人工智能
不确定性知识表示与推理

陈 川
中山大学 计算机学院
2024年

中山大學
SUN YAT-SEN UNIVERSITY

不确定性知识表示与推理

- 概率统计、独立性、贝叶斯规则
- 贝叶斯推断
- 贝叶斯网络

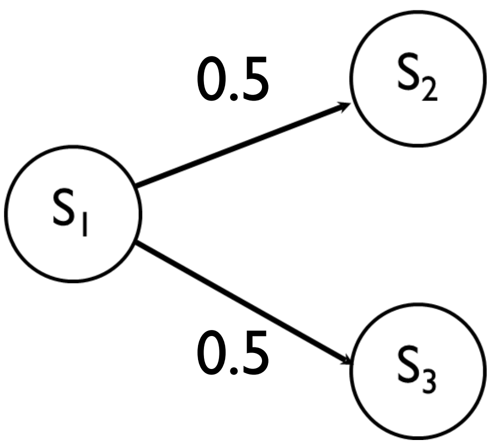
Uncertainty 不确定性

- In search, we viewed actions as being deterministic.
  - executing action A in state S_1 causes transition to state S_2
- Furthermore, there was a fixed initial state S_0.
- So after executing any sequence of actions, we know exactly what state we have arrived at.
- These assumptions are sensible in some domains, but in many domains they are not true.

The world is a very uncertain place

Uncertainty 不确定性

- We might not know exactly what state we start off in
  - *e.g.*, we can't see our opponents' cards in a poker game
  - We don't know what a patient's ailment is.
- We might not know all of the effects of an action
  - The action might have a random component, like rolling dice.
  - We might not know all of the long term effects of a drug.
  - An action might fail



Based on what we can see, there's a 30% chance we're in cell S_1, 30% in S_s and 40% in S_3....

Move(S_1,'N') = S_2 50% of the time

Move(S_1,'N') = S_3 50% of the time

Uncertainty 不确定性

- In such domains we still need to act,
- but we can't act solely on the basis of known true facts.
- We have to "gamble".
- But how do we **gamble rationally**?

An example

设想下：We have to go to the airport. But we don't know for certain what the traffic will be like on the way to the airport. When do we leave?

- If we must arrive at the airport at 9 pm on a week night
  - we could "safely" leave for the airport 1 hour before.
  - Some probability of the trip taking longer, but the probability is low.
- If we must arrive at the airport at 6:30pm on Friday
  - we most likely need 1.5 hour or more to get to the airport.

Acting rationally under uncertainty typically corresponds to maximizing one's Expected Utility 期望效用.

Expected Utility Example

- Probability distribution over outcomes (also called a "joint distribution")

Event	Go to Bloor St.	Go to Queen Street
Find Ice Cream	0.5	0.2
Find donuts	0.4	0.1
Find live music	0.1	0.7

- Utilities of outcomes

Event	Utility
Ice Cream	10
Donuts	5
Music	20

Expected Utility Example

- Maximum Expected Utility?

Event	Go to Bloor St.	Go to Queen Street
Ice Cream	0.5 * 10	0.2 *10
Donuts	0.4 * 5	0.1 * 5
Music	0.1 * 20	0.7 * 20
Utility	**9.0**	**16.5**

- Here, it's "Go to Queen Street"
- If the utility of Donuts of Ice Cream had been higher, however, it might have been "Go to Bloor Street".

Uncertainty

- To act rationally under uncertainty, we must be able to evaluate how likely certain things are.
  在面对不确定性时，要做出理性的行为，我们必须能够评估某些事情发生的可能性。
- By weighing likelihoods of events (probabilities), we can develop mechanisms for acting rationally under uncertainty.
  通过权衡事件的可能性（概率），我们可以开发出在不确定性下行动的机制。

前置概念 Probability (over Finite Sets)

A probability is a function defined over a set of atomic events U.

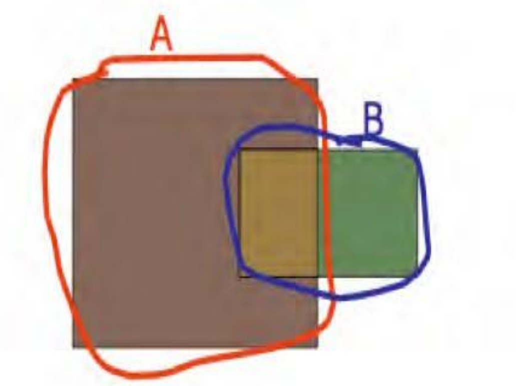
U represents the universe of all possible events.

- It assigns a value $Pr(e)$ to each event $e \in U$, in the range [0,1].

