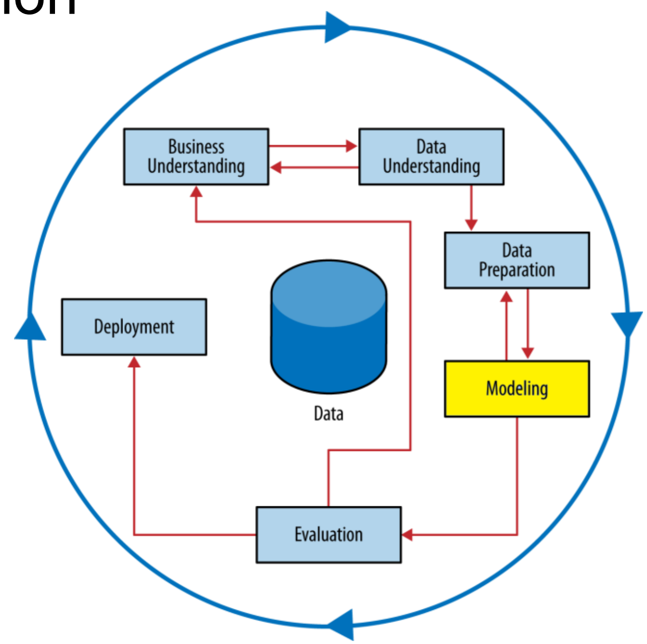
Evidence and Probabilities

PF9

Learning Goals

- Explicit evidence combination with **Bayes' Rule**
  - Take features as evidences, while the classification result as conditional result for the evidence(s).

- Probabilistic reasoning via the assumption of **conditional independence** among multiple evidences.
  - Naive Bayes



- Techniques:
  - Naive Bayes Classification
  - Evidence Lift

Example: An Upscale Hotel Chain

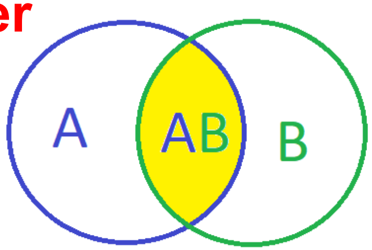
- **Target**: whether a consumer will book a room in our hotel?
- **Features**: whether the customer (1) visited our website; (2) did online shopping; (3) bought a travel book on Amazon or not.

- **Probabilistic reasoning**: combine multiple evidences probabilistically.
  - A feature as an evidence, all evidences as a feature vector $E = [e_1, \ldots, e_k]$.
    - The customer has visited our website.
    - The customer has been doing online shopping.
    - The customer has bought a travel book from Amazon.
  - What is the probability that a customer will book a room (c_i) given the feature values (E) about him/her? $P(c_i|E)$

- **Evidence Lift**: how would an evidence ($\boldsymbol{e_i}$) increase the estimated probability of booking $\boldsymbol{P(c_i|E)}$, in comparison to the baseline?
  - The baseline (prior) probability considers no evidence: $p(\boldsymbol{c_i})$

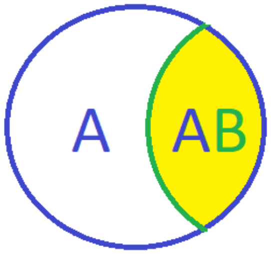
Conditional and Joint Probability

- **The area AB**: **joint probability A & B occur together**



$$p(A \cap B) \quad \text{or} \quad p(AB)$$

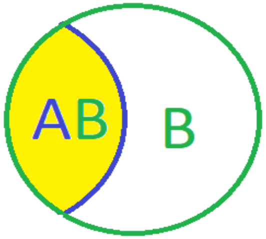
- **Ratio of area AB to A**: **conditional probability of B given A**



