Probability review

Adopted from notes of Andrew W. Moore and Eric Xing from CMU

So far our classifiers are deterministic!

- For a given X, the classifiers we learned so far give a single predicted y value
- In contrast, a probabilistic prediction returns a probability over the output space

  P(y=0|X), P(y=1|X)
- We can easily think of situations when this would be very useful!
  - Given P(y=1|X) =0.49, P(y=-1|X)=0.51, how would you predict?
  - What if I tell you it is much more costly to miss an positive example than the other way around?

Discrete Random Variables

- A is a Boolean-valued random variable if A denotes an event, and there is some degree of uncertainty as to whether A occurs.
- Examples
- A = The US president in 2023 will be male
- A = You wake up tomorrow with a headache
- A = You have Ebola

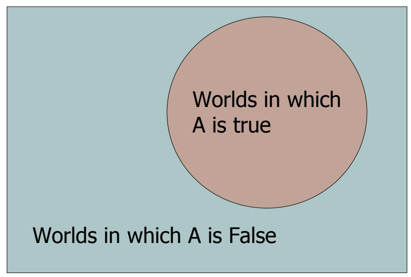
Probabilities

- We write P(A) as "the fraction of possible worlds in which A is true"
- We could at this point spend 2 hours on the philosophy of this.
- But we won't.

Visualizing A

Event space of all possible worlds

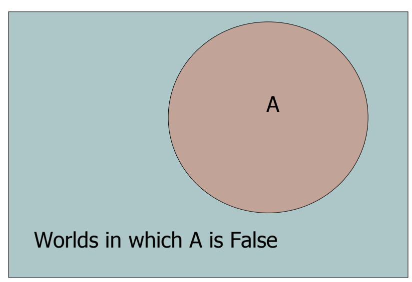
Its area is 1

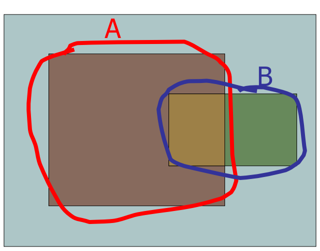


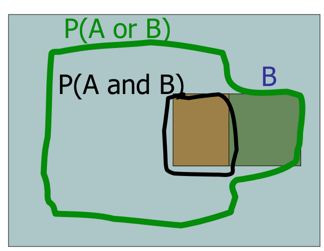
P(A) = Area of reddish oval

Basic axioms

- 0 <= P(A) <= 1
- P(True) = 1
- P(False) = 0
- P(A or B) = P(A) + P(B) - P(A and B)







Simple addition and subtraction

Elementary Probability Theorems

- P(~A) + P(A) = 1
- P(B) = P(B ^ A) + P(B ^ ~A)

Multivalued Random Variables

- Suppose A can take on more than 2 values
- A is a *random variable with arity k* if it can take on exactly one value out of $\{v_1, v_2, .. v_k\}$
- Thus…

$$P(A = v_i \wedge A = v_j) = 0 \text{ if } i \neq j$$
$$P(A = v_1 \vee A = v_2 \vee A = v_k) = 1$$

An easy fact about Multivalued Random Variables:

- Using the axioms of probability…

  0 <= P(A) <= 1, P(True) = 1, P(False) = 0

  P(A or B) = P(A) + P(B) - P(A and B)
- And assuming that A obeys…

$$P(A = v_i \wedge A = v_j) = 0 \text{ if } i \neq j$$
$$P(A = v_1 \vee A = v_2 \vee A = v_k) = 1$$

- It's easy to prove that

$$P(A = v_1 \vee A = v_2 \vee A = v_i) = \sum_{j=1}^{i} P(A = v_j)$$

An easy fact about Multivalued Random Variables:

- Using the axioms of probability…

  0 <= P(A) <= 1, P(True) = 1, P(False) = 0

  P(A or B) = P(A) + P(B) - P(A and B)
- And assuming that A obeys…

$$P(A = v_i \wedge A = v_j) = 0 \text{ if } i \neq j$$
$$P(A = v_1 \vee A = v_2 \vee A = v_k) = 1$$

- It's easy to prove that

$$P(A = v_1 \vee A = v_2 \vee A = v_i) = \sum_{j=1}^{i} P(A = v_j)$$

- And thus we can prove

$$\sum_{j=1}^{k} P(A = v_j) = 1$$

Another fact about Multivalued Random Variables:

- Using the axioms of probability…
  0 <= P(A) <= 1, P(True) = 1, P(False) = 0
  P(A or B) = P(A) + P(B) - P(A and B)
- And assuming that A obeys…

$$P(A=v_i \wedge A=v_j)=0 \text{ if } i \neq j$$
$$P(A=v_1 \vee A=v_2 \vee A=v_k)=1$$

- It's easy to prove that

$$P(B \wedge [A=v_1 \vee A=v_2 \vee A=v_i]) = \sum_{j=1}^{i} P(B \wedge A=v_j)$$

Another fact about Multivalued Random Variables:

