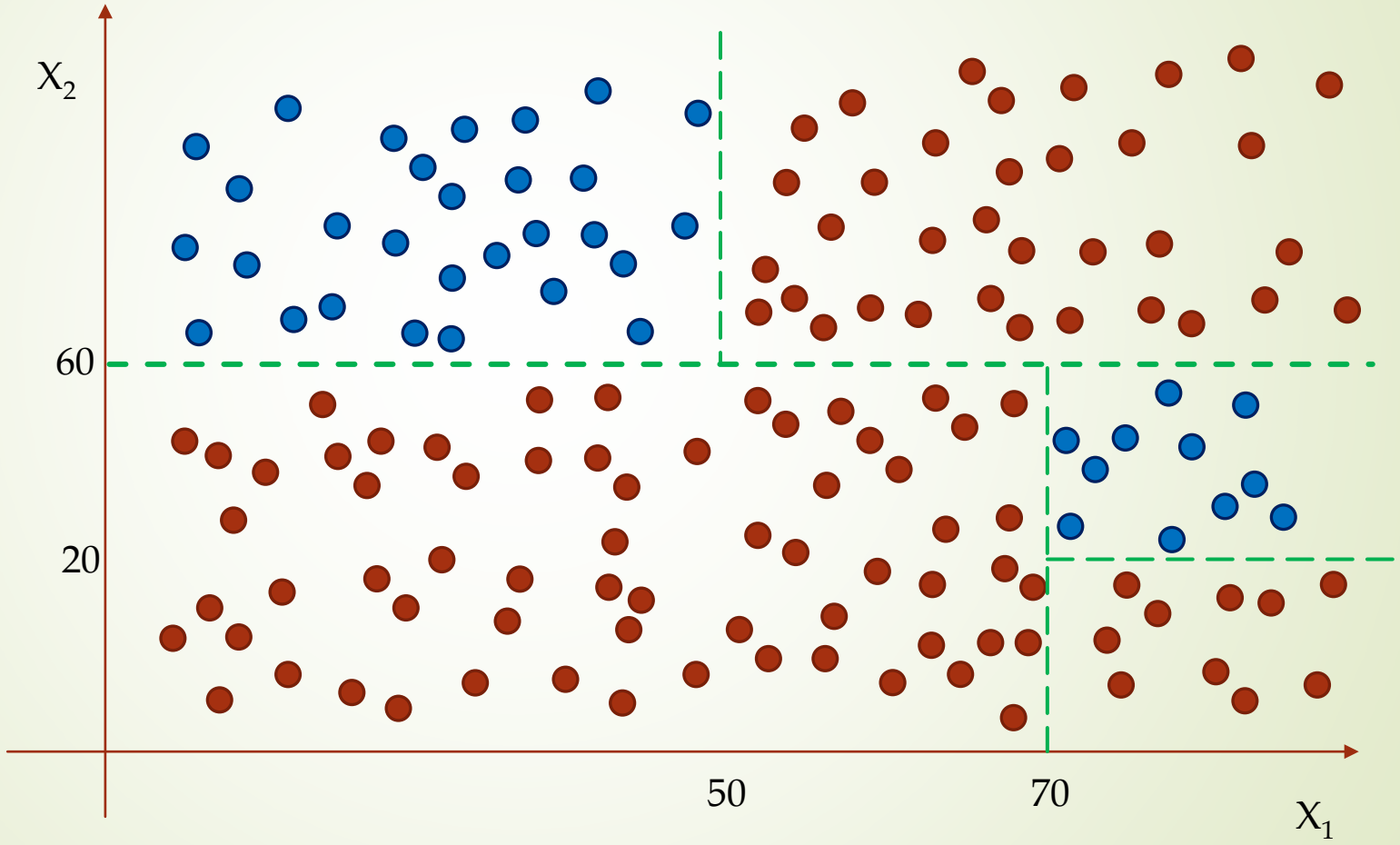
健康檢查



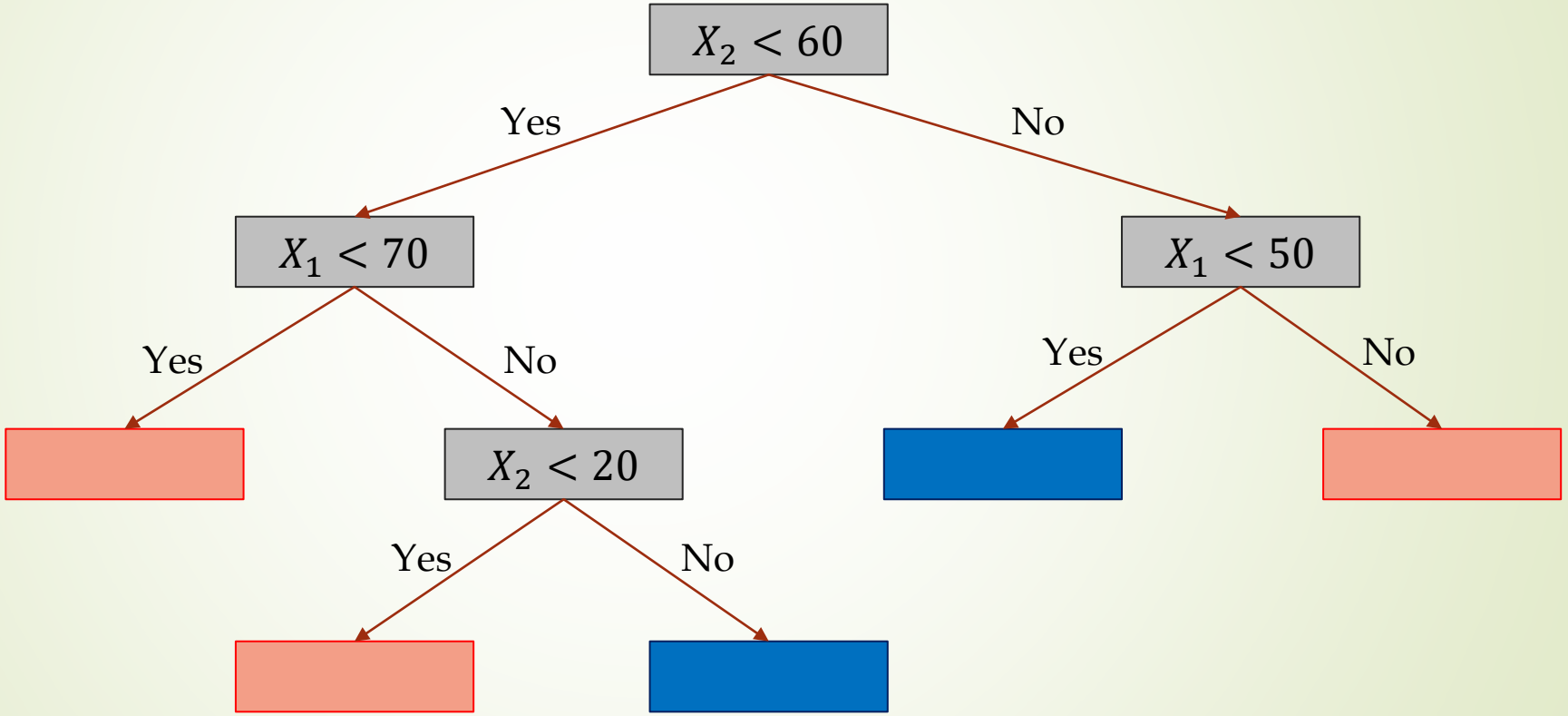
$$\frac{99}{99+9999} = 0.0098$$

決策樹

Decision Tree



Decision Tree

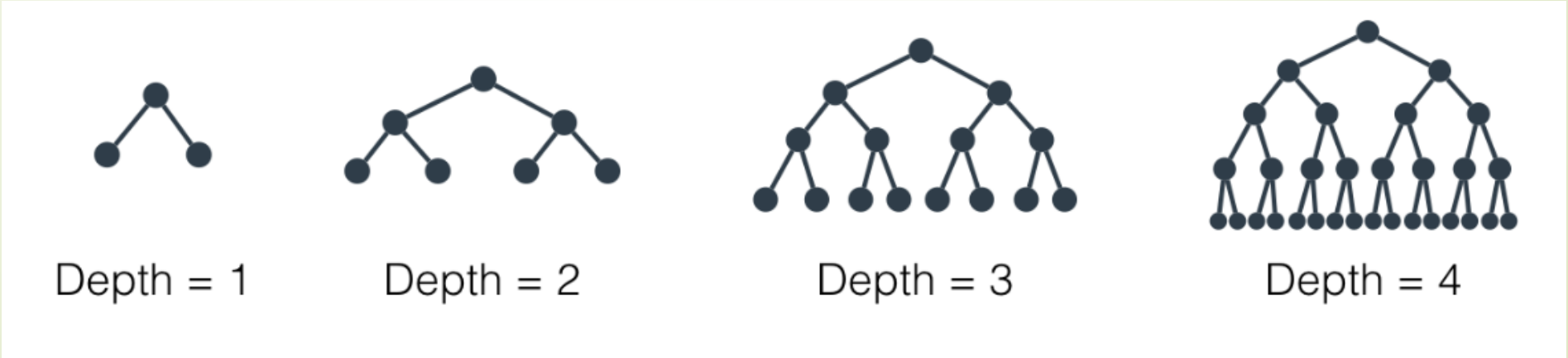


Hyperparameters for Decision Trees

- Maximum Depth
  - max_depth
- Minimum number of samples to split
  - min_samples_split
- Minimum number of samples per leaf
  - min_samples_leaf

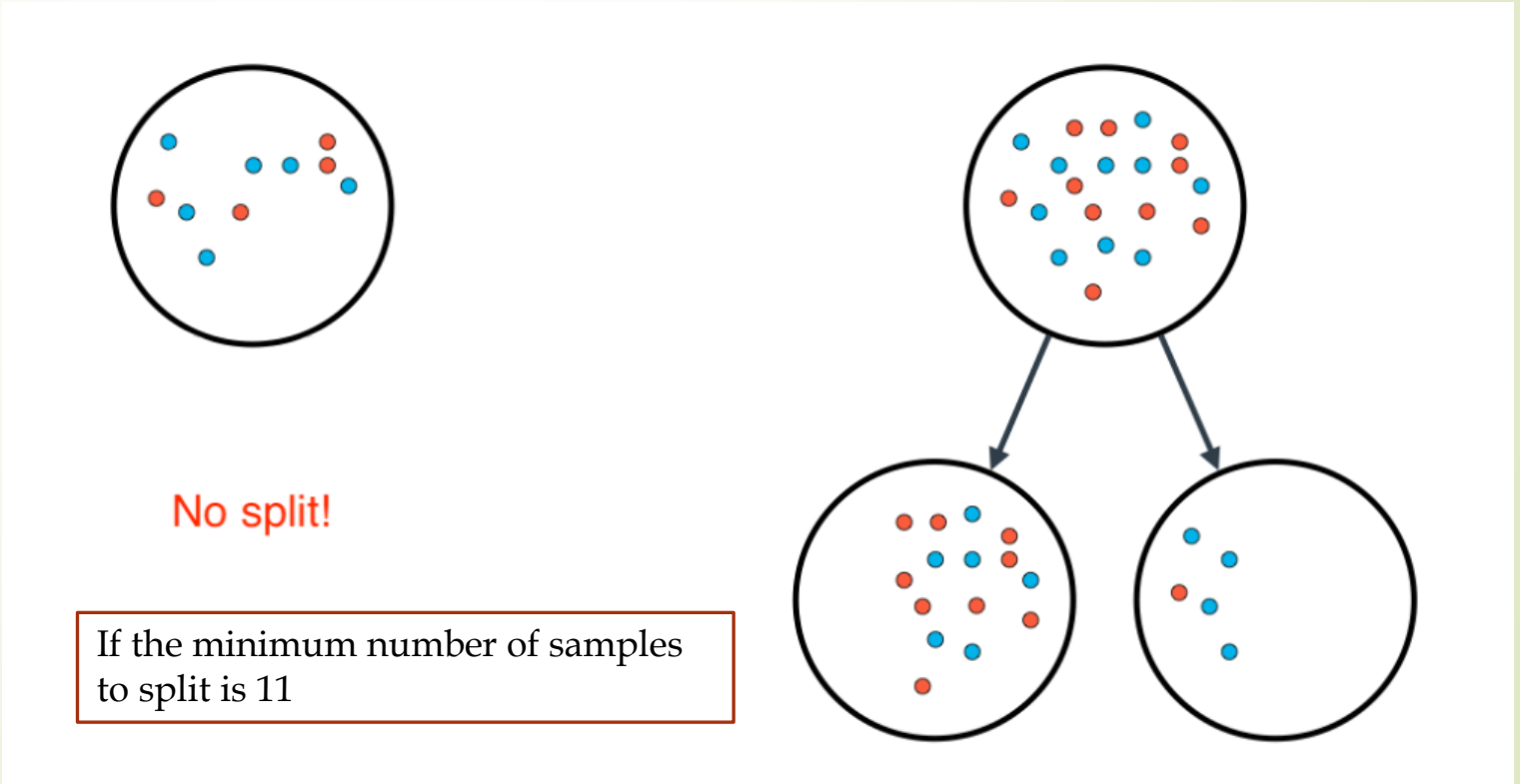
Maximum Depth

- The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.
- A tree of maximum length k can have at most 2^k leaves.



