COS10022 –DATA SCIENCE PRINCIPLES

Dr Pei-Wei Tsai (Lecturer, Unit Convenor)
ptsai@swin.edu.au, EN508d

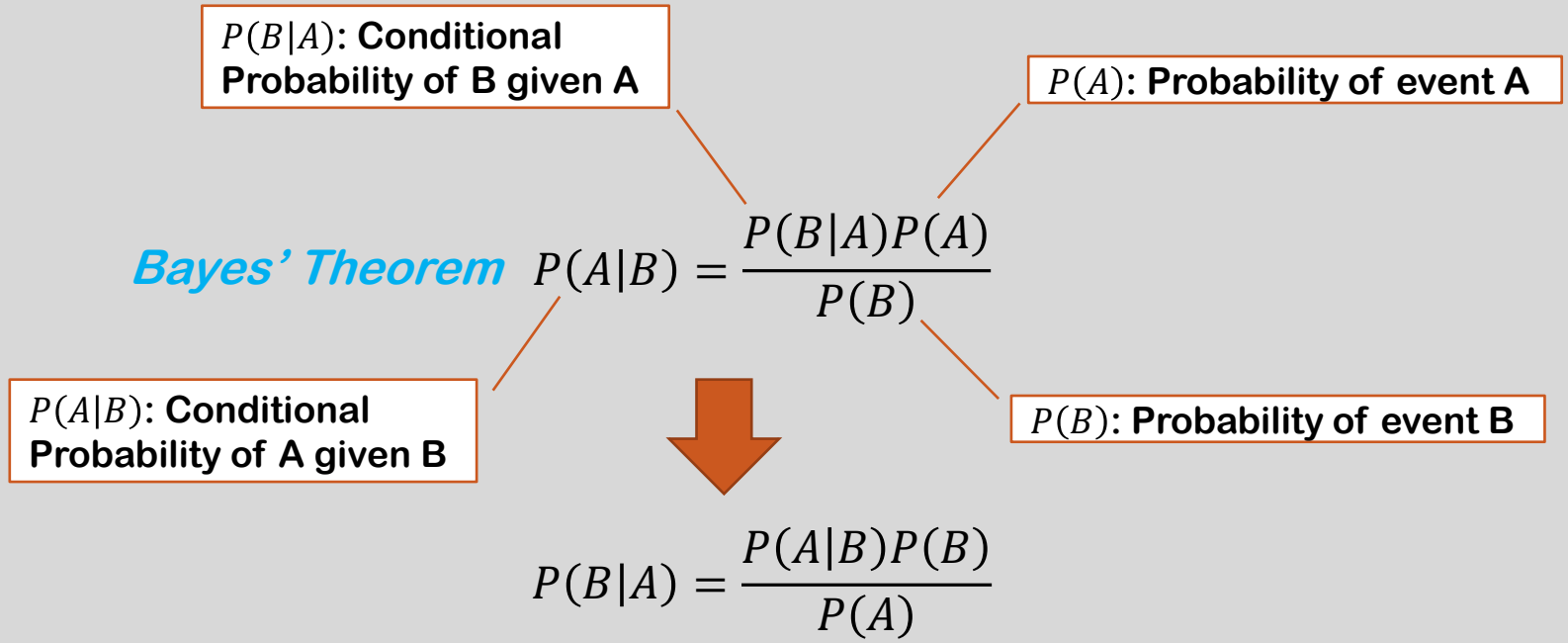
SWINBURNE UNIVERSITY OF TECHNOLOGY

BAYES' THEOREM

SWINBURNE UNIVERSITY OF TECHNOLOGY

Bayes' Theorem

- If you know the probability of A happens conditional on B happens, the probabilities of A happens, and B happens, independently, you can derive the probability of B happens conditional on A happens.



Why Using Bayes' Theorem?

Cost of Data Collection.

Timeless.

Exercise 1

- If we want to find out a patient's probability of having liver disease if he or she is an alcoholic.
- **A** could mean the event "patient has liver disease."
  - Past data tells you that 10% of patients entering your clinic have liver disease. → $P(A) = 0.1$
- **B** could mean the litmus test that "patient is an alcoholic."
  - Past data tells you that 5% of patients are alcoholics. → $P(B) = 0.05$
- You might also know that among those patients diagnosed with liver disease, 7% are alcoholics.
  - The probability that a patient is alcoholic given that he/she has liver disease is 7%.
    → $P(B|A) = 0.07$
- Bayes' Theorem tells you that $P(A|B) = \frac{(0.07 \times 0.1)}{0.05} = 0.14$
  - Thus, if the patient is an alcoholic, the chance of the patient having liver disease is 14%.

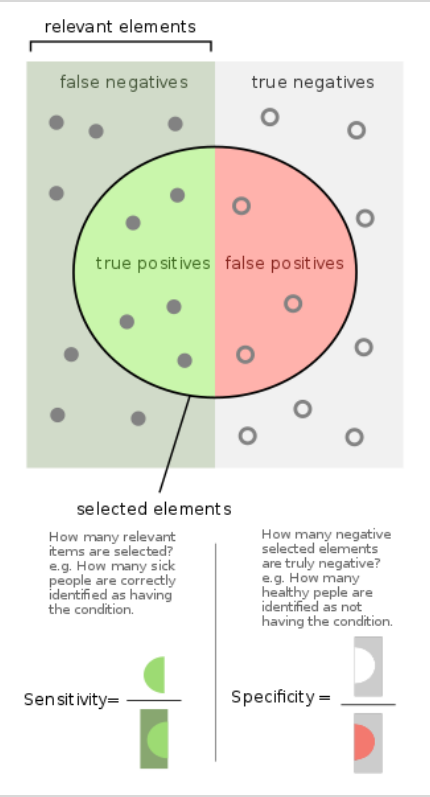
MODEL EVALUATION

SWINBURNE UNIVERSITY OF TECHNOLOGY

Supervised Models: Metrics and Methods

Popular metrics for evaluating the performance of supervised models:

1. **Accuracy**
2. **True Positive Rate (TPR) – also called Sensitivity or Recall.**
3. **True Negative Rate (TNR) – also called Specificity or Selectivity.**
4. **False Positive Rate (FPR) – also called Fall-out.**
5. **False Negative Rate (FNR) – also called Missing rate.**
6. **Precision:** $\frac{\text{True positive}}{\text{Predictive Condition Positive}}$
7. **Area Under the Curve (AUC) – also called the plasma concentration-time profile.**



These metrics can be calculated by utilising a **confusion matrix**.

https://en.wikipedia.org/wiki/Sensitivity_and_specificity

Supervised Models: Metrics and Methods

True Positives (TP):	the number of **positive** instances that a classifier correctly classifies as **positive**.
False Positives (FP):	the number of instances that a classifier identified as **positive** but in reality, are **negative**.
True Negatives (TN):	the number of **negative** instances that a classifier correctly identifies as **negative**.
False Negatives (FN):	the number of instances classified as **negative** but in reality, are **positive**.

TP and TN are correct predictions. A good classifier should have large TP and TN, and small (ideally, zero) numbers of FP and FN.

Supervised Models: Metrics and Methods

Example 1. A confusion matrix of Naïve Bayes classifier for 100 customers in predicting whether they would subscribe to the term deposit.

		Predicted Class		
		Subscribed	Not Subscribed	Total
Actual Class	Subscribed	3 (correct prediction)	8 (wrong prediction)	11
	Not Subscribed	2 (wrong prediction)	87 (correct prediction)	89
Total		5	95	100

QUESTION

Which type of error is more important?

COVID-19 CRP Test

- Even in different stage, the important index is also different.