$$\boldsymbol{p(B|A) = \frac{p(AB)}{p(A)}}$$

  - The joint probability $\mathbf{p(AB) = p(B|A) \times p(A)}$

- **Ratio of area AB to B**: **conditional probability of A given B**

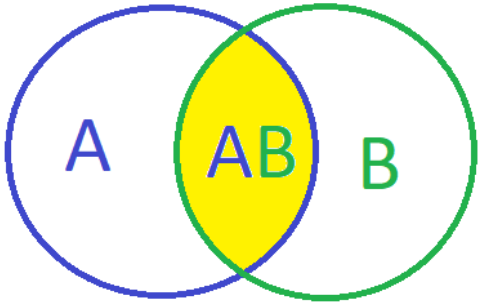


$$\boldsymbol{p(A|B) = \frac{p(AB)}{p(B)}}$$

  - The joint probability $\mathbf{p(AB) = p(A|B) \times p(B)}$

$$p(AB) = p(B|A) \times p(A) = p(A|B) \times p(B)$$

Example: Conditional and Joint Probability



- Assume there are 5 days per week:
  - Student A comes to school 4 days every week;
  - Student B comes to school 3 days every week;
  - Student A and B come to school together 2 days every week.

- Please answer below questions:
  - What is the probability that A come today? $^4/_5$
  - What is the probability that B come today? $^3/_5$

    Unconditional probability (Prior probability)
  - What is the probability that A and B come together? $^2/_5$ **Joint probability**
  - What is the probability B would come if A is here? $^2/_4$
  - What is the probability A would come if B is here? $^2/_3$

    Conditional probability

Exercise: Gender vs. Hair Color

- Unconditional probability:
  - $P(Red) = \frac{16}{30}$
  - $P(Yellow) = \frac{14}{30}$
  - $P(Male) = \frac{18}{30}$
  - $P(Female) = \frac{12}{30}$

	Yellow Hair	Red Hair	Total
Male	8	10	18
Female	6	6	12
Total	14	16	30

- Compute the probabilities below:
  - $P\ (Male \cap Yellow)$: joint probability to see a male person with yellow hair
  - $P\ (Yellow \mid Male)$: conditional probability of yellow hair given he is a male
  - $P\ (Male \mid Yellow)$: conditional probability of being male given s/he has yellow hair

Independence

- Independence is a special case of conditional probability.
  - If A and B are independent, knowing A tells nothing about B.

- If student A and B are independent:
  - Conditional probability to see A given B is

    $p\ (A \mid B) = p(A)$ = 80%

  - Conditional probability to see B given A is

    $p(B \mid A) = p(B)$ = 60%

  - Joint probability to see A and B together should be

    $p(AB)$ = $p(A) \times p(B)$ = 80% x 60% = 48%

  - Joint probability can also be rewritten as:
    - $p(AB)$ = $p(A) \times p(B) = p(A|B) \times p(B)$
    - $p(AB)$ = $p(B) \times p(A) = p(B|A) \times p(A)$

Student A comes to school 4 days every week

Student B comes to school 3 days every week

Higher than the joint probability (40%) observed

$$p(AB) = p(B|A) \times p(A) = p(A|B) \times p(B)$$

Bayes' Rule for Classification

- **Bayes' Rule**:

$$p(AB) = p(B|A) \times p(A) = p(A|B) \times \ p(B)$$

- **Conditional Probability of B on A**:

$$p(B|A) = \frac{p(A|B) \times p(B)}{p(A)}$$

- Replace A with $\boldsymbol{E}$ (i.e., evidence), B with $\boldsymbol{c_i}$ (i.e., a classification result):
  - Usually evidences E are easily observed, while the classification result $\boldsymbol{c_i}$ is something we are interested but NOT easily observed.

$$\boldsymbol{p(c_i\,|E) = \frac{p(E|c_i\,)\ \times\ p(c_i\,)}{p(E)}}$$

Estimate Class Probability With Bayes' Rule

- Unconditional probability:
  - $P(Red) = \frac{16}{30}$
  - $P(Yellow) = \frac{14}{30}$
  - $P(Male) = \frac{18}{30}$
  - $P(Female) = \frac{12}{30}$

	Yellow Hair	Red Hair	Total
Male	8	10	18
Female	6	6	12
Total	14	16	30

- The probability someone is male (c_i) given s/he has yellow hair (***E***) :

$$P(Male|Yellow) = \frac{P(Yellow|Male) \times P(Male)}{P(Yellow)}$$

$$= \frac{\frac{8}{18} \times \frac{18}{30}}{\frac{14}{30}} = \frac{8}{14}$$

Compute $p(E)$ can be difficult.