- It assigns a value to every set of events F by s probabilities of the members of that set:
- $Pr(F) = \sum_{e \in F} Pr(e)$

- Thus $Pr(U) = 1, Pr(\emptyset) = 0$

- $Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$



前置概念 Probability (over Finite Sets)

给定一个集合U（ universe ），概率函数是定义在U的子集上的函数，它将每个子集映射到实数，并且满足概率公理

- $Pr(U) = 1$
- $Pr(A) \in [0, 1]$
- $Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$

Properties and sets 属性与集合

Any set of events A can be interpreted as a property: the set of events with property A 任何一组事件A都可以解释为一个属性：具有属性A的事件集. Hence, we often write

- $A \vee B$ to represent the set of events with either property A or B: the set $A \cup B$
- $A \wedge B$ to represent the set of events with both property A and B: the set $A \cap B$
- $\neg A$ to represent the set of events that do not have property A: the set $U - A$ ($i.e.$, the complement of A wrt the universe of events U)

Probability over Feature Vectors 多变量概率

- As we move forward, we will model sets of events in our universe as vectors of feature values.
- We have
  - a set of variables $V_1, V_2, \ldots, V_n$
  - a finite domain of values for each variable, Dom[V_1], Dom[V_2],. . . , Dom[V_n].
- The universe of events U is the set of all vectors of values for the variables $\{(d_1, d_2, \ldots, d_n) \mid d_i \in Dom[V_i]\}$
- This event space has size $\prod_i |\boldsymbol{Dom}\ [\boldsymbol{V_i}]|$, *i.e.*, the product of the domain sizes.
- e.g., if |Dom[Vi]| = 2, we have 2^n distinct atomic events. (Exponential!)

Probability over Feature Vectors 多变量概率

- Asserting that some subset of variables have particular values allows us to specify a useful collection of subsets of U, *e.g.*
  - $\{V_1 = a\}$ = set of all events where $V_1 = a$
  - $\{V_1 = a, V_3 = d\}$ = set of all events where $V_1 = a$ and $V_3 = d$.
- If we had Pr of every atomic event (full instantiation of the variables) we could compute Pr of any other set, *e.g.*

$$Pr(\{V_1 = a\}) = \sum_{x_2 \in D[V_2]} \ldots \sum_{x_n \in D[V_n]} Pr(V_1 = a, V_2 = x_2, \ldots, V_n = x_n)$$

Review: Probability over Feature Vectors

Example:

$$P(\{V_1 = 1\}) = \sum_{x_2 \in \mathrm{Dom}[V_2]}, \sum_{x_3 \in \mathrm{Dom}[V_3]} P(\{V_1 = 1, V_2 = x_2, V_3 = x_3\}).$$

(V1 = 1, V2 = 1, V3 = 1)	(V1 = 2, V2 = 1, V3 = 1)	(V1 = 3, V2 = 1, V3 = 1)
(V1 = 1, V2 = 1, V3 = 2)	(V1 = 2, V2 = 1, V3 = 2)	(V1 = 3, V2 = 1, V3 = 2)
(V1 = 1, V2 = 1, V3 = 3)	(V1 = 2, V2 = 1, V3 = 3)	(V1 = 3, V2 = 1, V3 = 3)
(V1 = 1, V2 = 2, V3 = 1)	(V1 = 2, V2 = 2, V3 = 1)	(V1 = 3, V2 = 2, V3 = 1)
(V1 = 1, V2 = 2, V3 = 2)	(V1 = 2, V2 = 2, V3 = 2)	(V1 = 3, V2 = 2, V3 = 2)
(V1 = 1, V2 = 2, V3 = 3)	(V1 = 2, V2 = 2, V3 = 3)	(V1 = 3, V2 = 2, V3 = 3)
(V1 = 1, V2 = 3, V3 = 1)	(V1 = 2, V2 = 3, V3 = 1)	(V1 = 3, V2 = 3, V3 = 1)
(V1 = 1, V2 = 3, V3 = 2)	(V1 = 2, V2 = 3, V3 = 2)	(V1 = 3, V2 = 3, V3 = 2)
(V1 = 1, V2 = 3, V3 = 3)	(V1 = 2, V2 = 3, V3 = 3)	(V1 = 3, V2 = 3, V3 = 3)

Review: Probability over Feature Vectors

Example:

$$P(\{V_1 = 1, V_3 = 2\}) = \sum_{x_2 \in \mathrm{Dom}[V_2]} P(\{V_1 = 1, V_2 = x_2, V_3 = 2\}).$$

(V1 = 1, V2 = 1, V3 = 1)	(V1 = 2, V2 = 1, V3 = 1)	(V1 = 3, V2 = 1, V3 = 1)
(V1 = 1, V2 = 1, V3 = 2)	(V1 = 2, V2 = 1, V3 = 2)	(V1 = 3, V2 = 1, V3 = 2)
(V1 = 1, V2 = 1, V3 = 3)	(V1 = 2, V2 = 1, V3 = 3)	(V1 = 3, V2 = 1, V3 = 3)
(V1 = 1, V2 = 2, V3 = 1)	(V1 = 2, V2 = 2, V3 = 1)	(V1 = 3, V2 = 2, V3 = 1)
(V1 = 1, V2 = 2, V3 = 2)	(V1 = 2, V2 = 2, V3 = 2)	(V1 = 3, V2 = 2, V3 = 2)
(V1 = 1, V2 = 2, V3 = 3)	(V1 = 2, V2 = 2, V3 = 3)	(V1 = 3, V2 = 2, V3 = 3)
(V1 = 1, V2 = 3, V3 = 1)	(V1 = 2, V2 = 3, V3 = 1)	(V1 = 3, V2 = 3, V3 = 1)
(V1 = 1, V2 = 3, V3 = 2)	(V1 = 2, V2 = 3, V3 = 2)	(V1 = 3, V2 = 3, V3 = 2)
(V1 = 1, V2 = 3, V3 = 3)	(V1 = 2, V2 = 3, V3 = 3)	(V1 = 3, V2 = 3, V3 = 3)

In these examples we are **"summing out" 总结法则** some variables, which is also known as "marginalizing" our distribution

Problem and solution 存在的问题与解决思路

Problem

- This is an exponential number of atomic probabilities to specify. 需要指定指数级别的原子概率
- Requires summing up an exponential number of items.

