

CS 4824/ECE 4424: Probability & Estimation

Acknowledgement:

Many of these slides are derived from Tom Mitchell, Pascal Poupart, Pieter Abbeel, Eric Eaton, Carlos Guestrin, William Cohen, and Andrew Moore.

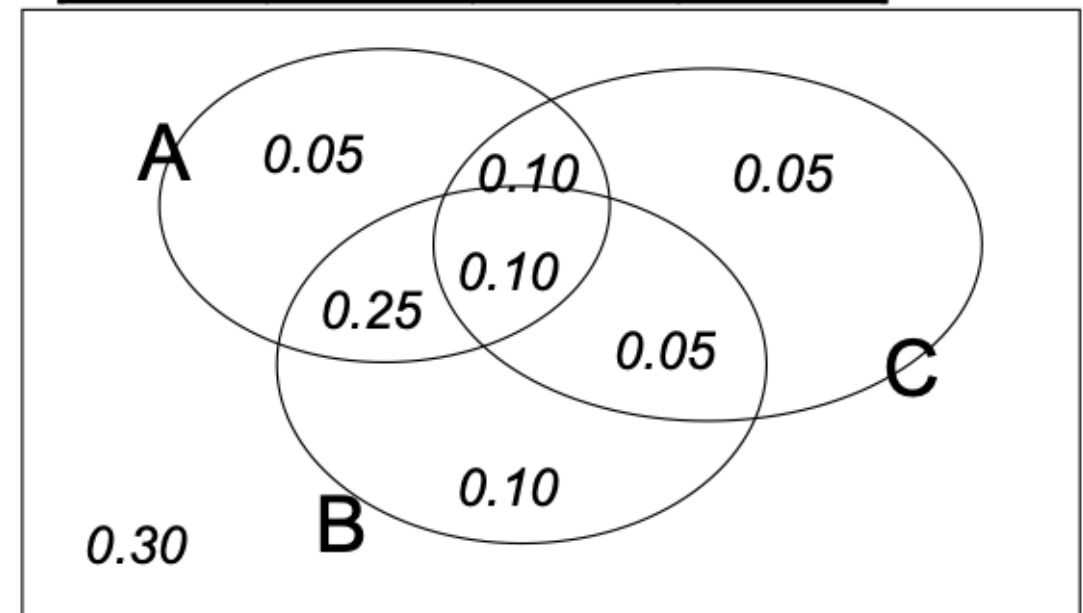
Let's turn probabilistic

- A probabilistic view of supervised function approximation
 - Instead of learning $f: X \rightarrow Y$
 - Learn $P(Y \mid X)$

The joint distribution

- Steps for coming up with a joint distribution:
 - Make a table listing all combinations of values of your variables (if there are M boolean variables then the truth table will have 2^M rows).
 - For each combination of values, say how likely it is.
 - By axioms of probability, these values must sum to 1.

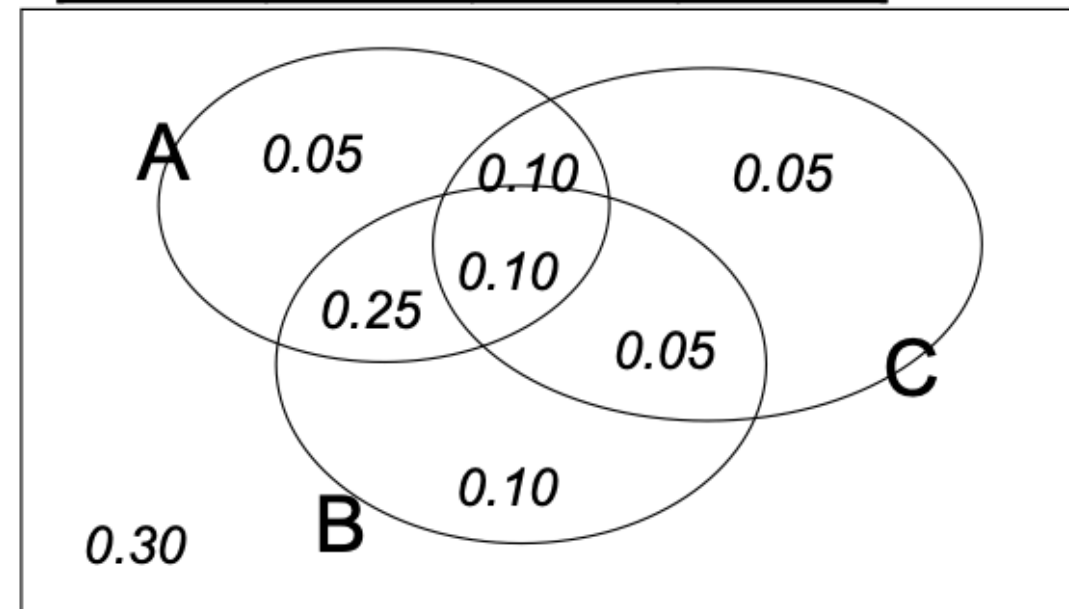
A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10



Using the joint distribution






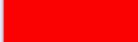


- After you have a joint distribution, you can ask for the probability of any any logical expression involving these variables.
- e.g., $P(A)$, $P(AB)$, $P(A|B)$

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



Inference with the joint distribution

- Suppose we want to learn the function $f: \langle G, H \rangle \rightarrow W$
- Or $P(W \mid G, H)$
- Learn joint distribution from data:
calculate $P(W \mid G, H)$

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 

- $P(W=\text{rich} \mid G = \text{female}, H = 40.5-) =$

This sounds like the solution to learning $P(Y|X)$!
Are we done?

**This sounds like the solution to learning
 $P(Y|X)$ or equivalently $f: X \rightarrow Y$**

Are we done?

Learning $P(Y|X)$ may require more data than we have

- Consider a joint distribution with 100 boolean attributes
 - # rows in this tables?
 - # people on earth?
 - fraction of rows with 0 training examples?

Learning $P(Y|X)$ may require more data than we have

- Consider a joint distribution with 100 boolean attributes
 - # rows in this tables?
 - # people on earth?
 - fraction of rows with 0 training examples?

The issue of
Data Sparsity!

What to do?

- Well, we need to be:
 - 1. smart about estimating probabilities from sparse data
 - maximum likelihood estimation (MLE)
 - maximum a posteriori estimation (MAP)
 - 2. smart about how to represent joint distribution
 - graphical models

Let's start by looking at how to be smart about estimating probabilities...

Elevator trials

- Given a coin, estimate the probability that it will turn up heads ($X=1$) or tails ($X=0$)
- Test A: 100 flips, 51 heads ($X=1$), 49 tail ($X=0$)
- Test B: 3 flips, 2 heads ($X=1$), 1 tail ($X=0$)

Elevator trials

- Test C: keep flipping and develop a single (online) learning algorithm that gives reasonable estimate after each flip.