Minimum number of samples to split

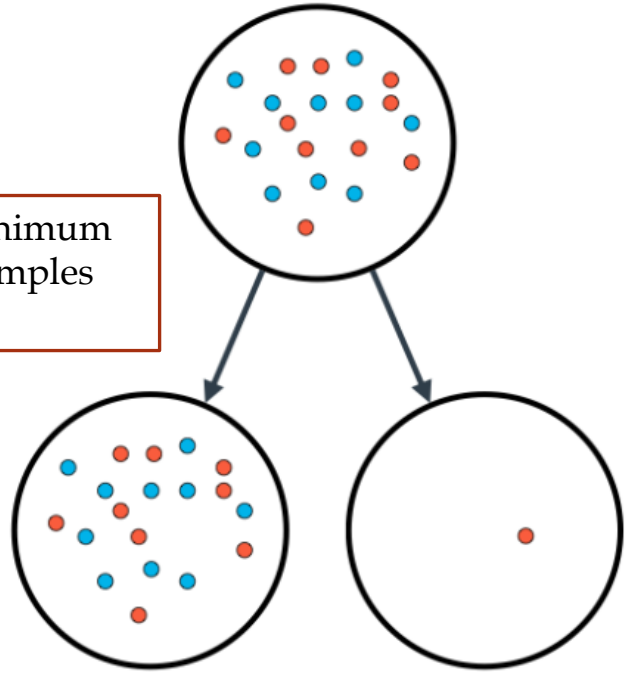
- The minimum number of samples required to split an internal node.

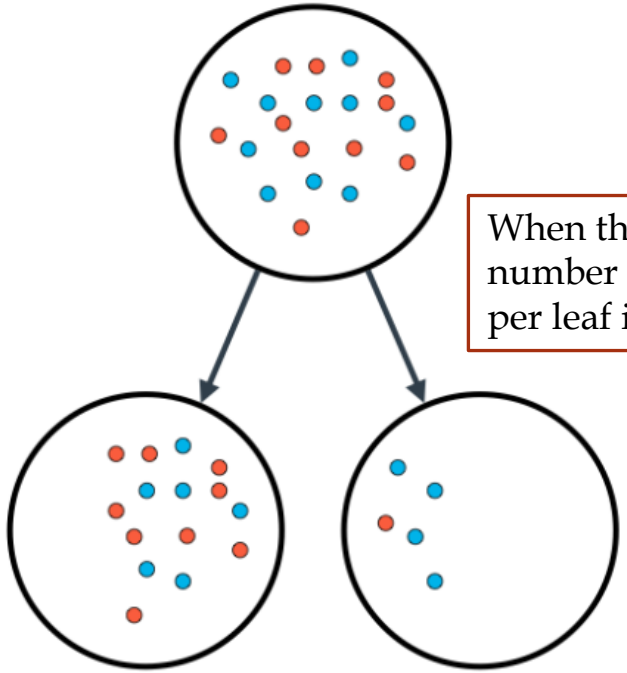


Minimum number of samples per leaf

- The minimum number of samples required to be at a leaf node. A split point at any depth will only be considered if it leaves at least min_samples_leaf training samples in each of the left and right branches. This may have the effect of smoothing the model, especially in regression.

