Chapter 8
Bayesian Learning:

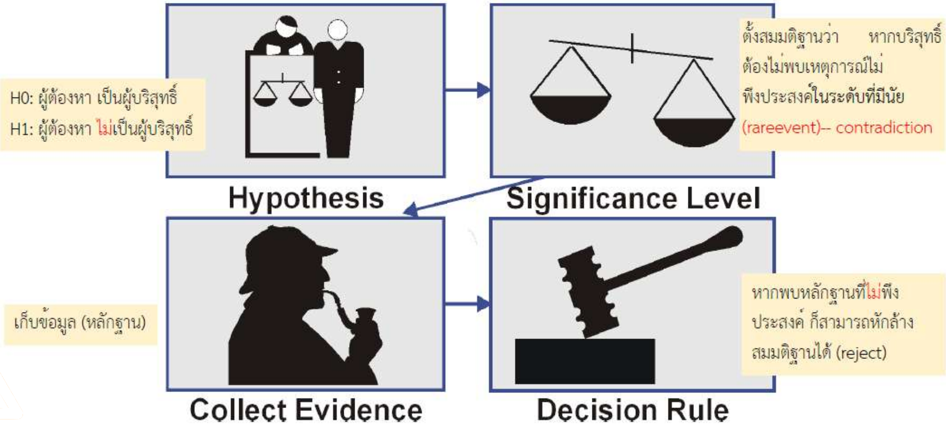
Associate Professor Yachai Limpiyakorn, Ph.D.

Part 1

Naïve Bayes

Applications:
- In finance, rate the risk of lending money to potential borrowers
- In medicine, determine the effectiveness of medication

Hypothesis Testing

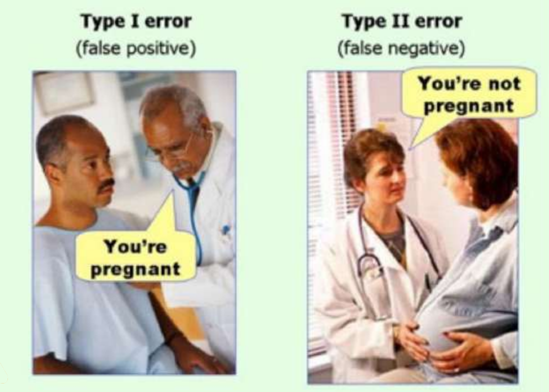


DECISION BASED ON SAMPLE	TRUTH ABOUT THE POPULATION: H0 is true	H0 is false
FAIL TO REJECT H0	Correct Decision (prob = 1 - α)	Type II Error -fail to reject H0 when it is false (prob = β)
REJECT H0	Type I Error -rejecting H0 when it is true (prob = α)	Correct Decision (prob = 1 – β)

Error in Hypothesis Testing

- Level of significance (α) is the probability of rejecting the null hypothesis when it is true
- α = P(Type I error) = P(Reject H_0 | H_0 is true)
- α = 0.05 indicates willing to accept a 5% chance that you are wrong when you reject the null hypothesis. To lower this risk, you must use a lower value for α. However, using a lower value alpha means it is less likely to detect a true difference if one really exists..
- Power of the test (1-β) is the probability of rejecting the null hypothesis when it is false
- Probability of making a type II error is β, which depends on the power of the test.

Type I and Type II Errors



- H0: You are normal
- H1: You are not normal (pregnant)
- Type I
  - ความผิดพลาดในการ reject มากเกินไป
  - ทั้ง ๆ ที่ความจริง ปกติ
- Type II
  - ความผิดพลาดในการ reject น้อยเกินไป
  - ทั้ง ๆ ที่ความจริงไม่ปกติ

Confusion Matrix: Performance of Classifier

		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** (Recall) $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

- *A* and *B* denote two events.
- Events could be that *it will rain tomorrow*; *a person has cancer*.
- *P(A |B)* = probability that *A* occurs given *B* is true
- *P(B|A)* = probability of observing *B* given *A* occurs
- *P(A)* and *P(B)* are probability that *A* and *B* occur, respectively
- Bayes Theorem provides a principled way for calculating conditional probabilities, called a *posterior probability*.

