Machine Learning

Topic: Basic Probability

Axioms of Probability

- Let there be a space S composed of a countable number of events

$$S \equiv \{e_1, e_2, e_3, \ldots.e_n\}$$

- The probability of each event is between 0 and 1

$$0 \leq P(e_1) \leq 1$$

- The probability of the whole sample space is 1

$$P(S) = 1$$

- **When two events are mutually exclusive**, their probabilities are additive

$$P(e_1 \vee e_2) = P(e_1) + P(e_2)$$

Discrete Random Variable

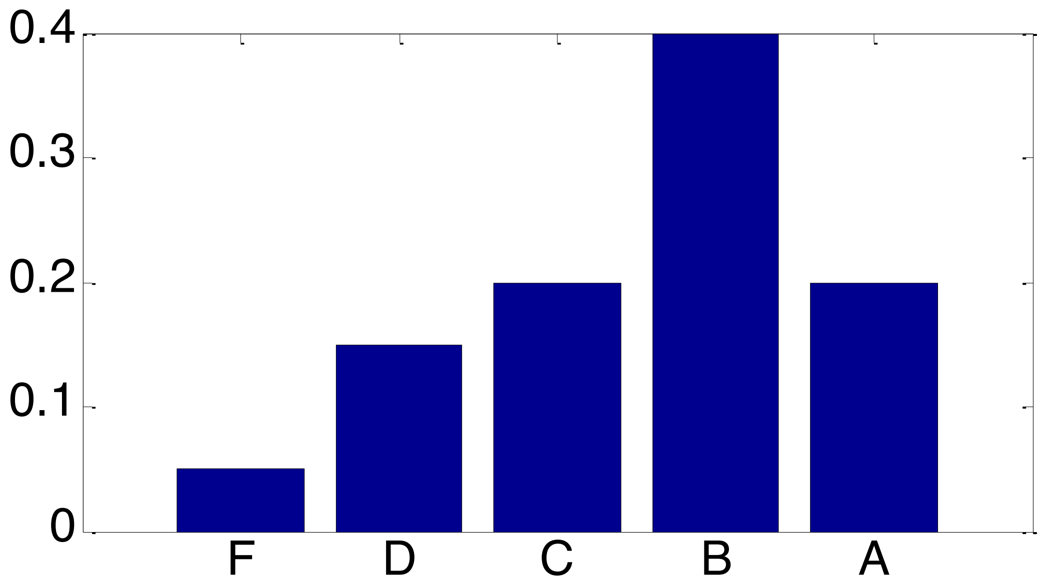
* Discrete random variable X represents some experiment.

* $P(X)$ is the probability distributions over $\{x_1, \ldots, x_n\}$, the set of possible outcomes for X.

* These outcomes are mutually exclusive.