When the minimum number of samples per leaf is 1





When the minimum number of samples per leaf is 5

The Features

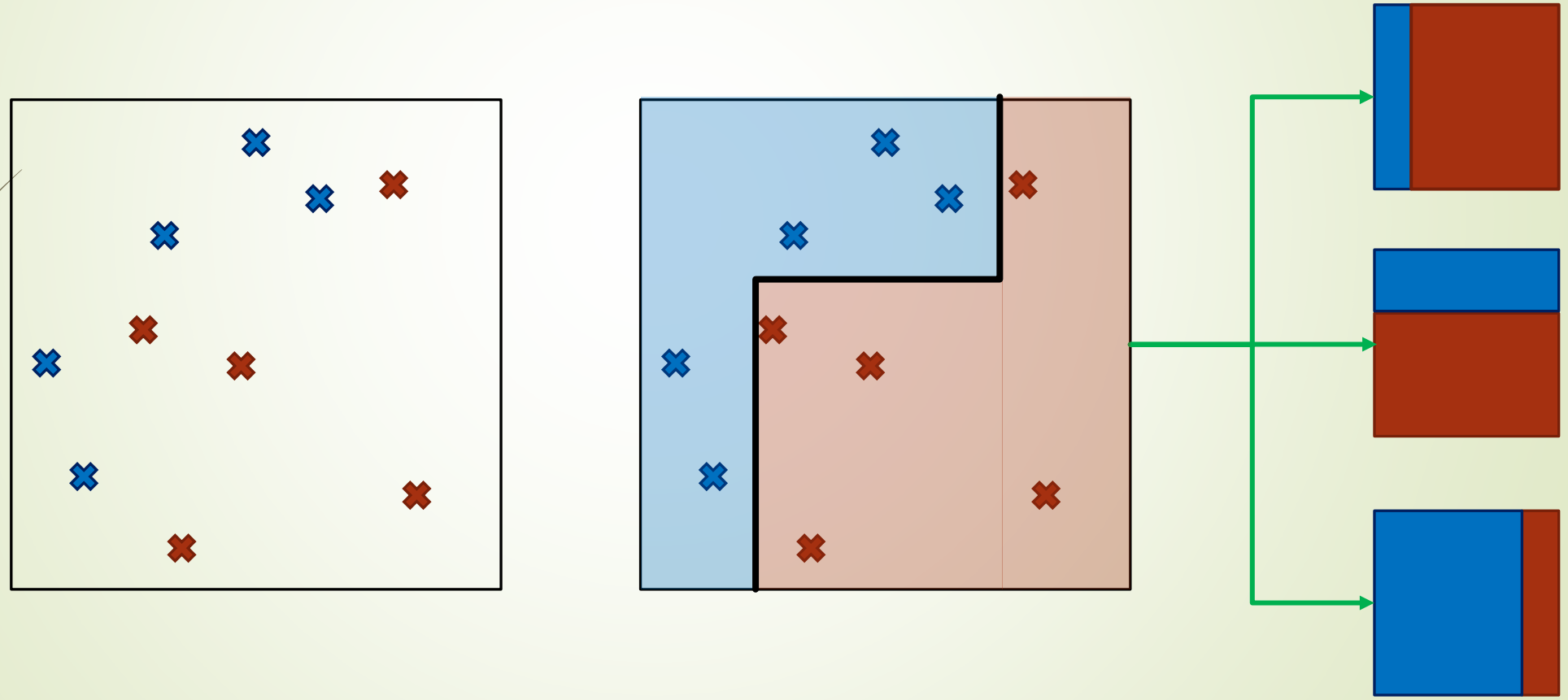
Feature	Underfitting / Overfitting
Small maximum depth	Underfitting
Large maximum depth	Overfitting
Small minimum samples per split	Overfitting
Large minimum samples per split	Underfitting

隨機森林

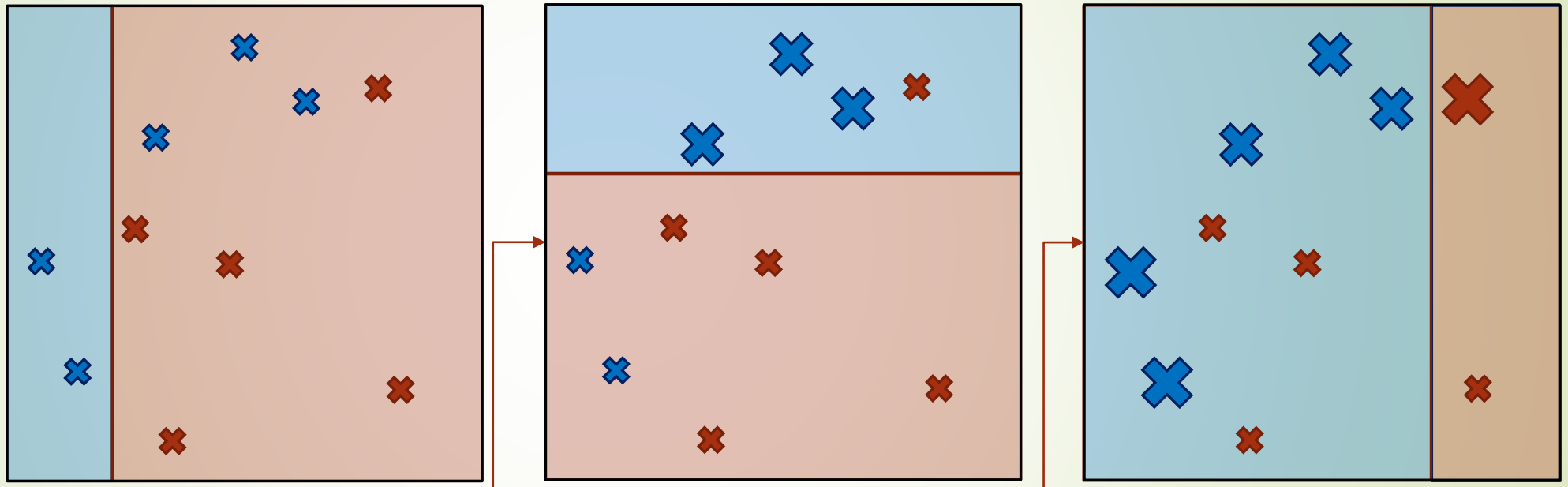
集成學習 (Ensemble learning)

- Ensemble Learning基本條件
  - 每個分類器之間應該要有差異
  - 每個分類器準確率需大於0.5
- Bagging (**Random Forest** : Bagging + Decision tree) [Bootstrap aggregating]
  - 從訓練資料中隨機抽取(取出後放回，n<N)樣本，訓練多個分類器，每個分類器的權重一致，最後用投票方式(Majority vote)得到最終結果
- Boosting (**GBDT** : Gradient Boost + Decision tree)
  - 將很多個弱的分類器(weak classifier)進行合成，變成一個強分類器(Strong classifier)，和Bagging不同的是分類器之間是有關聯性的。
  - 透過將舊分類器的錯誤資料權重提高，然後再訓練新的分類器，這樣新的分類器就會學習到錯誤分類資料(misclassified data)的特性，進而提升分類結果。
- AdaBoost (**Boosting Tree** : AdaBoost + Decision tree)
  - 是一種改進的Boosting分類算法

Bagging

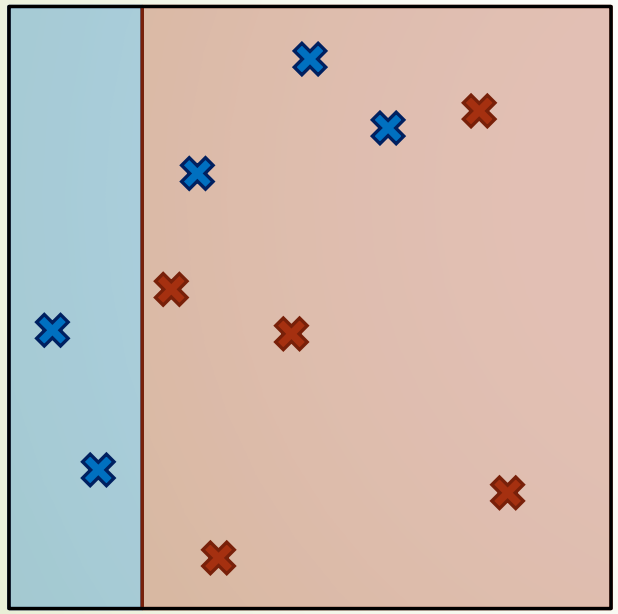


AdaBoost (training)