Example 1 – Compute $p(E)$

- There are 3 coins with the same look and weight.
  - 2 coins are fair: P(Head | F) = 0.5
  - 1 coin is magic: P(Head | M) = 0.75

- One coin is picked randomly. How likely this is a magic coin?
  - P(M): the unconditional/prior probability – the baseline

P(M) = 1/3

- Now, we throw it once: we see a head. What is the probability that this is a magic coin?
  - Evidence: Head
  - Outcome: Magic

P(M | Head)

Example 1 – Compute $p(E)$

- Unconditional probability:
  - $P(M) = 1/3\,;\ P(F) = 2/3$
- Conditional probability: likelihood to see a Head given a magic/fair coin:
  - $P(Head|M) = 3/4\,;\ P(Head|F) = 1/2$

- According to Bayes' rule, the probability of being a magic coin given a head is observed:

$$\boldsymbol{P(M|Head) = \frac{P(Head|M) \times P(M)}{P(Head)}}$$

  - $P(Head) = P(M)\ \times\ P(Head|M) + P(F)\ \times\ P(Head|F)$

$$= {}^{1}/_{3} \times {}^{3}/_{4} + {}^{2}/_{3} \times {}^{1}/_{2} = {}^{7}/_{12}$$

  - $\boldsymbol{P(\mathrm{M}|\mathrm{Head}) = \frac{\mathrm{P(Head|M) \times P(M)}}{\mathrm{P(Head)}}}$

$$= ({}^{3}/_{4} \times {}^{1}/_{3})/({}^{7}/_{12}) \quad = \quad {}^{3}/_{7}$$

More likely this is a magic coin if a head is observed (3/7 > 1/3)

Example 2 – Compute $p(E)$

- There are 3 coins with the same look and weight.
  - 2 coins are fair: P(Head | F) = 0.5
  - 1 coin is magic: P(Head | M) = 0.75

- One coin is picked randomly, and we throw it *twice*. The outcome is {Head,Head}. What is the probability that this is a magic coin?
  - Evidence: {Head, Head}
  - Outcome: Magic

P(M | HH)

Recap: Bayes' Rule for Classification

$$p(c_i|E) = \frac{p(E|c_i) \times p(c_i)}{p(E)}$$

- c_i is a class label (i.e., one target value).
- $E = [e_1, e_2, \dots, e_k]$ is a k-dimensional feature vector, with e_i is a feature value.

- $p(c_i|E)$: the (posterior) **conditional probability** of class c_i.
  - The probability an instance belongs to class c_i given evidence E observed.
- $p(c_i)$: the (prior) **unconditional probability** of class c_i (i.e., baseline).
  - The probability an instance is classified as c_i without any evidence.
  - Usually inferred from data: i.e., proportion of class c_i in training data.
- $p(E|c_i)$: the **conditional probability** to see evidence E in class c_i.
  - i.e., proportion of instances in class c_i with evidence E.

How can we handle multiple features/evidences?

Naive Bayes: Conditional Independence

- According to the formula of joint probability:

$$p(AB) = p(B|A) \times p(A)$$

- If there are three events:

$$p(ABC) = p(B|AC) \times p(A|C) \times p(C)$$

$$p(AB|C) = \frac{p(ABC)}{p(C)} = p(B|AC) \times p(A|C)$$

- If we assume A and B are **conditionally independent** given C, then

$$p(AB \mid C) = p(A \mid C) \times p(B \mid C)$$

- **Naive Bayes** assumes conditional independence among features given class $\boldsymbol{c_i}$.
  - $\boldsymbol{p(E|c_i) = p(e_1, e_2, \ldots, e_k|c_i) = p(e_1|c_i) \times p(e_2|c_i) \times \cdots \times p(e_k|c_i)}$
  - $\boldsymbol{p(e_i|c_i)}$ can be estimated directly from the data.