Bayes Theorem

$$P(A\,|B) = \frac{P(B|A)P(A)}{P(B)}$$

Example 1

คนไข้คนหนึ่งไปตรวจหามะเร็ง ผลการตรวจเป็นบวก อยากทราบว่า เราควรวินิจฉัยโรคคนไข้คนนี้ว่าเป็นมะเร็งจริงหรือไม่? ความเป็นจริง คือ

- ผลการตรวจเมื่อเป็นบวกจะให้ความถูกต้อง 98% กรณีที่มีโรคนั้นอยู่จริง
- ผลการตรวจเมื่อเป็นลบจะให้ความถูกต้อง 97% กรณีที่ไม่มีโรคนั้น
- 0.008 ของประชากรทั้งหมดเป็นโรคมะเร็ง

จากความเป็นจริงที่กำหนดให้ข้างต้น เราจะทราบค่าความน่าจะเป็นต่อไปนี้

P(cancer) = ... P(~cancer) = ...

P(+|cancer) = ... P(-|cancer) = ...

P(+|~cancer) = ... P(-|~cancer) = ...

P(cancer | +)
=P(+ | cancer)P(cancer)

=……………..………………

P(~cancer | +)
=P(+ | ~cancer)P(~cancer)

=……………………………

เนื่องจากผลรวมของ P(cancer|+) กับ P(~cancer|+)

เท่ากับ…………………จึงสามารถ normalize ค่าของ

P(cancer|+)=……………………………………………………
และ
P(~cancer|+)=………………………………………………

เนื่องจาก P(~cancer|+) มีค่ามากกว่า
สมมติฐานที่ว่า คนไข้ไม่เป็นมะเร็ง เมื่อทราบผลตรวจ
เป็นบวก จึงถูกเลือก

Example 2 cancer screening test scenario

- It reports **80** out of **100** cancer patients are correctly diagnosed, while the other **20** are not; cancer is falsely detected in **900** out of **9,900** healthy people.
- Given a positive screening result, the chance that the subject has cancer is ……….………, compared to ………………… where without undergoing the screening.

$$P(C|Pos) = \frac{P(Pos|C)P(C)}{P(Pos)}$$

Example3:

- Three machines *A*, *B*, *C* in a factory account for 35%, 20%, 45% of bulb production.
- The fraction of defective bulbs produced by each machine is 1.5%, 1%, and 2% respectively.
- A bulb produced by this factory was identified defective, denoted as **event *D***.
- This bulb was most likely manufactured by which machine?

Bayes Theorem for Modeling Hypotheses

- Bayes Theorem is a useful tool in applied machine learning. It provides a probabilistic model to describe the relationship between data (D) and a hypothesis (h);

$$P(h|D) = P(D|h) * P(h) / P(D)$$

- **Bayes**: maps the probabilities of observing input features given belonging classes, to the probability distribution over classes based on **Bayes theorem**.

How Naïve Bayes works

- Given a data sample *x* with *n* features, ***x*** = <*x1*, *x2*, …, *xn*>
- Goal of NB is to determine the probabilities that this sample belongs to each of K possible classes *y1*, *y2*, …, *yK*, that is *P(yk |**x**)*
- Consider ***x***, or *x1*, *x2*, …, *xn*, is a **joint event** that the sample has features with values *x1*, *x2*, …, *xn*, respectively, *yk* is an event that the sample belongs to class *k*.
- We can apply Bayes' theorem as:

$$P(y_k|x) = \frac{P(x|y_k)P(y_k)}{P(x)}$$

Mechanics of Naïve Bayes

- *P* (*yk*) portrays how classes are distributed, provided with no further knowledge of observation features. Thus, it is also called **prior** in Bayesian probability terminology. Prior can be either predetermined (usually in a *uniform* manner where each class has an equal chance of occurrence) or *learned from a set of training samples*.
- *P*(*yk* |***x***), in contrast to prior *P* (*yk*), is the **posterior** with extra knowledge of observation.
- *P*(***x*** | *yk*), or *P*(*x1*, *x2*, …, *xn*|*yk*) is the **joint distribution** of *n* features, given the sample belongs to class *yk*. This is how likely the features with such values cooccur. Obviously, the likelihood will be difficult to compute as the number of features increases.
- In NB, this is solved thanks to the feature independence assumption. The joint conditional distribution of n features can be expressed as the joint product of individual feature conditional distributions:

$$P(x|y_k) = P(x_1|y_k) * P(x_2|y_k) * \cdots * P(x_n|y_k)$$

Each conditional distribution can be efficiently learned from a set of training samples.

$$P(y_k|x) \propto P(x|y_k)P(y_k) = P(x_1|y_k) * P(x_2|y_k) * \cdots * P(x_n|y_k) * P(y_k)$$

Magazine Promotion	Watch Promotion	Life Insurance Promotion	Credit Card Insurance	Age	Sex
Yes	No	No	No	45	Male
Yes	Yes	Yes	Yes → NO	40	Female
No	No	No	No	42	Male
Yes	Yes	Yes	Yes	30	Male
Yes	No	Yes	No	38	Female
No	No	No	No	55	Female
Yes	Yes	Yes	Yes	35	Male
No	No	No	No	27	Male
Yes	No	No	No	43	Male
Yes	Yes	Yes	No	41	Female

Naïve Bayes Learning

Naïve Bayes Algorithm

Naïve_Bayes_Learn (examples)

For each target value C_i

$P'(C_i) \leftarrow$ estimate $P(C_i)$

For each attribute value $A_j = a_j$ of each target value C_i

$P'(A_j = a_j \mid C_i) \leftarrow$ estimate $P(A_j = a_j \mid C_i)$

Classify_New_Example (x)

$$C = \mathbf{Max}_{i=1}^{m} P'(C_i) \prod_{j=1}^{n} P'(A_j = a_j \mid C_i)$$

sampleID	hair color	eye color	weight	apply lotion	sun burn
S1	black	dark	overweight	no	-
S2	red	dark	normal	no	+
S3	blonde	light	overweight	no	+
S4	red	light	underweight	no	+
S5	black	dark	overweight	yes	-
S6	blonde	dark	overweight	no	+
S7	red	light	underweight	yes	-
S8	black	dark	normal	no	-
S9	blonde	dark	normal	yes	+
S10	red	light	normal	yes	+
S11	black	light	normal	yes	+
S12	blonde	light	underweight	no	+
S13	red	dark	normal	yes	-
S14	black	light	underweight	no	+

Learn_Naive_Bayes_Text (*Examples, V*)

1. collect all words and other tokens that occur in *Examples*
   - *Vocabulary* ← all distinct words and other tokens in *Examples*
2. calculate the required $P(v_j)$ and $P(w_k | v_j)$ probability terms
   - For each target value v_j in V do
     - $docs_j$ ← subset of *Examples* for which the target value is v_j
     - $P(v_j)$ ← | $docs_j$ | / | *Examples* |
     - $Text_j$ ← a single document created by concatenating all members of $docs_j$
     - n ← total number of words in $Text_j$ (counting duplicate words each)
     - for each word w_k in *Vocabulary*
       - n_k ← number of times word w_k occurs in $Text_j$
       - $P(w_k | v_j)$ ← $\dfrac{n_k + 1}{n + |Vocabulary|}$

Classify_Naive_Bayes_Text(Doc)

- *positions* ← all word positions in *Doc* that contain tokens found in *Vocabulary*
- Return v_{NB} where

$$v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \prod_{i \in positions} P(a_i | v_j)$$

m-estimate

Example classifying with Naïve Bayes

	ID	Terms in email	Is spam
Training data	1	click win prize	yes
	2	click meeting setup meeting	no
	3	prize free prize	yes
	4	click prize free	yes
Testing case	5	free setup meeting free	?