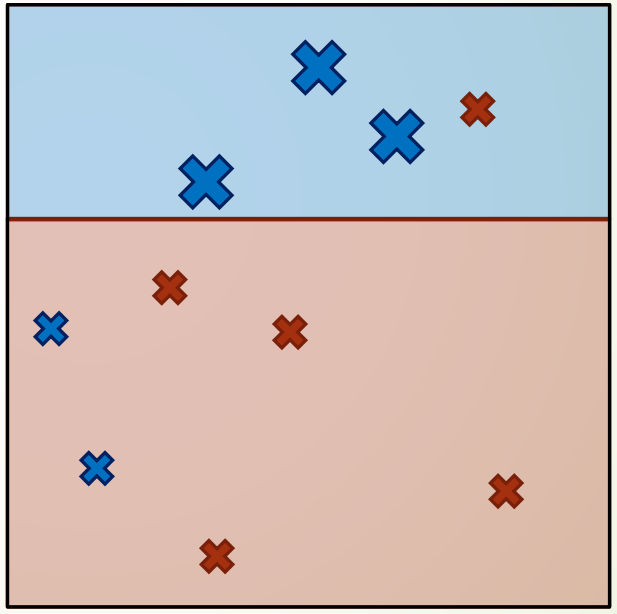


AdaBoost (weight)

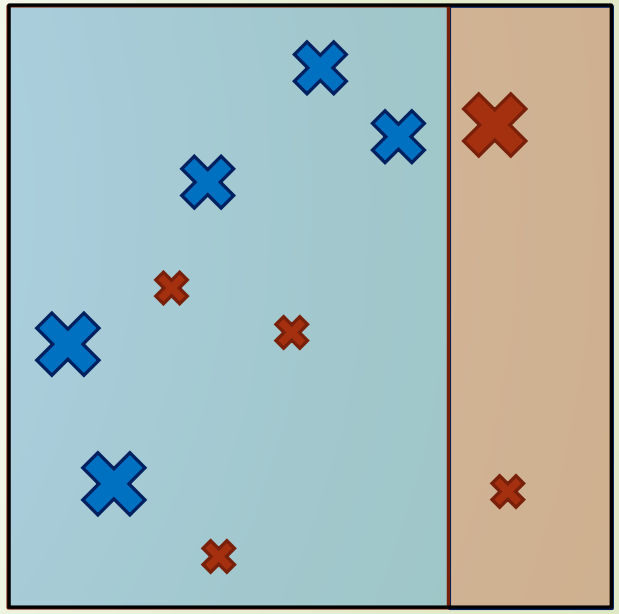
$$weight = \ln\left(\frac{accuracy}{1 - accuracy}\right) = \ln(\frac{\#correct}{\#incorrect})$$



$$weight = \ln\left(\frac{7}{3}\right) = 0.84$$



$$weight = \ln\left(\frac{11}{3}\right) = 1.30$$



$$weight = \ln\left(\frac{19}{3}\right) = 1.84$$

Bagging vs. Boosting

- 訓練樣本:
  - Bagging: 每一次的訓練集是隨機抽取(每個樣本權重一致)，抽出可放回，以獨立同分布選取的訓練樣本子集訓練弱分類器。
  - Boosting: 每一次的訓練集不變，訓練集之間的選擇不是獨立的，每一次選擇的訓練集都是依賴上一次學習得結果，根據錯誤率(給予訓練樣本不同的權重)取樣。
- 分類器:
  - Bagging: 每個分類器的權重相等。
  - Boosting: 每個弱分類器都有相應的權重，對於分類誤差小的分類器會有更大的權重。
- 每個分類器的取得:
  - Bagging: 每個分類器可以並行生成。
  - Boosting: 每個弱分類器只能依賴上一次的分類器順序生成。

Reference

分類模型性能評價與選擇

Introduction

- Problem ↔ Tools ↔ Measurement Tools
- Regression Measures
  - Mean Absolute Error (MAE)
  - Mean Squared Error (MSE)
  - R2 Score
- Classification Measures
  - Accuracy
  - Precision
  - Recall
  - F-Beta Score
  - ROC Curve & AUC

混淆矩陣 (Confusion Matrix)

$$Accruacy\ Rate = \frac{Correct}{Total}$$

$$Error\ Rate = \frac{Incorrect}{Total}$$

	Diagnosed Sick	Diagnosed Healthy
Sick	1,000 True Positives	200 False Negatives
Healthy	800 False Positives	8,000 True Negatives

$$Accruacy\ Rate = \frac{9{,}000}{10{,}000} = 90\%$$

	Spam Folder	Inbox
Spam	100 True Positives	170 False Negatives
Not Spam	30 False Positives	700 True Negatives

$$Accruacy\ Rate = \frac{800}{1{,}000} = 80\%$$

準確率悖論 (Accuracy Paradox)

另例，如果A類的發病率占主導地位，在99%的病例中發現，則預測每個病例都是A類，其準確度為99%

	Diagnosed Sick	Diagnosed Healthy
Sick	100 True Positives	50 False Negatives
Healthy	150 False Positives	9,700 True Negatives

$$Accruacy\ Rate = \frac{9{,}800}{10{,}000} = 98\%$$

	Diagnosed Sick	Diagnosed Healthy
Sick	0 True Positives	150 False Negatives
Healthy	0 False Positives	9,850 True Negatives

$$Accruacy\ Rate = \frac{9{,}850}{10{,}000} = 98.5\%$$

偽陽性與偽陰性 (Fales Positives & Fales Negatives)

False Positives ok
False Negatives NOT ok

	Diagnosed Sick	Diagnosed Healthy
Sick	1,000 True Positives	200 False Negatives
Healthy	800 False Positives	8,000 True Negatives

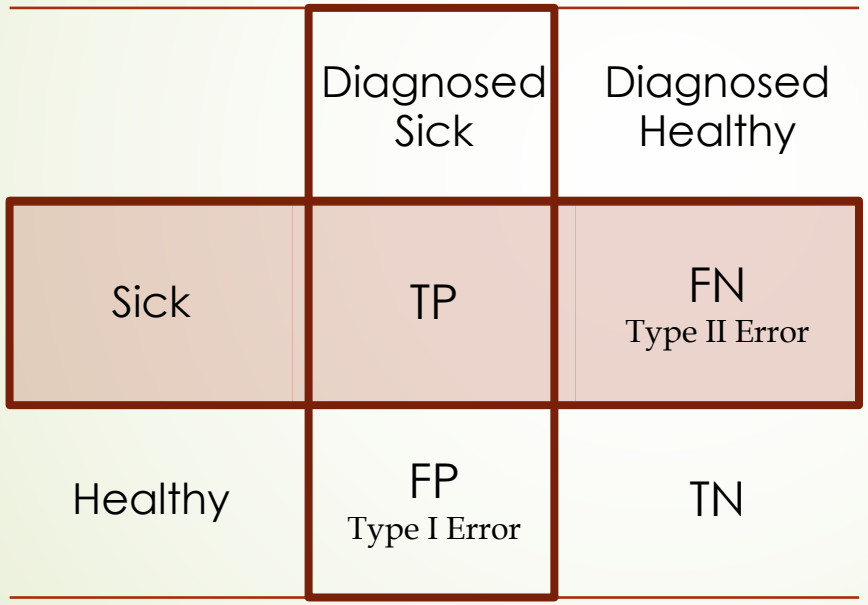
High Recall

False Positives NOT ok
False Negatives ok

	Spam Folder	Inbox
Spam	100 True Positives	170 False Negatives
Not Spam	30 False Positives	700 True Negatives

High Precision

其他指標



$$Accuracy = (TP + TN)/Total$$

$$Precision = TP/(TP + FP)$$

重視Type I Error, ex 門禁系統

$$Recall = TP/(TP + FN)$$

重視Type II Error, ex 疾病檢查

$$F_\beta = (1 + \beta^2) \times \frac{precision \times recall}{(\beta^2 \times precision) + recall}$$