Naive Bayes for Classification

- **The Naive Bayes Equation:**

$$\boldsymbol{p(c_i|\,E) = \frac{p(E|c_i)\times p(c_i)}{p(E)}}$$

- $p(E|c_i) = p(e_1|c_i) \times p(e_2|c_i) \times \cdots \times p(e_k|c_i)$

- $p(E) = p(E \cap c_0) + p(E \cap c_1)$

  Assume binary classification

  $= p(E \mid c_0) \times p(c_0) + p(E \mid c_1) \times p(c_1)$

  $= p(e_1|c_0) \times \cdots \times p(e_k|c_0) \times p(c_0) + p(e_1|c_1) \times \cdots \times p(e_k|c_1) \times p(c_1)$

- Note 1: if we aim for classification, no need to compute $p(E)$ as it is the same for all classes.
- Note 2: with multiple features, $p(E)$ based on naive bayes rule is NOT equal to the proportion of instances with feature E in the data. Same logic applies to $p(E|c_i)$ and $p(c_i|\,E)$.

Naive Bayes Classifier

- 3 features:
  - Height = s, m, t
  - Weight = h, l, n
  - Long_hair = y, n
- Target variable:
  - Gender = F, M
- For a person who is tall, normal weight, long hair, what is the predicted gender?

ID	Height	Weight	Long_hair	Gender
1	m	n	n	M
2	s	l	y	F
3	t	h	n	M
4	s	n	y	F
5	t	n	y	F
6	s	l	n	F
7	s	h	y	M
8	m	n	n	F
9	m	l	y	F
10	t	n	n	M

$$\boldsymbol{p\,(G = F|\, H = t, W = n, L = y)}$$

Is it 100%?

Naive Bayes Classifier

- Compute the likelihood of being Male given the person is tall, normal weight and long hair.
  - The probability of $G = M$ given the three feature values is:

$$P(G = M | H = t, W = n, L = y)$$

Bayes' Rule

$$= \frac{P(H = t, W = n, L = y | G = M) \times P(G = M)}{P(H = t, W = n, L = y)}$$

Conditional Independence assumption

$$= \frac{[P(H = t | G = M) \times P(W = n | G = M) \times P(L = y | G = M)] \times P(G = M)}{P(H = t, W = n, L = y)}$$

Naive Bayes Classifier

- Compute the likelihood of being Female given the person is tall, normal weight and long hair.
  - The probability of $G = F$ given the three feature values is:

$$P(G = F | H = t, W = n, L = y)$$

Bayes' Rule

$$= \frac{P(H = t, W = n, L = y | G = F) \times P(G = F)}{P(H = t, W = n, L = y)}$$

Conditional Independence assumption

$$= \frac{[P(H = t | G = F) \times P(W = n | G = F) \times P(L = y | G = F)] \times P(G = F)}{P(H = t, W = n, L = y)}$$

Naive Bayes Classifier

- Conditional probabilities of being tall:
  - $P(H = t|G = M) = \frac{2}{4}$
  - $P(H = t|G = F) = \frac{1}{6}$
- Conditional probabilities of having normal weight:
  - $P(W = n|G = M) = \frac{2}{4}$
  - $P(W = n|G = F) = \frac{3}{6}$
- Conditional probabilities of having long hair:
  - $P(L = y|G = M) = \frac{1}{4}$
  - $P(L = y|G = F) = \frac{4}{6}$

ID	Height	Weight	Long_hair	Gender
1	m	n	n	M
2	s	l	y	F
3	t	h	n	M
4	s	n	y	F
5	t	n	y	F
6	s	l	n	F
7	s	h	y	M
8	m	n	n	F
9	m	l	y	F
10	t	n	n	M

Naive Bayes Classifier

- The probability of $G = M$ given the three features is:

$$P(G = M|H = t, W = n, L = y)$$

$$= \frac{[P(H = t|G = M) \times P(W = n|G = M) \times P(L = y|G = M)] \times P(G = M)}{P(H = t, W = n, L = y)}$$

$$= \frac{\frac{2}{4} \times \frac{2}{4} \times \frac{1}{4} \times \frac{4}{10}}{\boldsymbol{P(H = t, W = n, L = y)}}$$

$$= \frac{\frac{1}{40}}{\boldsymbol{P(H = t, W = n, L = y)}}$$

- As $p(\mathrm{E})$ is the same for both class, no need to compute it if classification is the goal.