  需要对指数级别的项求和。

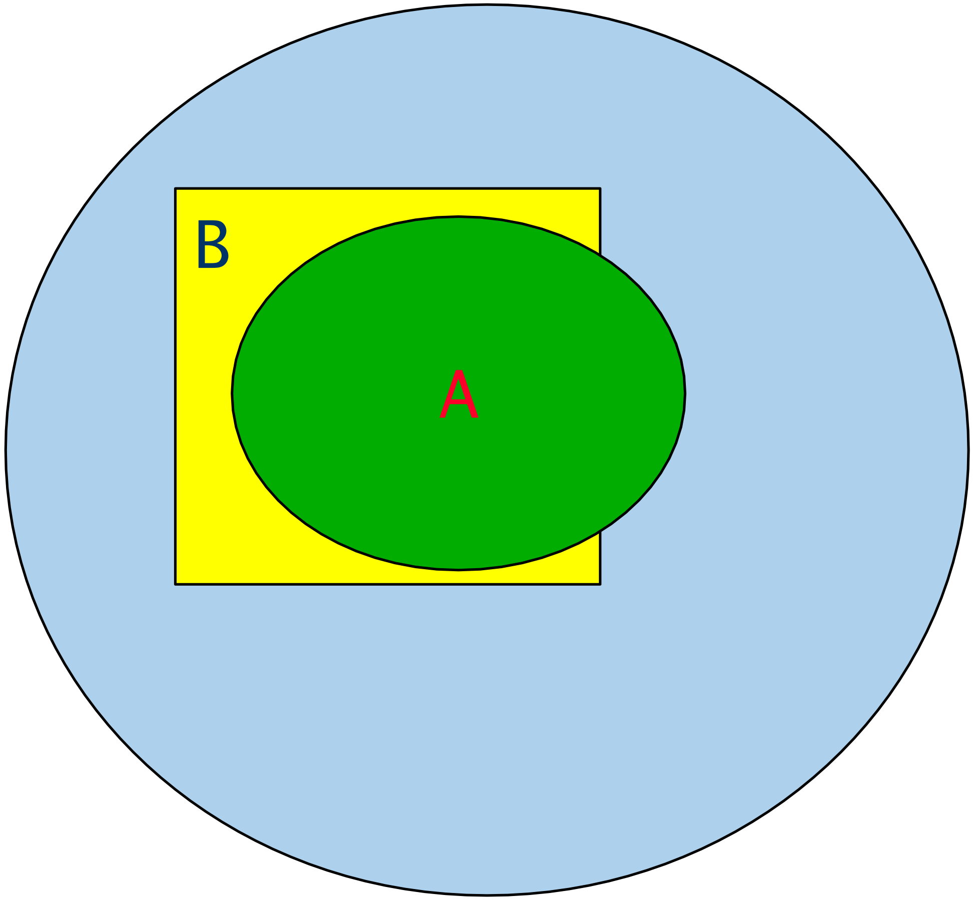
Solution

- Make use of probabilistic independence, especially conditional independence. 利用概率独立性，特别是条件独立性。

Conditional probabilities 条件概率

- Before we get to conditional independence, we need to define the meaning of conditional probabilities.
  - Say that A is a set of events such that $Pr(A) > 0$.
  - Then one can define a conditional probability wrt the event A: $Pr(B|A) = Pr(B \cap A)/Pr(A)$
  - Conditioning on A, corresponds to restricting one's attention to the events in A.

An example



B covers about 30% of the entire space (U), but covers over 80% of A.

So $Pr(B) = 0.3$, but $Pr(B|A) = 0.8$

These capture conditional information about the influence of any one variable's value on the probability of others'.

Summing out rule 总结法则

- Say that $B_1, B_2, \ldots, B_k$ form a partition of the universe U.
  - $B_i \cap B_j = \emptyset, \quad i \neq j$ (mutually exclusive 相互排斥)
  - $B_1 \cup B_2 \cup \ldots \cup B_k = U$ (exhaustive 周全)
- In probabilities:
  - $Pr(B_i \cap B_j) = 0, i \neq j$
  - $Pr(B_1 \cup B_2 \cup \ldots \cup B_k) = 1$
- Given any other set of events A, we have that
  $Pr(A) = Pr(A \cap B_1) + \ldots + Pr(A \cap B_k)$
- In conditional probabilities:
  $Pr(A) = Pr(A|B_1)Pr(B_1) + \ldots + Pr(A|B_k)Pr(B_k)$
- Often we know $Pr(A|B_i)$, so we can compute $Pr(A)$ this way.