- Using the axioms of probability…
  0 <= P(A) <= 1, P(True) = 1, P(False) = 0
  P(A or B) = P(A) + P(B) - P(A and B)
- And assuming that A obeys…

$$P(A=v_i \wedge A=v_j)=0 \text{ if } i \neq j$$
$$P(A=v_1 \vee A=v_2 \vee A=v_k)=1$$

- It's easy to prove that

$$P(B \wedge [A=v_1 \vee A=v_2 \vee A=v_i]) = \sum_{j=1}^{i} P(B \wedge A=v_j)$$

- And thus we can prove

$$P(B) = \sum_{j=1}^{k} P(B \wedge A=v_j)$$

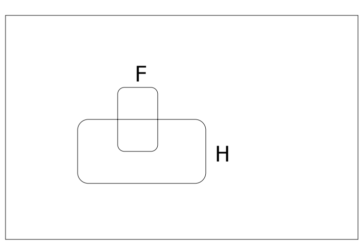
Elementary Probability in Pictures

$$\sum_{j=1}^{k} P(A = v_j) = 1$$

$$P(B) = \sum_{j=1}^{k} P(B \wedge A = v_j)$$

Conditional Probability

- P(A|B) = Fraction of worlds in which B is true that also have A true

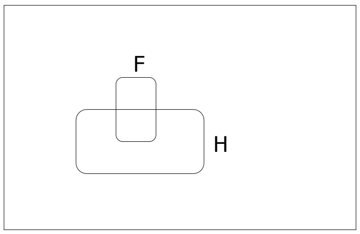


H = "Have a headache"
F = "Coming down with Flu"

P(H) = 1/10
P(F) = 1/40
P(H|F) = 1/2

"Headaches are rare and flu is rarer, but if you're coming down with 'flu there's a 50-50 chance you'll have a headache."

Conditional Probability



H = "Have a headache"
F = "Coming down with Flu"

P(H) = 1/10
P(F) = 1/40
P(H|F) = 1/2

P(H|F) = Fraction of flu-inflicted worlds in which you have a headache

= #worlds with flu and headache

#worlds with flu

= Area of "H and F" region

Area of "F" region

= P(H ^ F)

P(F)

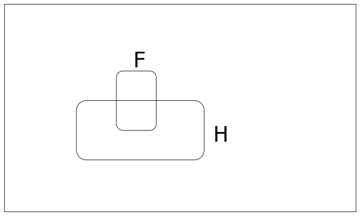
Definition of Conditional Probability

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

Corollary: The Chain Rule

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

Probabilistic Inference



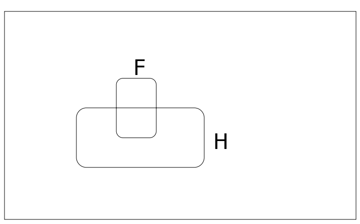
H = "Have a headache"
F = "Coming down with Flu"

P(H) = 1/10
P(F) = 1/40
P(H|F) = 1/2

One day you wake up with a headache. You think: "Drat! 50% of flus are associated with headaches so I must have a 50-50 chance of coming down with flu"

Is this reasoning good?

Probabilistic Inference



H = "Have a headache"
F = "Coming down with Flu"

P(H) = 1/10
P(F) = 1/40
P(H|F) = 1/2

P(F ^ H) = …

P(F|H) = …

What we just did…

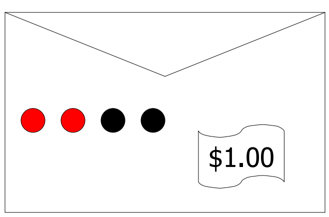
$$P(B|A) = \frac{P(A \wedge B)}{P(A)} = \frac{P(A|B)\,P(B)}{P(A)}$$

This is Bayes Rule

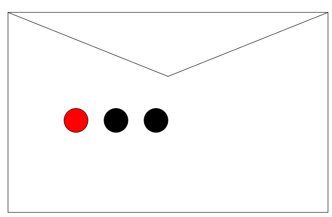
Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

 Slide 19

Using Bayes Rule to Gamble



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

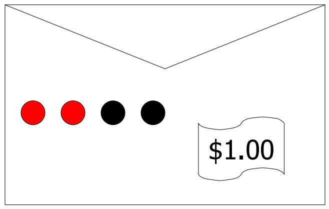


The "Lose" envelope has three beads and no money

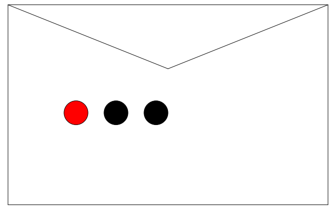
Trivial question: someone draws an envelope at random and offers to sell it to you. How much should you pay?

 Slide 20

Using Bayes Rule to Gamble



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

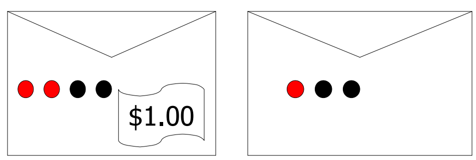


The "Lose" envelope has three beads and no money

Interesting question: before deciding, you are allowed to see one bead drawn from the envelope.