Naive Bayes Classifier

- The probability of $G = F$ given the three features is:

$$P(G = F|H = t, W = n, L = y)$$

$$= \frac{[P(H = t|G = F) \times P(W = n|G = F) \times P(L = y|G = F)] \times P(G = F)}{P(H = t, W = n, L = y)}$$

$$= \frac{\frac{1}{6} \times \frac{3}{6} \times \frac{4}{6} \times \frac{6}{10}}{\boldsymbol{P(H = t, W = n, L = y)}}$$

$$= \frac{\frac{1}{30}}{\boldsymbol{P(H = t, W = n, L = y)}}$$

$1/30 > 1/40$

- **Conclusion**: the gender of this person (tall, normal weight, long hair) is predicted as as Female.
  - $p(\mathrm{E})$ is the same for both class, no need to compute it.

Estimate Class Probability

- Let's look at the denominator $p(\mathrm{E})$:

$$\boldsymbol{P(H = t, W = n, L = y)}$$

$$= P(H = t, W = n, L = y|\boldsymbol{G} = \boldsymbol{M}) \times \boldsymbol{P}(\boldsymbol{G} = \boldsymbol{M}) + P(H = t, W = n, L = y|\boldsymbol{G} = \boldsymbol{F}) \times \boldsymbol{P}(\boldsymbol{G} = \boldsymbol{F})$$

Conditional Independence assumption

$$= [P(H = t|G = M) \times P(W = n|G = M) \times P(L = y|G = M)] \times P(G = M) + [P(H = t|G = F) \times P(W = n|G = F) \times P(L = y|G = F)] \times P(G = F)$$

$$= \left(\frac{2}{4} \times \frac{2}{4} \times \frac{1}{4}\right) \times \frac{4}{10} + \left(\frac{1}{6} \times \frac{3}{6} \times \frac{4}{6}\right) \times \frac{6}{10} = \frac{7}{120}$$

- Then
  - $P(G = M|H = t, W = n, L = y) = \dfrac{\frac{1}{40}}{\boldsymbol{P(H=t,W=n,L=y)}} = \dfrac{\frac{1}{40}}{\frac{7}{120}} = \dfrac{3}{7}$
  - $P(G = F|H = t, W = n, L = y) = \dfrac{\frac{1}{30}}{\boldsymbol{P(H=t,W=n,L=y)}} = \dfrac{\frac{1}{30}}{\frac{7}{120}} = \dfrac{4}{7}$

Naive Bayes Classifier

- Advantages
  - A very simple classifier: store the counts of classes and feature as each instance is seen.
    - $\boldsymbol{p(c_i)}$: proportion of instances of class c_i in the dataset.
    - $\boldsymbol{p(e_i|c_i)}$: proportion of instances in class c_i with feature value e_i.
  - **Incremental learning** allows the model to learn from new data without forgetting previously learnt information.
    - Useful when data is continuously generated (e.g., streaming platforms) or too large to be processed all at once.

- Disadvantages
  - Class probability estimation may not be accurate, ranking is fine.
    - Probability is overestimated for correct class but underestimated for incorrect class, as the *assumption of conditional independence* is usually unrealistic.

Gaussian Naive Bayes

- Gaussian NB works for data with **continuous features**.
  - e.g., petal width, petal length for the iris data

- Estimate the prior probability for each class $\boldsymbol{c_i}$.
  - i.e., proportion of instances of class $\boldsymbol{c_i}$ in the training set.

$p(\boldsymbol{c_i})$

- Compute each feature’s mean and standard deviation in each class $\boldsymbol{c_i}$, assuming normal distribution.
- When making prediction for a new instance X:
  - compute the conditional probability for X takes on feature value $\boldsymbol{e_i}$ in class $\boldsymbol{c_i}$, using probability density function of continuous random variables.