Independence 独立性

- It could be that the density of B on A is identical to its density on the entire set.
  - **Probability density** is a measure of likelihood: pick an element at random from the entire set. How likely is it that the picked element is in the set B?
- Alternately, the density of B on A could be much different from its density on the whole space.
- In the first case $P(B|A) = P(B)$, we say that B is **independent** of A . While in the second case, it is B is **dependent** on A. ($P(B|A) \neq P(B)$).
- In this case, knowing an element belongs to A does not tell us anything more about whether it also belongs to B

Conditional independence 条件独立性

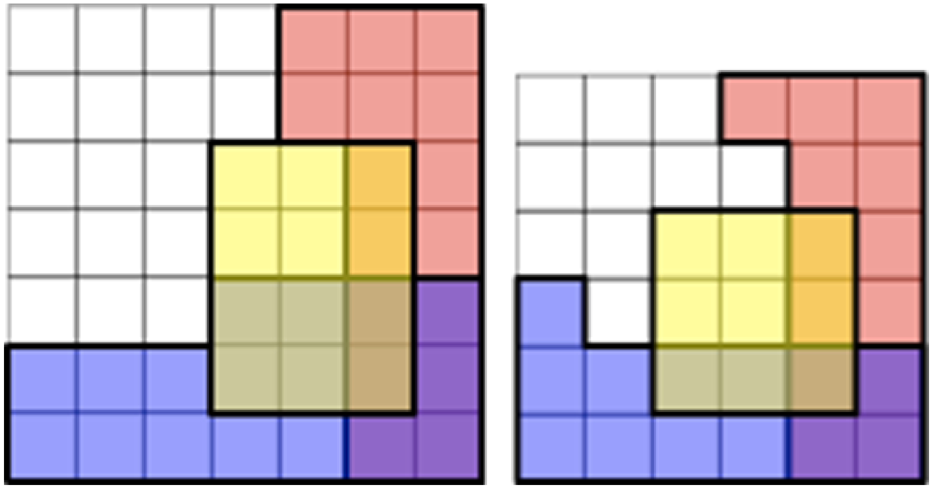
- Say we have already learned that a randomly picked element has property A.
- We want to know whether or not the element has property B:
  - $Pr(B|A)$ expresses the probability of this being true.
- Now we learn that the element also has property C. Does this give us more information about B-ness?
  - $Pr(B|A \cap C)$ expresses the probability of this being true under the additional information.

Conditional independence 条件独立性

- If $Pr(B|A \cap C) = Pr(B|A)$, then we have not gained any additional information from knowing that the element is in C.
- In this case we say that B is conditionally independent of C given A.
- That is, once we know A, additionally knowing C is irrelevant (it will give us no more information as to the value of whether or not B is true).
- Conditional independence is independence in the conditional probability space $Pr(\bullet|A)$.

Conditional independence 条件独立性

Note !



These pictures represent the probabilities of event sets A, B and C by the areas shaded red, blue and yellow respectively with respect to the total area. In both examples A and B are conditionally independent given C because:

$$P(A \wedge B|\ C) = P(A|C)P(B|C)$$

BUT A and B are NOT conditionally independent given $\neg C$, as:

$$P(A \wedge B|\neg C) \neq P(A|\neg C)P(B|\neg C)$$

Computational Impact of Independence独立性下的计算

- If A and B are independent, then $Pr(A \cap B) = Pr(A) \cdot Pr(B)$

  Proof:

  P(B|A) = P(B) (def'n of independence)

  P(A∧B)/P(A) = P(B)

  P(A∧B) = P(B) * P(A)

- If given A, B and C are conditionally independent, then $Pr(B \cap C|A) = Pr(B|A) \cdot Pr(C|A)$

  Proof:

  P(B|C∧A) = P(B|A) (def'n of conditional independence)

  P(B∧C∧A)/P(C∧A) = P(B∧A)/P(A)

  P(B∧C∧A)/P(A) = P(C∧A)/P(A) * P(B∧A)/P(A)

  P(B∧C|A) = P(B|A) * P(C|A) .

Computational Impact of Independence独立性下的计算

- Independence property allows us to "break" up the computation of a conjunction "$P(A \wedge B)$" into two separate computations "$P(A)$" and "$P(B)$". $P(B \wedge C|A)$ into $P(B|A)$ and $P(C|A)$, $P(B|A \wedge C)$ into $P(B|A)$
- This can yield great computational savings.

Review: Chain rule 链式法则

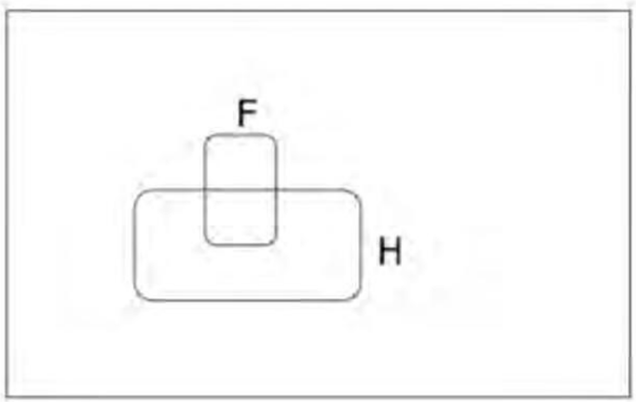
$$P(A_1 \wedge A_2 \wedge \ldots \wedge A_n) =$$
$$P(A_1 | A_2 \wedge \ldots \wedge A_n) * P(A_2 | A_3 \wedge \ldots \wedge A_n)$$
$$* \ldots * P(A_{n-1} | A_n) * P(A_n)$$