Suppose it's black: How much should you pay?
Suppose it's red: How much should you pay?

Calculation…

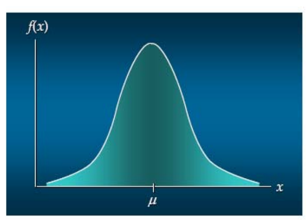


Continuous Probability Distribution

- A continuous random variable x can take any value in an interval on the real line
  - X usually corresponds to some real-valued measurements, e.g., today's lowest temperature
  - It is not possible to talk about the probability of a continuous random variable taking an exact value --- P(x=56.2)=0
  - Instead we talk about the probability of the random variable taking a value within a given interval $P(x \in [50, 60])$

PDF: probability density function

- The probability of X taking value in a given range [x1, x2] is defined to be the area under the PDF curve between x1 and x2
- We use *f*(x) to represent the PDF of x
- Note:
  - *f*(x) ≥0
  - *f*(x) can be larger than 1
  - $\int_{-\infty}^{\infty} f(x)dx = 1$
  - $P(X \in [x1, x2]) = \int_{x1}^{x2} f(x)dx$



What is the intuitive meaning of *f(x)*?

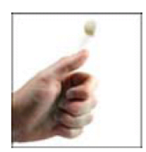
If f(x1)=α*a and f(x2)=a

Then when x is sampled from this distribution, you are α times more likely to see that x is "very close to" x1 than that x is "very close to" x2

Some commonly used distributions

Bernoulli distribution: Ber(p)

$$P(x)=\begin{cases}1-p & \text{for } x=0\\ p & \text{for } x=1\end{cases} \quad\Rightarrow\quad P(x)=p^x(1-p)^{1-x}$$



Binomial distribution: Binomial(n , p)

the probability to see x heads out of n flips

$$P(x)=\frac{n(n-1)\cdots(n-x+1)}{x!}p^x(1-p)^{n-x}$$

Multinomial distribution: Multinomial(n , [x_1, x_2, …, x_k])

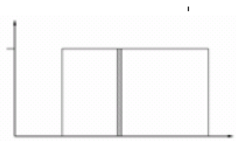
The probability to see x_1 ones, x_2 twos, etc, out of n dice rolls

$$P([x_1,x_2,\ldots,x_k])=\frac{n!}{x_1!x_2!\cdots x_k!}\theta_1^{x_1}\theta_2^{x_2}\cdots\theta_k^{x_k}$$

Continuous Distributions

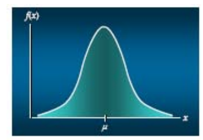
Uniform Probability Density Function

$$f(x) = 1/(b-a) \quad \text{for } a \le x \le b$$
$$= 0 \quad \text{elsewhere}$$



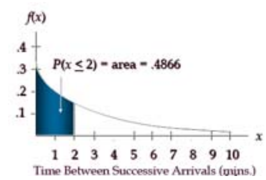
Normal (Gaussian) Probability Density Function

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



Exponential Probability Distribution

$$f(x) = \frac{1}{\mu} e^{-x/\mu}$$



The Joint Distribution

Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:

The Joint Distribution

Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M rows).

A	B	C
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

The Joint Distribution

Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M rows).
2. For each combination of values, say how probable it is.

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10

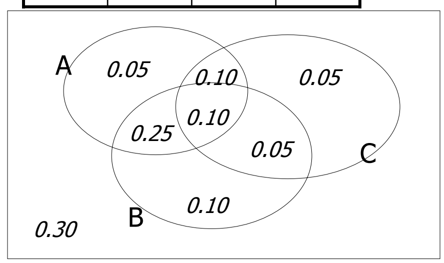
The Joint Distribution

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M rows).
2. For each combination of values, say how probable it is.
3. If you subscribe to the axioms of probability, those numbers must sum to 1.

Example: Boolean variables A, B, C

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10



Using the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

One you have the JD you can ask for the probability of any logical expression involving your attribute

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Using the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

P(Poor Male) = 0.4654

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Inference with the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

$$P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$$

Inference with the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

$$P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$$

P(Male | Poor) = 0.4654 / 0.7604 = 0.612

So we have learned that

- Joint distribution is extremely useful!
  we can do all kinds of cool inference
  - I've got a sore neck: how likely am I to have meningitis?
  - Many industries grow around Beyesian Inference: examples include medicine, pharma, Engine diagnosis etc.
- But, **HOW** do we get them?
  - We can learn from data

Learning a joint distribution

Build a JD table for your attributes in which the probabilities are unspecified

A	B	C	Prob
0	0	0	?
0	0	1	?
0	1	0	?
0	1	1	?
1	0	0	?
1	0	1	?
1	1	0	?
1	1	1	?

The fill in each row with

$$\hat{P}(\text{row}) = \frac{\text{records matching row}}{\text{total number of records}}$$

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	**0.25**
1	1	1	0.10

Fraction of all records in which A and B are True but C is False