* Their probabilities sum to one: $\sum_{i=1}^{n} P(x_i) = 1$

An Example: Your grade



GPA value	Letter grade	Probability
4	A	0.2
3	B	0.4
2	C	0.2
1	D	0.15
0	F	0.05

Boolean Random Variable

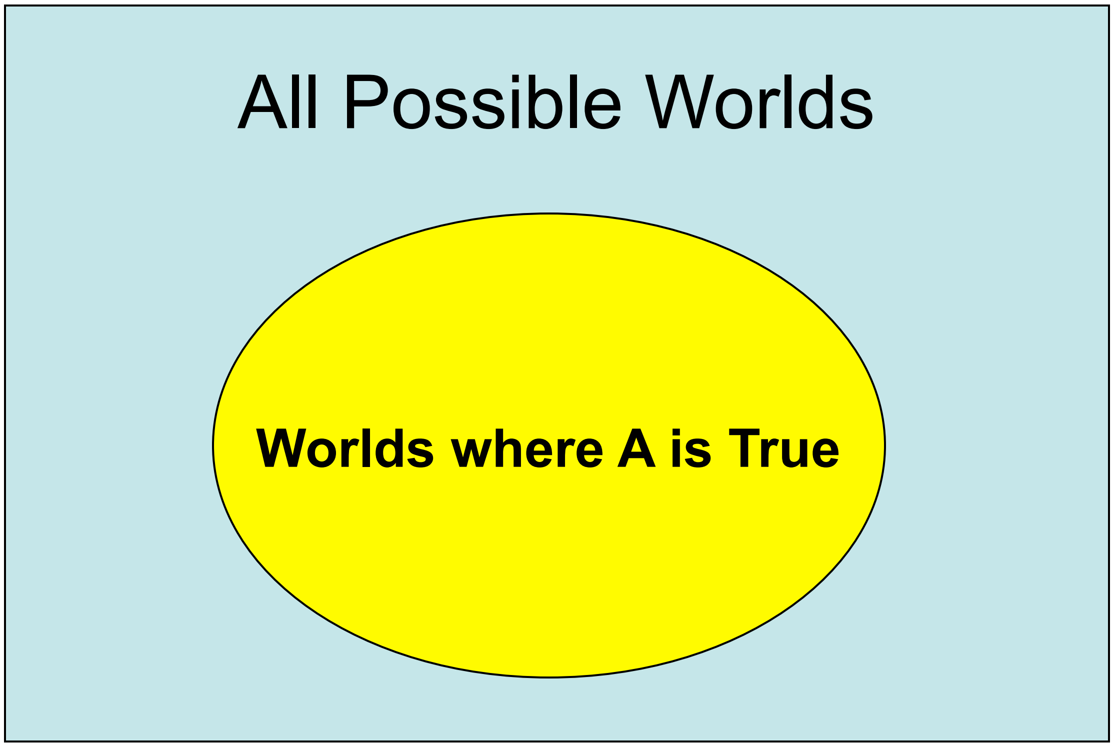
- Boolean random variable: A random variable that has only two possible outcomes

  e.g.

  X = "Tomorrow's high temperature > 60" has only two possible outcomes

  As a notational convention, **P(X)** for a Boolean variable will mean **P(X="true")**, since it is easy to infer the rest of the distribution.