$p(e_i|\boldsymbol{c_i})$

  - Estimate class probability for X

$$\frac{(\boldsymbol{e_1}|c_i) \times \cdots \times \boldsymbol{p}(\boldsymbol{e_k}|c_i) \times \boldsymbol{p}(\boldsymbol{c_i})}{\boldsymbol{p}(\boldsymbol{E})}$$

Evidence Lift

- Assuming **unconditional independence** for all features:
  - $p(\mathrm{E}) = p(e_1) \times\ p(e_2) \times \cdots \times p(e_k)$
  - $p(c_i|E) = \frac{[p(e_1|c_i)\times p(e_2|c_i)\times\cdots\times p(e_k|c_i)]\times p(c_i)}{p(\mathrm{E})} = \frac{[p(e_1|c_i)\times p(e_2|c_i)\times\cdots\times p(e_k|c_i)]\times p(c_i)}{p(e_1)\times p(e_2)\times\cdots\times p(e_k)}$

Naive-Naive Bayes

- **Evidence lift** for a feature value (e_i) is:

$$lift_{c_i}(e_i) = \frac{p(e_i|c_i)}{p(e_i)}, \qquad i = 1, \dots, k$$

$$p(c_i|E) = p(c_i) \times lift_{c_i}(e_1)\ \times \cdots \times\ lift_{c_i}(e_k)$$

- Each feature value raises/lowers the prior probability $p(c_i)$ by its lift score.
  - If $lift_{c_i}(e_i)$ > 1, then $p(c_i|E)\ >\ p(c_i)$.
  - If $lift_{c_i}(e_i)$ < 1, then $p(c_i|E)\ <\ p(c_i)$.

Evidence Lift

#	X	target
1	M	c_1
2	M	c_1
3	M	c_1
4	F	c_1
5	M	c_0
6	F	c_0
7	F	c_0
8	F	c_0
9	M	c_1
10	M	c_0

	c_1	c_0	
M	4	2	6
F	1	3	4
	5	5	

$p(c_1) = \frac{1}{2}$

The prior probability

$$lift_{c_1}(X=F) = \frac{p(X=F|c_1)}{p(X=F)} = \frac{\frac{1}{5}}{\frac{4}{10}} = \frac{1}{2}$$

$$P(c_1|X=F) = p(c_1) \times lift_{c_1}(X=F) = \frac{1}{2} * \frac{1}{2} = \frac{1}{4}$$

Rewritten as $lift_{c_1}(X=F) = \frac{P(c_1|X=F)}{p(c_1)}$

$lift_{c_i}(e_i)$ tells how much more prevalent class c_i is in a subpopulation (i.e., $p(c_i|e_i)$), compared to its overall prevalence in general population (i.e., $p(c_i)$).

Evidence Lift of Facebook "Likes"

- Kosinski et al. (2013) find that what people "like" on FB is predictive of personal traits (usually not apparent):
  - How they score on psychometric tests (extrovert or conscientious)?
  - How they score on intelligence tests?
  - Whether they are (openly) gay?

- What are the *Likes* that have strong evidence lifts for "high IQ" (IQ>130)?
  - If a person likes "Lord of the Rings", the probability he has high-IQ is 69% higher than the baseline (i.e., proportion of high-IQ in general population).

Like	Lift	Like	Lift
Lord Of The Rings	1.69	Wikileaks	1.59
One Manga	1.57	Beethoven	1.52
Science	1.49	NPR	1.48
Psychology	1.46	*Spirited Away*	1.45
The Big Bang Theory	1.43	Running	1.41
Paulo Coelho	1.41	Roger Federer	1.40
The Daily Show	1.40	*Star Trek*	1.39
Lost	1.39	Philosophy	1.38
Lie to Me	1.37	*The Onion*	1.37
How I Met Your Mother	1.35	*The Colbert Report*	1.35
Doctor Who	1.34	*Star Trek*	1.32
Howl's Moving Castle	1.31	Sheldon Cooper	1.30
Tron	1.28	*Fight Club*	1.26
Angry Birds	1.25	*Inception*	1.25
The Godfather	1.23	*Weeds*	1.22