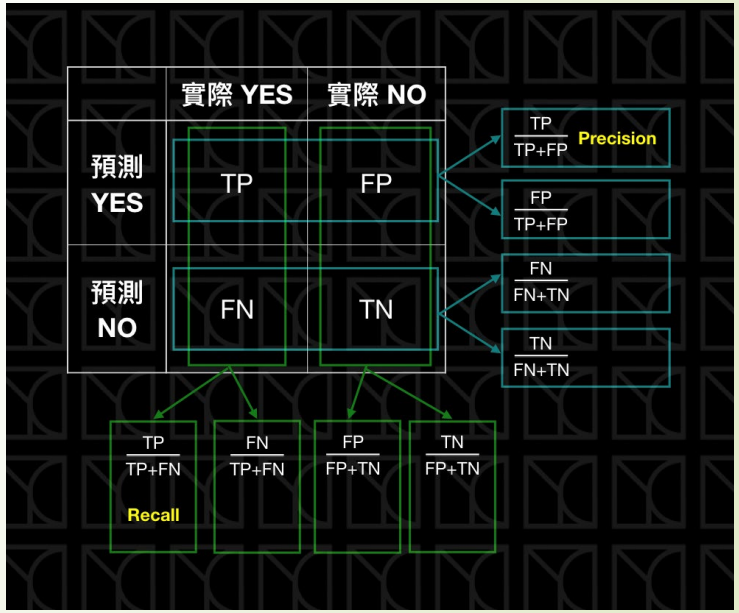
β > 1, 為了提高F_β，偏向提升Recall
β = 1, F1 score
β < 1, 為了提高F_β，偏向提升Precision

F-beta Score

Precision	$F_{0.5}$ Score	F_1 Score	F_2 Score	Recall
$\beta = 0$	門禁系統 垃圾郵件		疾病檢查 飛機零件	$\beta \to \infty$
	Focus on Type I (False Alarm)		Focus on Type II (False Pass)	
	Precision		Recall	

CAP & ROC

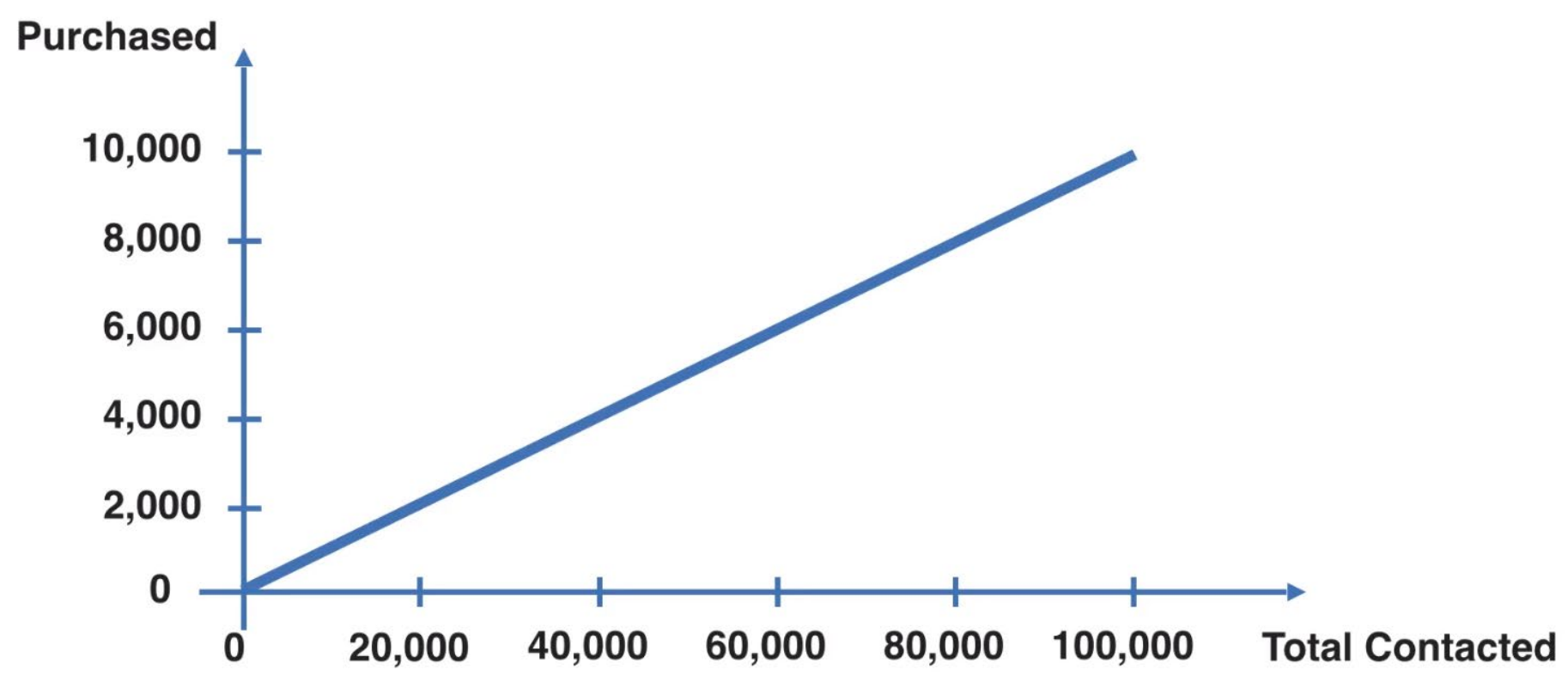
- CAP
  - Cumulative Accuracy Profile
  - TP vs. (TP+TN+FP+FN)
- ROC
  - Receiver Operating Characteristic
  - TP/(TP+FN) vs. FP/(FP+TN)



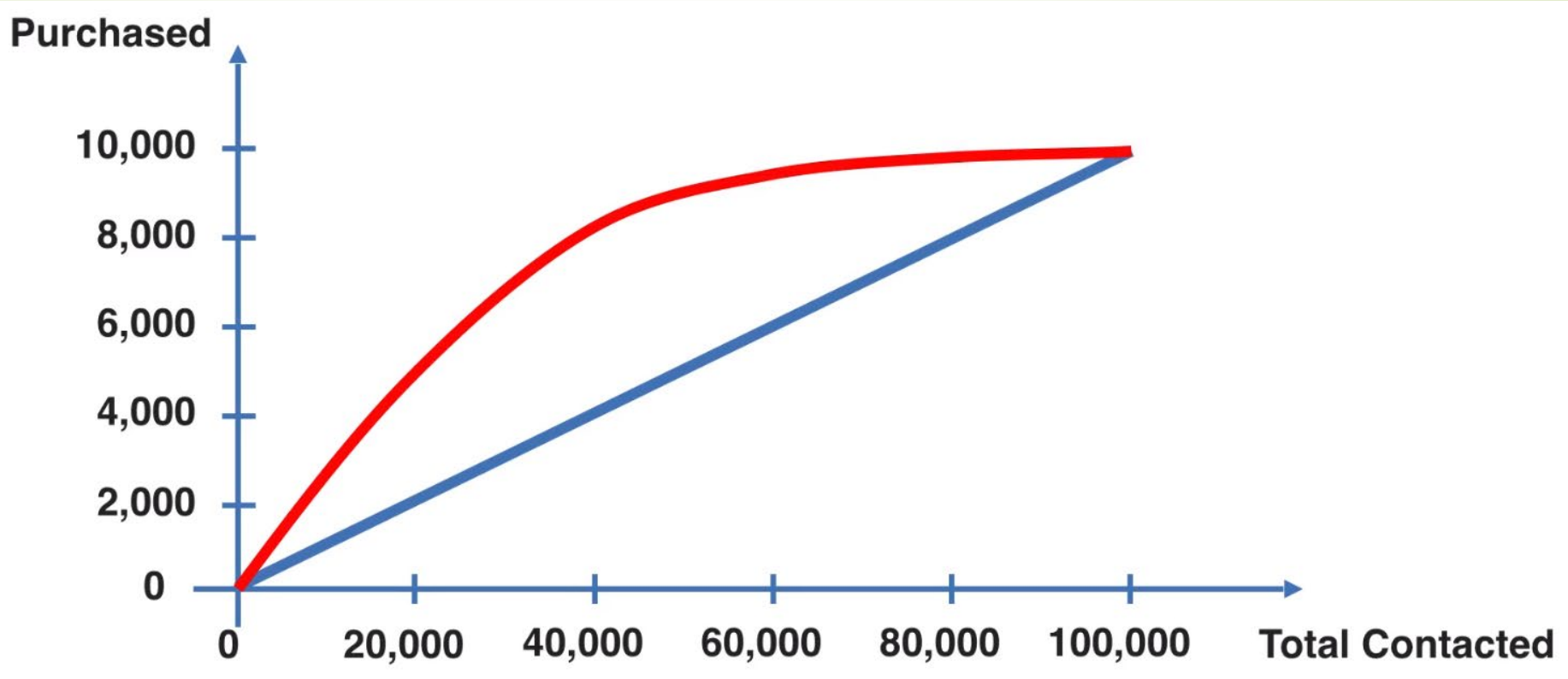
The CAP is distinct from the receiver operating characteristic (ROC) curve

- ROC plots the true-positive rate against the false-positive rate.
- CAPs are used in robustness evaluations of classification models.