Vizualizing P(A) for a Boolean variable

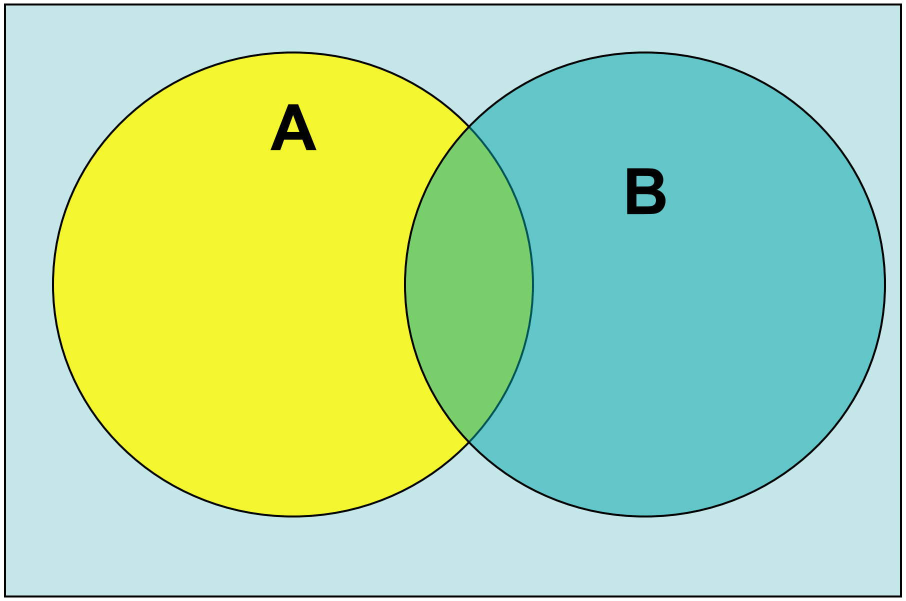


$0 \leq P(A) \leq 1$

If a value is over 1 or under 0, it isn't a probability

$$P(A) = \frac{\text{area of yellow oval}}{\text{area of blue rectangle}}$$

Vizualizing Stuff for two Booleans



$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

Independence

- variables A and B are said to be *independent* iff…

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

Bayes Rule

- Definition of Conditional Probability

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

- Corollary:
  The Chain Rule

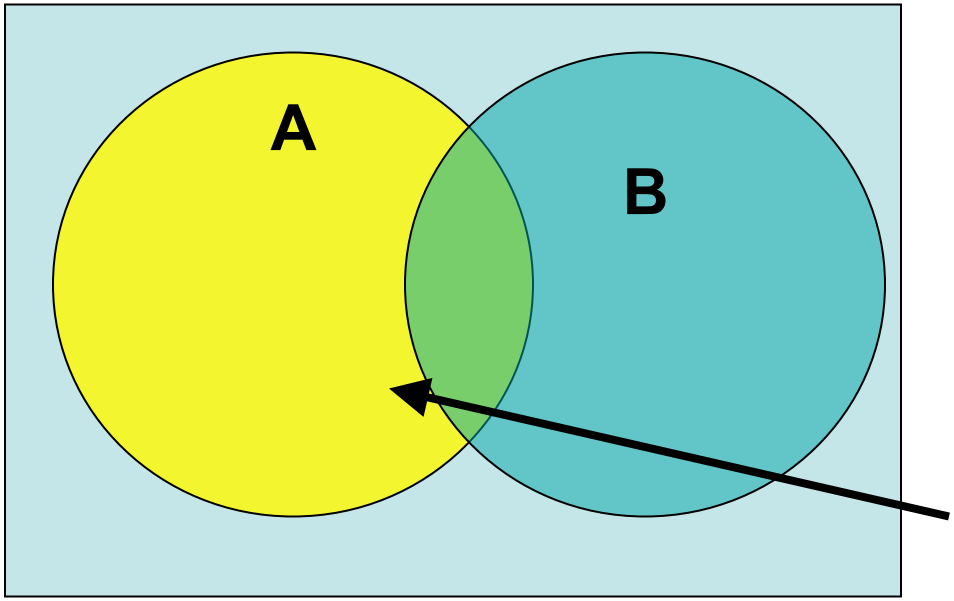
$$P(A \mid B)P(B) = P(A \wedge B)$$

- Bayes Rule
  (Thomas Bayes, 1763)

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

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

Conditional Probability



The conditional probability of A given B is represented by the following formula

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

NOT Independent

Can we do the following?

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

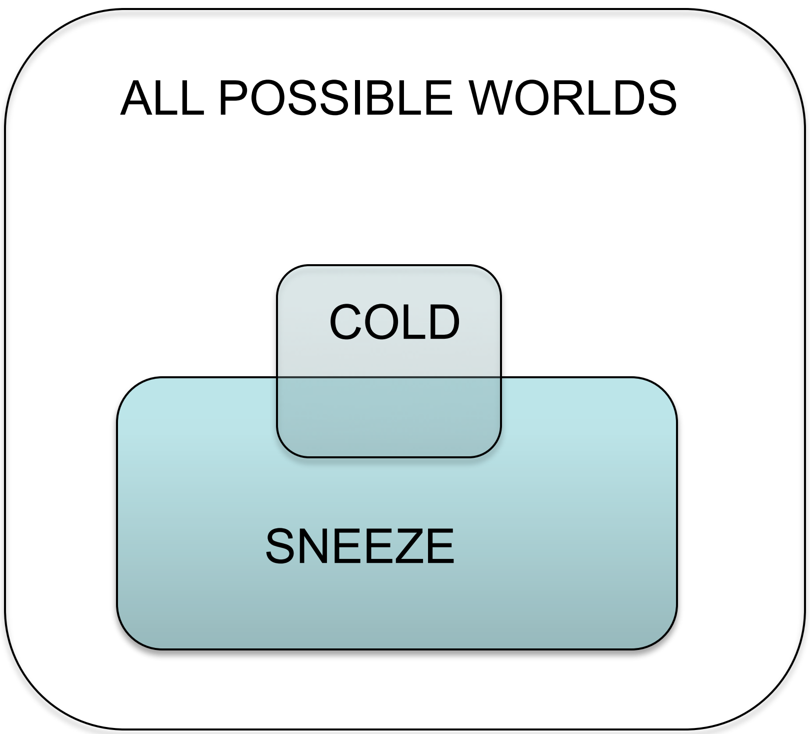
Only if A and B are ***independent***

Probabilistic Inference

0.1 = P(S) = probability of sneeze

0.05 = P(C) = probability of cold

0.5 = P(S|C) = probability of sneeze, given cold



You sneeze. Your friend says "half of all colds are associated with sneezing. You have a 50% chance of having a cold."