Proof:

$$P(A_1 | A_2 \wedge \ldots \wedge A_n) * P(A_2 | A_3 \wedge \ldots \wedge A_n) * \ldots * P(A_{n-1} | A_n)$$
$$= P(A_1 \wedge A_2 \wedge \ldots \wedge A_n) / P(A_2 \wedge \ldots \wedge A_n) * P(A_2 \wedge \ldots \wedge A_n) / P(A_3 \wedge \ldots \wedge A_n) * \ldots * P(A_{n-1} \wedge A_n) / P(A_n) * P(A_n)$$

$$Pr(A_1 \cap A_2 \cap A_3)$$
$$= Pr(A_1 | A_2 \cap A_3) Pr(A_2 \cap A_3)$$
$$= Pr(A_1 | A_2 \cap A_3) Pr(A_2 | A_3) Pr(A_3)$$

流感案例分析



P(Headache=true) = 1/10

P(Flu=true) = 1/40

P(Headache=true|Flu=true) = 1/2

Headaches are rare and having flu is rarer. But, given flu, there is a 50% chance you have a headache.

What is P(Flu=true|Headache=true)?

$$
\begin{aligned}
P(Flu|Headache) &= P(Flu \wedge Headache)/P(Headache) \\
&= P(Flu \wedge Headache)/P(Flu) * P(Flu)/P(Headache) \\
&= \mathbf{P(Headache|\ Flu)P(Flu)/P(Headache)}
\end{aligned}
$$

What we just did

We Derived Bayes' Rule.

$$P(Y|X) = P(X|Y)P(Y)/P(X)$$

$$\begin{aligned} P(Y|X) &= P(Y \wedge X)/P(X) \\ &= P(Y \wedge X)/P(Y) * P(Y)/P(X) \\ &= P(X|Y)P(Y)/P(X) \end{aligned}$$

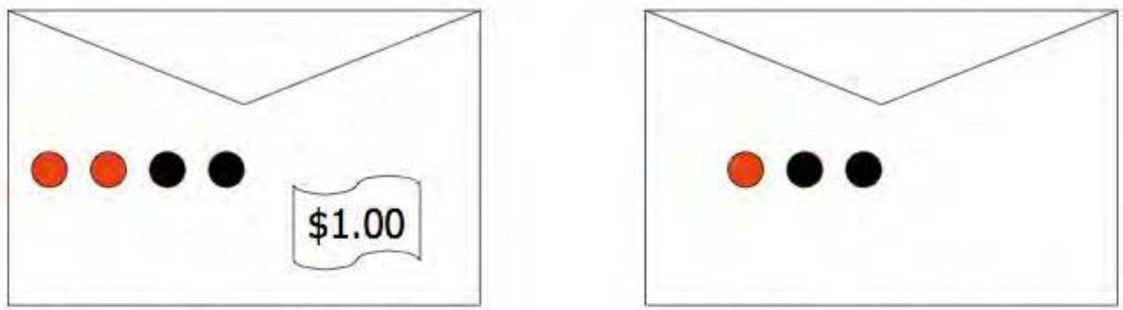
What we just did, more formally

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

This is Bayes Rule

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

Using Bayes Rule to Gamble

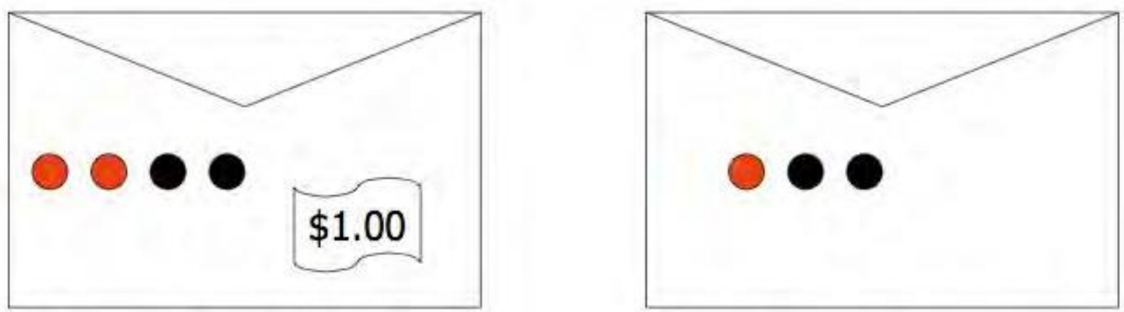


The "Win" envelope has a dollar and four beads in it

The "Lose" envelope has three beads and no money

Trivial question: Someone picks an envelope and random and asks you to bet as to whether or not it holds a dollar. What are your odds?