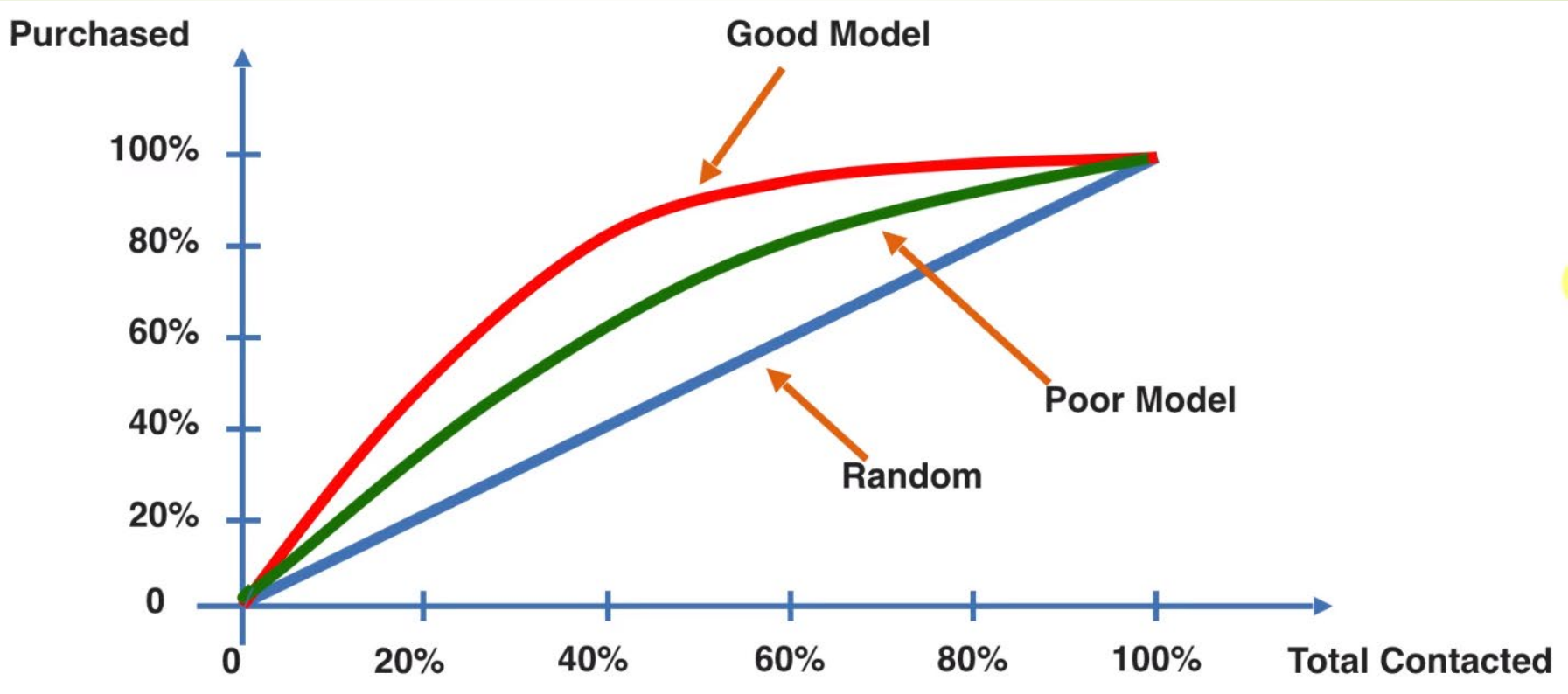
累計準確曲線 (Cumulative Accuracy Profile Curve)



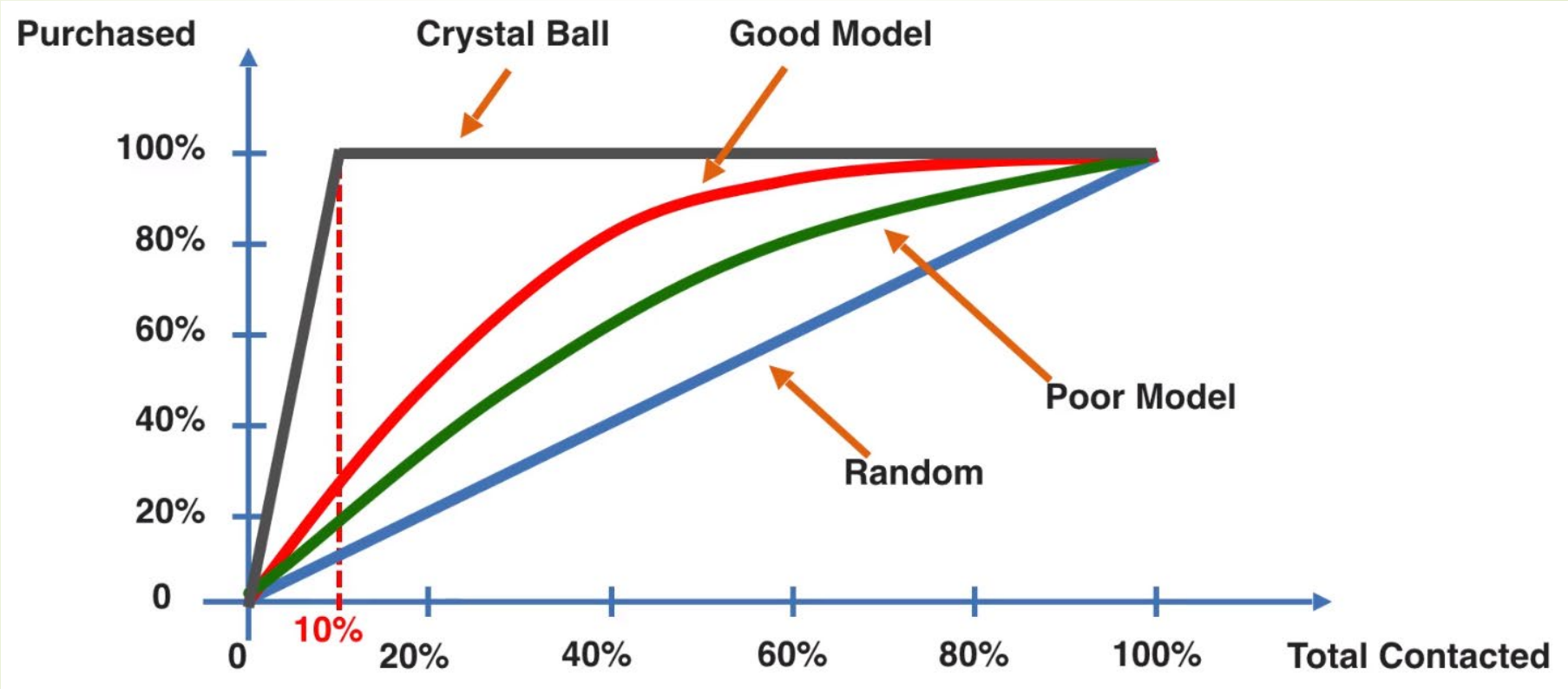
累計準確曲線


Purchased
10,000
8,000
6,000
4,000
2,000
0
0
20,000
40,000
60,000
80,000
100,000
Total Contacted