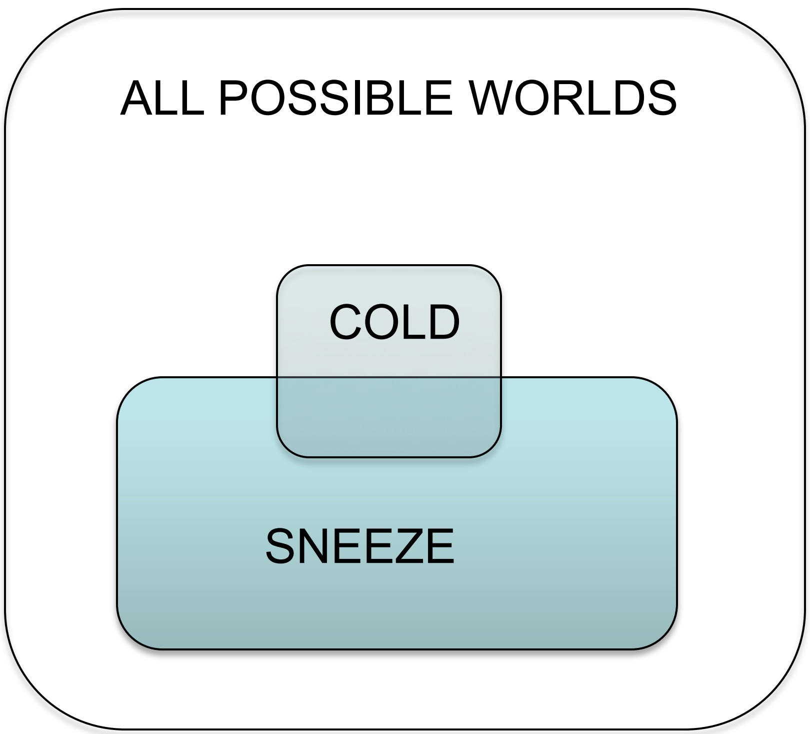
Is this reasoning sound?

Probabilistic Inference

0.1 = P(S) = probability of sneeze

0.05 = P(C) = probability of cold

0.5 = P(S|C) = probability of sneeze, given cold



Let's apply Baye's rule and see….

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

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

The Joint Distribution

- Make a truth table listing all combinations of variable values
- Assign a probability to each row
- Make sure the probabilities sum to 1

A	B	C	Prob
0	0	0	0.1
0	0	1	0.2
0	1	0	0.1
0	1	1	0.05
1	0	0	0.05
1	0	1	0.2
1	1	0	0.25
1	1	1	0.05

Using The Joint Distribution

- Find P(A)
- Sum the probabilities of all rows where A=1

P(A) = 0.05+ 0.2 + 0.25+ 0.05

= 0.55

This can be done for any set of variables and assignments…

e.g.

P(Name =“bob” ^ Age >3)

A	B	C	Prob
0	0	0	0.1
0	0	1	0.2
0	1	0	0.1
0	1	1	0.05
1	0	0	0.05
1	0	1	0.2
1	1	0	0.25
1	1	1	0.05

Using The Joint Distribution

- Find P(A|B)

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

$$= \frac{0.25 + 0.05}{0.1 + 0.05 + 0.25 + 0.05}$$

$$= \frac{0.3}{0.45}$$

$$= .666667$$

A	B	C	Prob
0	0	0	0.1
0	0	1	0.2
0	1	0	0.1
0	1	1	0.05
1	0	0	0.05
1	0	1	0.2
1	1	0	0.25
1	1	1	0.05

Using The Joint Distribution

- Are A and B Independent?

$$P(A \wedge B) = 0.3$$

$$P(A) = 0.55$$

$$P(B) = 0.45$$

$$P(A)P(B) = 0.55 * 0.45$$

$$P(A \wedge B) \neq P(A)P(B)$$

NO. They are NOT independent

A	B	C	Prob
0	0	0	0.1
0	0	1	0.2
0	1	0	0.1
0	1	1	0.05
1	0	0	0.05
1	0	1	0.2
1	1	0	0.25
1	1	1	0.05

How to learn the joint distribution?

- Experts tell us the value for each row?
- Learn from data?(e.g. US Census data)

Size of Household	Zip Code	Rent vs. Own	Name
1	60201	rent	M. Mathers
2	60201	own	W. Smith
2	90210	own	J. Lopez
1	60201	own	K. Frog
2	90210	own	O. Henry
10	60201	rent	N. Suleman

How to learn the joint distribution?

- Count all occurrences of event X.
- Divide by total number of events.