Using Bayes Rule to Gamble



The "Win" envelope has a dollar and four beads in it

The "Lose" envelope has three beads and no money

Not trivial question: Someone lets you take a bead out of the envelope before you bet. If it is black, what are your odds? If it is red, what are your odds?

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

后验概率　先验概率　调整因子

Bayes rule贝叶斯法则

- Bayes rule is a simple mathematical fact. But it has great implications wrt how probabilities can be reasoned with.

$$Pr(Y|X) = Pr(X|Y)Pr(Y)/Pr(X)$$

- *e.g.*, from treating patients with heart disease we might be able to estimate the value of
- *Pr*(*high Cholesterol*高胆固醇|*heart disease*心脏病)
- With Bayes rule, we can turn this around into a predictor for heart disease
- *Pr*(*heart disease*心脏病 | *high Cholesterol*高胆固醇)
- With a simple blood test we can determine "high Cholesterol", and use it to help estimate the likelihood of heart disease.

Bayes Rule Example 贝叶斯法则案例

- Disease ∈ {*malaria*疟疾, cold感冒 , flu流感};
- Symptom = fever 发烧
- Must compute $Pr(Disease|fever)$ to prescribe treatment. Why not assess this quantity directly?
  - $Pr(mal|fever)$ – is not natural to assess. It does not reflect the underlying “causal mechanism” malaria ⇒ fever 难以直接评估，无法反映内在因果机制
  - $Pr(mal|fever)$ – is not “stable”: a malaria epidemic changes this quantity (for example) 随环境不断变化
- So we use Bayes rule:
- $Pr(mal|fever) = Pr(fever|mal)Pr(mal)/Pr(fever)$

Pr(mal)表示当前环境下患**疟疾概率**，反映的就是环境的影响

Bayes Rule Example 贝叶斯法则案例

- What about ***Pr(fever)***
- Say that malaria, cold and flu are the only possible causes of fever, *i.e.*, $Pr(fever \mid \neg malaria \wedge \neg cold \wedge \neg flu) = 0$, and they are mutually exclusive.
- Then $\boldsymbol{Pr(fever)} = \boldsymbol{Pr(malaria \wedge fever)} + \boldsymbol{Pr(cold \wedge fever)} + \boldsymbol{Pr(flu \wedge fever)}$

  其中 $Pr(malaria \wedge fever) = Pr(fever \mid mal)Pr(mal)$
- Similarly, we can obtain $Pr(cold \wedge fever)$ and $Pr(flu \wedge fever)$

Useful equations 常用的计算公式总结

- Conditional probability: $Pr(B|A) = Pr(B \cap A)/Pr(A)$
- Summing out rule:
  Say that $B_1, B_2, \ldots, B_k$ form a partition of U. Then
  $Pr(A) = Pr(A \cap B_1) + \ldots + Pr(A \cap B_k)$
- If A and B are independent, then
  $Pr(A \cap B) = Pr(A) \cdot Pr(B)$
- If given A, B and C are conditionally independent, then
  $Pr(B \cap C|A) = Pr(B|A) \cdot Pr(C|A)$
- Bayes rule: $Pr(Y|X) = Pr(X|Y)Pr(Y)/Pr(X)$
- Chain rule: $Pr(A_1 \cap A_2 \cap \ldots \cap A_n) = Pr(A_1|A_2 \cap \ldots \cap A_n) \cdot$
  $Pr(A_2|A_3 \cap \ldots \cap A_n) \cdot \ldots \cdot Pr(A_{n-1}|A_n) \cdot Pr(A_n)$

扩展到连续空间分布

- $Pr(X)$ for variable X refers to the (marginal边际) distribution over X. $Pr(X|Y)$ refers to family of conditional distributions over X, one for each $y \in Dom(Y)$.
- For each $d \in Dom[Y]$, $Pr(X|Y = d)$ specifies a distribution over the values of X: $Pr(X = d_1|Y = d), Pr(X = d_2|Y = d), \ldots, Pr(X = d_n|Y = d)$, where $Dom[X] = \{d_1, d_2, \ldots, d_n\}$.
- Distinguish between $Pr(X)$ which is distribution and $Pr(X = d)$ ($d \in Dom[X]$) – which is a number.

Think of $Pr(X)$ as a function that accepts any $x \in Dom[X]$ as an argument and returns $Pr(X = x)$.

Similarly, think of $Pr(X|Y)$ as a function that accepts any $y \in Dom[Y]$ and returns a distribution $Pr(X|Y = y)$.