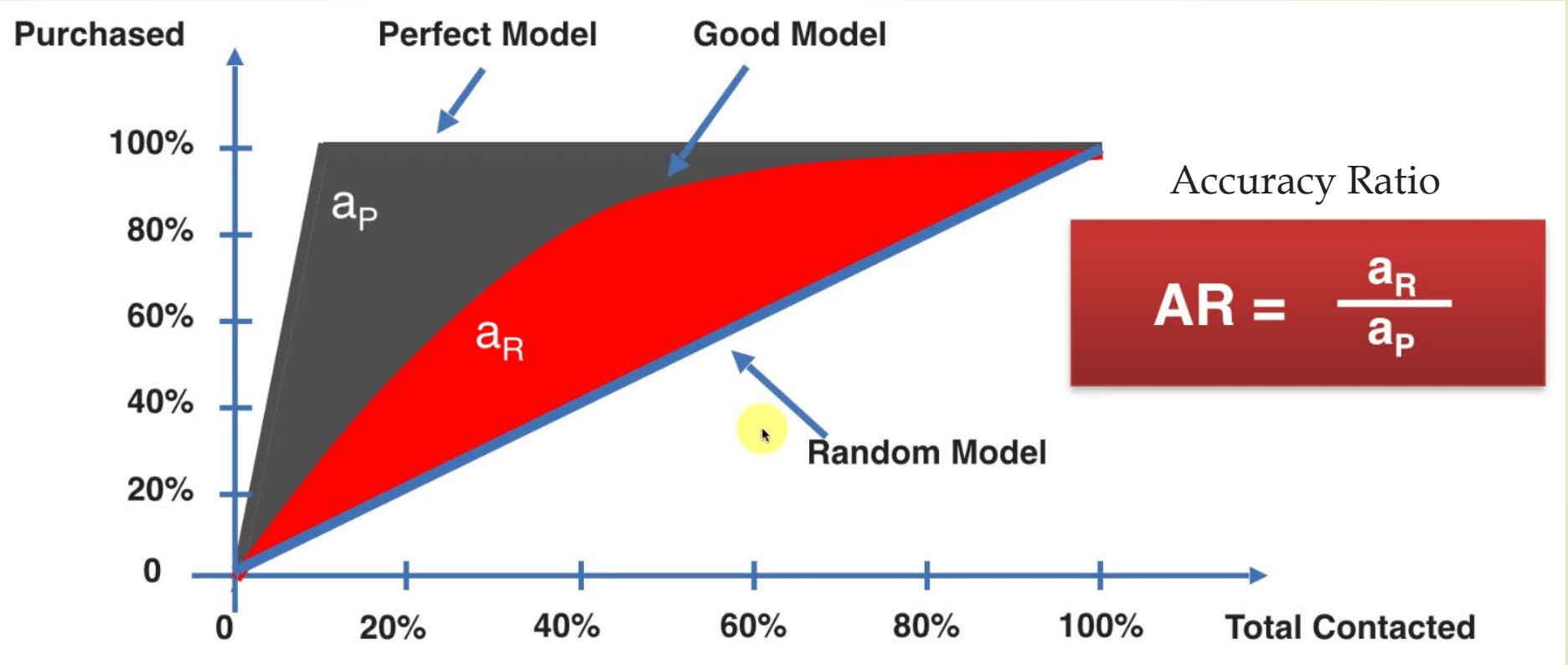
累計準確曲線


Purchased
Good Model
100%
80%
60%
40%
20%
0
Poor Model
Random
0
20%
40%
60%
80%
100%
Total Contacted

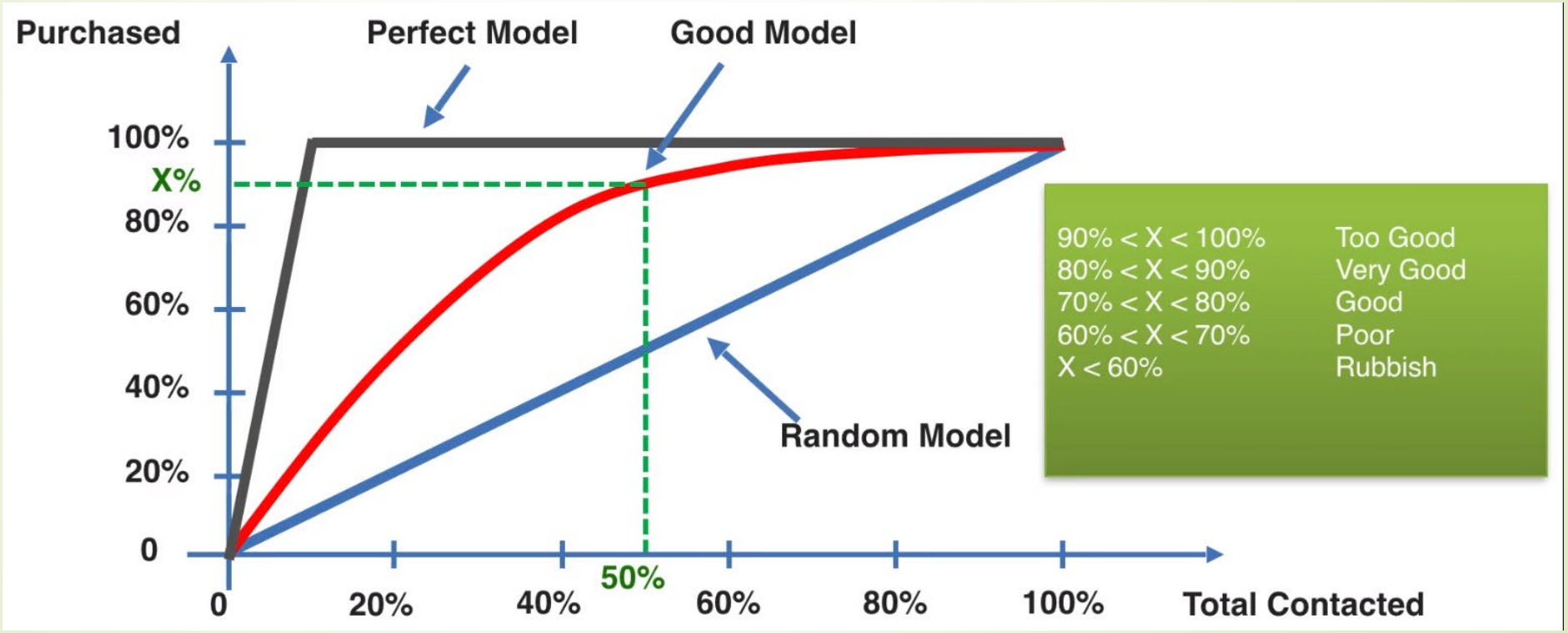
累計準確曲線


Purchased
Crystal Ball
Good Model
100%
80%
60%
40%
20%
0
Poor Model
Random
10%
0
20%
40%
60%
80%
100%
Total Contacted

累計準確曲線分析 (CAP Curve Analysis)