Numeric Data

Numeric data can be dealt with in a similar manner provided that the probability density function representing the distribution of the data is known. If a particular numerical attribute is normally distributed, we use the standard probability density function shown in Equation 10.13.

$$f(x) = 1/(\sigma\sqrt{2\pi})e^{-(x-\mu)^2/(2\sigma^2)} \quad (10.13)$$

where
- e = the *exponential function*
- μ = the class *mean* for the given numerical attribute
- σ = the class *standard deviation* for the attribute
- x = the attribute value

Although this equation looks quite complicated, it is very easy to apply. To demonstrate, consider the data in Table 10.6. This table displays the data in Table 10.4 with an added column containing numerical attribute *age*.

Let's use this new information to compute the conditional probabilities for the *male* and *female* classes for the following instance.

Magazine Promotion = *Yes*
Watch Promotion = *Yes*
Life Insurance Promotion = *No*
Credit Card Insurance = *No*
Age = *45*
Sex = *?*

- **Addition of Attribute *Age* to the Bayes Classifier Dataset**

Magazine Promotion	Watch Promotion	Life Insurance Promotion	Credit Card Insurance	Age	Sex
Yes	No	No	No	45	Male
Yes	Yes	Yes	Yes	40	Female
No	No	No	No	42	Male
Yes	Yes	Yes	Yes	30	Male
Yes	No	Yes	No	38	Female
No	No	No	No	55	Female
Yes	Yes	Yes	Yes	35	Male
No	No	No	No	27	Male
Yes	No	No	No	43	Male
Yes	Yes	Yes	No	41	Female

Male
$\mu = \frac{222}{6} = 37$

Femal
$\mu \ \frac{174}{4} = 43.5$

For the overall conditional probabilities we have:

$P(E \mid sex = male) = (4/6)\ (2/6)\ (4/6)\ (4/6)\ [P(age = 45 \mid sex = male)]$
$P(E \mid sex = female) = (3/4)\ (2/4)\ (1/4)\ (3/4)\ [P(age = 45 \mid sex = female)]$

To determine the conditional probability for *age* given *sex* = *male*, we assume *age* to be normally distributed and apply the probability density function. We use the data in Table 10.5 to find the mean and standard deviation scores. For the class *sex* = *male* we have: $\sigma = 7.69$, $\mu = 37.00$, and $x = 45$. Therefore the probability that *age* = *45* given *sex* = *male* is computed as:

$$P(age = 45 \mid sex = male) = 1/(\sqrt{2\pi}\,7.69)e^{-(45-37.00)^2/[2(7.69)^2]}$$

Making the computation, we have:

$P(age = 45 \mid sex = male) \approx 0.030$

To determine the conditional probability for *age* given *sex* = *female*, we substitute $\sigma = 7.77$, $\mu = 43.50$, and $x = 45$. Specifically,

$$P(age = 45 \mid sex = female) = 1/(\sqrt{2\pi}\,7.77)e^{-(45-43.50)^2/[2(7.77)^2]}$$

Making the computation, we have:

$P(age=45 \mid sex = female) \approx 0.050$

We can now determine the overall conditional probability values:

$P(E \mid sex = male) = (4/6)\ (2/6)\ (4/6)\ (4/6)\ (0.030) \approx .003$
$P(E \mid sex = female) = (3/4)\ (2/4)(1/4)\ (3/4)\ (0.050) \approx .004$

Finally, applying Equation 10.9 we have:

$P(sex = male \mid E) \approx (.003)\ (0.60)\ /\ P(E) \approx .0018\ /\ P(E)$
$P(sex = female \mid E) \approx (.004)\ (0.40)\ /\ P(E) \approx .0016\ /\ P(E)$

Once again, we ignore $P(E)$ and conclude that the instance belongs to the *male* class.