贝叶斯推断

对条件概率公式进行变形，可以得到

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

后验概率　　先验概率　调整因子

P(A)称为"先验概率"（Prior probability），即在B事件发生之前，我们对A事件概率的一个判断。

P(A|B)称为"后验概率"（Posterior probability），即在B事件发生之后，我们对A事件概率的重新评估。

P(B|A)/P(B)称为"可能性函数"（Likelyhood），这是一个调整因子，使得预估概率更接近真实概率。

贝叶斯推断的含义

条件概率可以理解为

后验概率 ＝ 先验概率 × 调整因子

贝叶斯推断含义：我们先预估一个"先验概率"，然后加入实验结果，看这个实验到底是增强还是削弱了"先验概率"，由此得到更接近事实的"后验概率"。

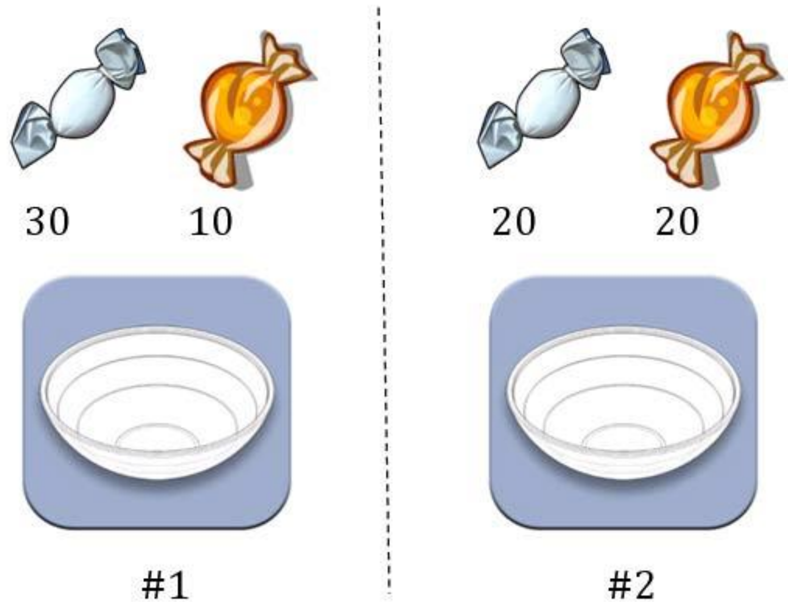
如果 "调整因子">1，意味着"先验概率"被增强，事件A的发生的可能性变大；

如果 "调整因子"=1，意味着B事件无助于判断事件A的可能性；

如果 "调整因子"<1，意味着"先验概率"被削弱，事件A的可能性变小。

【例子】水果糖问题

两个一模一样的碗，一号碗有30颗水果糖和10颗巧克力糖，二号碗有水果糖和巧克力糖各20颗。现在随机选择一个碗，从中摸出一颗糖，发现是水果糖。请问这颗水果糖来自一号碗的概率有多大?



【例子】水果糖问题

假定，H1表示一号碗，H2表示二号碗。由于这两个碗是一样的，所以P(H1)=P(H2)，也就是说，在取出水果糖之前，这两个碗被选中的概率相同。因此，P(H1)=0.5，我们把这个概率就叫做"先验概率"，即没有做实验之前，来自一号碗的概率是0.5。

再假定，E表示水果糖，所以问题就变成了在已知E的情况下，来自一号碗的概率有多大，即求P(H1|E)。我们把这个概率叫做"后验概率"，即在E事件发生之后，对P(H1)的修正。

由条件概率公式，有

$$P(H_1|E) = P(H_1)\frac{P(E|H_1)}{P(E)}$$

【例子】水果糖问题

已知，P(H1)等于0.5，P(E|H1)为一号碗中取出水果糖的概率，等于0.75，那么求出P(E)就可以得到答案。

根据全概率公式

$$P(E)=P(E|H_1)P(H_1)+P(E|H_2)P(H_2)$$

所以

$$P(E)=0.75\times0.5+0.5\times0.5=0.625$$

代入数据得

$$P(H1|E)=0.5\times\frac{0.75}{0.625}=0.6$$

这表明，来自一号碗的概率是0.6。也就是说，取出水果糖之后，H1事件的可能性得到了增强。

【例子】别墅与狗

一座别墅在过去的 20*365天里一共发生过 2 次被盗，别墅的主人有一条狗，狗平均每周晚上叫 3 次，在盗贼入侵时狗叫的概率被估计为 0.9，问题是：在狗叫的时候发生入侵的概率是多少？TRY

- 我们假设 A 事件为狗在晚上叫，B 为盗贼入侵，
  $P(A) = 3/7$
  $P(B)=2/(20\cdot 365)=2/7300$
  $P(A \mid B) = 0.9$
- 按照公式很容易得出结果：
  $P(B|A)=0.9*(2/7300)/(3/7)=0.00058$

【例子】假阳性问题

已知某种疾病的发病率是0.001，即1000人中会有1个人得病。现有一种试剂可以检验患者是否得病，它的准确率是0.99，即在患者确实得病的情况下，它有99%的可能呈现阳性。它的误报率是5%，即在患者没有得病的情况下，它有5%的可能呈现阳性。现有一个病人的检验结果为阳性，请问他确实得病的可能性有多大?

【例子】假阳性问题

假定A事件表示得病，那么P(A)为0.001。这就是"先验概率"，即没有做试验之前，我们预计的发病率。再假定B事件表示阳性，那么要计算的就是P(A|B)。这就是"后验概率"，即做了试验以后，对发病率的估计。

根据公式

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

用全概率公式改写分母

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

结果

$$P(A|B)=0.001\times\frac{0.99}{0.99\times 0.001+0.05\times 0.999}\approx 0.019$$

【例子】假阳性问题

P(A|B)约等于0.019。也就是说，即使检验呈现阳性，病人得病的概率，也只是从0.1%增加到了2%左右。这就是所谓的“假阳性”，即阳性结果完全不足以说明病人得病。为什么?