累計準確曲線分析

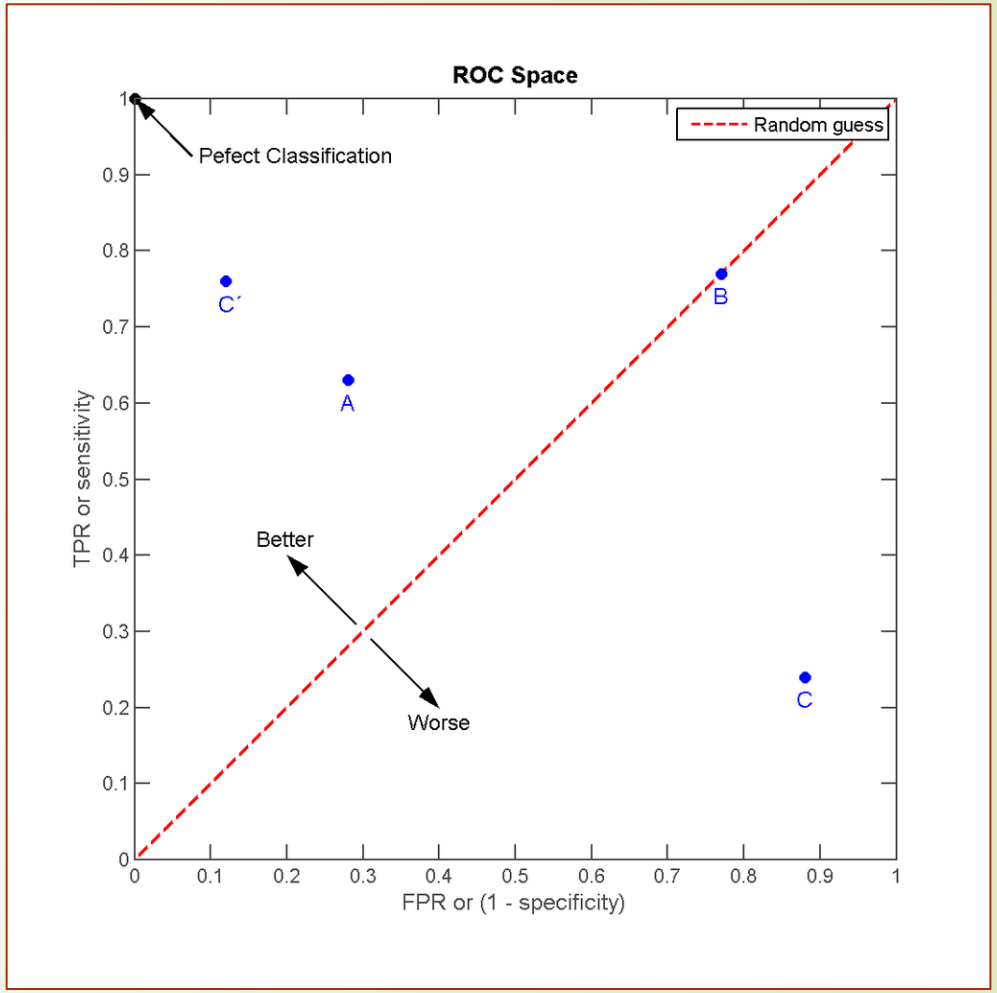

Purchased
Perfect Model
Good Model
100%
X%
80%
60%
40%
20%
0
0
20%
40%
50%
60%
80%
100%
Total Contacted
Random Model
90% < X < 100%
Too Good
80% < X < 90%
Very Good
70% < X < 80%
Good
60% < X < 70%
Poor
X < 60%
Rubbish

接收者操作特徵 (Receiver Operating Characteristic)

A			B			C		
TP=63	FP=28	91	TP=77	FP=77	154	TP=24	FP=88	112
FN=37	TN=72	109	FN=23	TN=23	46	FN=76	TN=12	88
100	100	200	100	100	200	100	100	200
TPR = 0.63			TPR = 0.77			TPR = 0.24		
FPR = 0.28			FPR = 0.77			FPR = 0.88		
ACC = 0.68			ACC = 0.50			ACC = 0.18		

$$\text{True Positive Rate} = \frac{True\ Positives}{All\ Positives}$$

$$\text{False Positive Rate} = \frac{False\ Positives}{All\ Negatives}$$

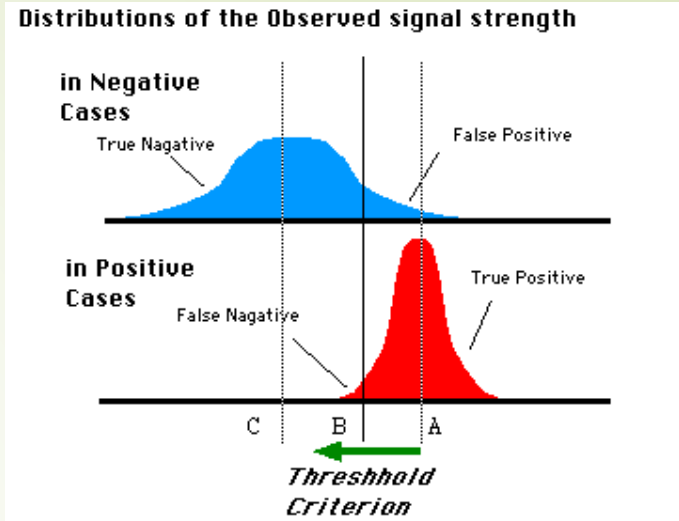


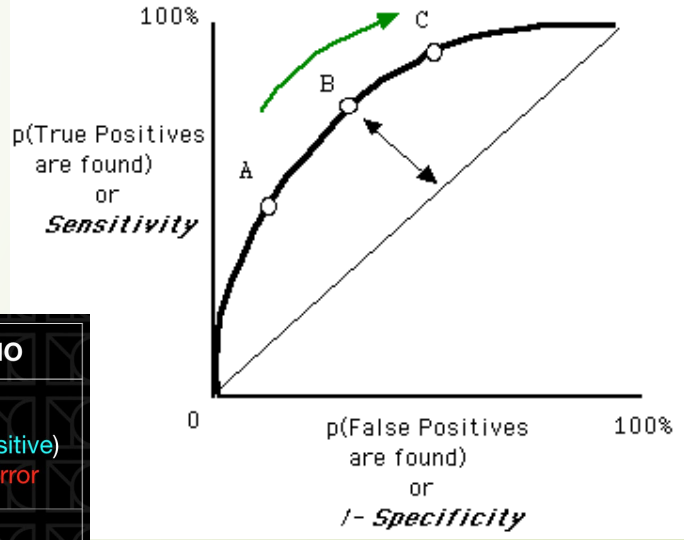
ROC曲線

- 將同一模型每個閾值 的 (FPR, TPR) 座標都畫在ROC空間裡，就成為特定模型的ROC曲線
  - 當閾值設定為最高時，必得出ROC座標系左下角的點 (0, 0)
  - 當閾值設定為最低時，必得出ROC座標系右上角的點 (1, 1)
  - 隨著閾值調低，ROC點 往右上（或右 / 或上）移動，或不動；但絕不會往左下(或左 / 或下)移動

$$\text{True Positive Rate} = \frac{True\ Positives}{All\ Positives}$$

$$\text{False Positive Rate} = \frac{False\ Positives}{All\ Negatives}$$

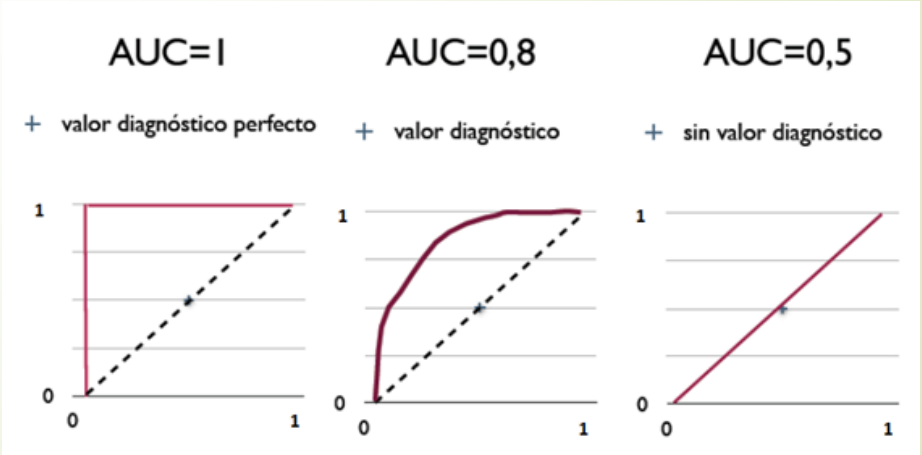




	實際 YES	實際 NO
預測 YES	TP (True Positive)	FP (False Positive) Type I Error
預測 NO	FN (False Negative) Type II Error	TN (True Negative)

AUC (Area Under the Curve)

- ROC曲線下方的面積
  - AUC = 1，是完美分類器，採用這個預測模型時，存在至少一個閾值能得出完美預測。絕大多數預測的場合，不存在完美分類器。
  - 0.5 < AUC < 1，優於隨機猜測。這個分類器（模型）妥善設定閾值的話，能有預測價值。
  - AUC = 0.5，跟隨機猜測一樣（例：丟銅板），模型沒有預測價值。
  - AUC < 0.5，比隨機猜測還差；但只要總是反預測而行，就優於隨機猜測。



感謝聆聽