原因：误报率太高、疾病发生概率低

$$P(A|B)=0.001\times\frac{0.99}{0.99\times0.001+0.05\times0.999}\approx0.019$$

Bayesian Classifiers 贝叶斯分类器

Consider each attribute and class label as random variables 比如给定一位学生人工智能、明清文学、宋明理学、常微分方程几门课的成绩attribute，判别他是计算机学院还是历史学院class label

Given a record with attributes (A_1, A_2,…,A_n)

- Goal is to predict class C
- Specifically, we want to find the value of C that maximizes P(C| A_1, A_2,…,A_n)

Can we estimate P(C| A_1, A_2,…,A_n) directly from data?

Bayesian Classifiers贝叶斯分类器

Approach:

- compute the posterior probability $P(C \mid A_1, A_2, \ldots, A_n)$ for all values of C using the Bayes theorem

$$\boxed{P(C \mid A_1 A_2 \ldots A_n) = \frac{P(A_1 A_2 \ldots A_n \mid C) P(C)}{P(A_1 A_2 \ldots A_n)}}$$

- Choose value of C that maximizes
  $P(C \mid A_1, A_2, \ldots, A_n)$

- Equivalent to choosing value of C that maximizes
  $P(A_1, A_2, \ldots, A_n \mid C)\, P(C)$

How to estimate $P(A_1, A_2, \ldots, A_n \mid C)$?

Naïve Bayes Classifier朴素贝叶斯分类器

Assume independence among attributes A_i when class is given:

- $P(A_1, A_2, \ldots, A_n | C) = P(A_1| C_j) P(A_2| C_j)\ldots P(A_n| C_j)$

- Can estimate $P(A_i| C_j)$ for all A_i and C_j.

- New point is classified to C_j if

  $$P(C_j) P(A_1| C_j) P(A_2| C_j)\ldots P(A_n| C_j)$$

  is maximal.

How to Estimate Probabilities from Data?

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

如何处理连续数值

Class: $P(C) = N_c/N$

- e.g., $P(No) = 7/10$,
  $P(Yes) = 3/10$

For discrete attributes:

$$P(A_i \mid C_k) = |A_{ik}| / N_{c\,k}$$

- where $|A_{ik}|$ is number of instances having attribute A_i and belongs to class C_k
- Examples:

P(Status=Married|No) = 4/7
P(Refund=Yes|Yes)=0

How to Estimate Probabilities from Data?

For continuous attributes:

- 离散化 Discretize the range into bins
  - Large interval number: too few training records for reliable estimate
  - Small interval number: aggregate records from different classes
- 概率密度估计 Probability density estimation:
  - Assume attribute follows a normal distribution
  - Use data to estimate parameters of distribution (e.g., mean and standard deviation)
  - Once probability distribution is known, can use it to estimate the conditional probability $P(A_i|C_j)$

How to Estimate Probabilities from Data?

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Normal distribution:

$$P(A_i \mid C_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(A_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

- One for each (A_i, C_i) pair

For (Income, Class=No):

- If Class=No
  - sample mean = 110
  - sample variance = 2975

$$P(Income = 120 \mid No) = \frac{1}{\sqrt{2\pi \times 2975}} e^{-\frac{(120-110)^2}{2 \times 2975}} = 0.0072$$

Example of Naïve Bayes Classifier

Given a Test Record:

$$X = (\text{Refund} = \text{No}, \text{Married}, \text{Income} = 120\text{K})$$

naive Bayes Classifier:

P(Refund=Yes|No) = 3/7
P(Refund=No|No) = 4/7
P(Refund=Yes|Yes) = 0
P(Refund=No|Yes) = 1
P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No)=1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/7
P(Marital Status=Divorced|Yes)=1/7
P(Marital Status=Married|Yes) = 0 ?

For taxable income:
If class=No: sample mean=110
sample variance=2975
If class=Yes: sample mean=90
sample variance=25

P(X|Class=No) = P(Refund=No|Class=No) $\times$ P(Married| Class=No) $\times$ P(Income=120K| Class=No)
= 4/7 $\times$ 4/7 $\times$ 0.0072 = 0.0024

P(X|Class=Yes) = P(Refund=No| Class=Yes) $\times$ P(Married| Class=Yes) $\times$ P(Income=120K| Class=Yes)
= 1 $\times$ 0 $\times$ 1.2 $\times$ 10^{-9} = 0

Since P(X|No)P(No) > P(X|Yes)P(Yes)

Therefore P(No|X) > P(Yes|X)
=> Class = No

Naïve Bayes Classifier

If one of the conditional probability is zero, then the entire expression becomes zero

Probability estimation:

$$\text{Original}: P(A_i \mid C) = \frac{N_{ic}}{N_c}$$

$$\text{Laplace}: P(A_i \mid C) = \frac{N_{ic} + 1}{N_c + c}$$

$$\text{m - estimate}: P(A_i \mid C) = \frac{N_{ic} + mp}{N_c + m}$$

c: number of classes

p: prior probability

m: parameter

Thanks

*Some Slides based on those of Prof. Sheila McIlraith, Prof. Dan Klein and Prof. Pieter Abbeel. Thanks for their support.