Observed Data

Size of Household	Zip Code	Rent vs. Own	Name
1	60201	rent	M. Mathers
2	60201	own	W. Smith
2	90210	own	J. Lopez
1	60201	own	K. Frog
2	90210	own	O. Henry
10	60201	rent	N. Suleman

Learned Joint Distribution

Size of Household	Zip Code	Rent vs. Own	Probability
1	60201	Rent	1/6
1	60201	Own	1/6
1	90210	Rent	0
1	90210	Own	0
2	60201	Rent	0
2	60201	Own	1/6
2	90210	Rent	0
2	90210	Own	1/3
10	60201	Rent	1/6
10	60201	Own	0
10	90210	Rent	0
10	90210	Own	0

Should these really all be 0?

Why not estimate like this?

- Given *m* boolean variables, we need to estimate 2^m values.

    20 yes-no questions = a million observable events

    How many observations do we need to feel confident estimating all these probabilities?

- Just because you didn’t observe it, doesn’t mean it never happens (black swan events).

- How do we get around this combinatorial explosion?
  - Assume independence of variables!!

…back to Independence

- Independence is DOMAIN Knowledge, often supplied by the problem designer
  - Is the probability I get a haircut independent of the probability you have an apple for lunch?
  - Is the probability Germany will close a nuclear power plant independent of the probability of a tidal wave in Japan?

- Independence implies you can learn the probability distribution for A without worrying about how it is influenced by B: $P(A \mid B) = P(A)$

- Independence also implies you can learn the probability of a joint event by multiplying the probabilities of the independent events: $P(A \wedge B \wedge C) = P(A)P(B)P(C)$

To learn the independent distributions

- Treat all variables as independent.
- Count all occurrences of event X.
- Divide by total number of events.

S: Size of Household	Probability
1	2/6
2	3/6
10	1/6

Observed Data

Size of Household	Zip Code	Rent vs. Own	Name
1	60201	rent	M. Mathers
2	60201	own	W. Smith
2	90210	own	J. Lopez
1	60201	own	K. Frog
2	90210	own	O. Henry
10	60201	rent	N. Suleman

Z: Zip Code	Probability
60201	4/6
90210	2/6

R:Rent/Own	Probability
rent	2/6
own	4/6

How to learn the joint distribution?

- Under the independence assumption, multiply independent events together to get the joint probability.

S: Size of Household	Probability
1	2/6
2	3/6
10	1/6

Z: Zip Code	Probability
60201	4/6
90210	2/6

R: Rent/Own	Probability
rent	2/6
own	4/6

Learned Joint Distribution

Size of Household	Zip Code	Rent vs. Own	Probability
1	60201	Rent	16/216
1	60201	Own	32/216
1	90210	Rent	8/216
1	90210	Own	16/216
2	60201	Rent	24/216
2	60201	Own	48/216
2	90210	Rent	12/216
2	90210	Own	24/216
10	60201	Rent	8/216
10	60201	Own	16/216
10	90210	Rent	4/216
10	90210	Own	8/216

Let's compare the distributions

Assume No Independence

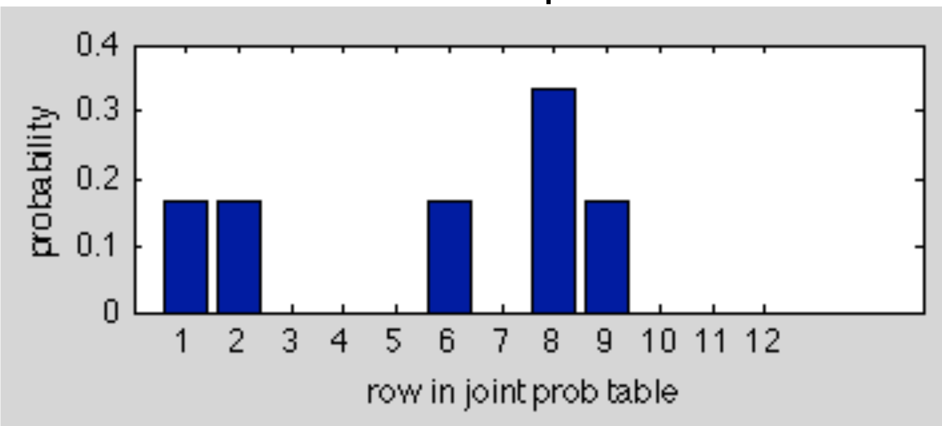
Size of Household	Zip Code	Rent vs. Own	Probability
1	60201	Rent	1/6
1	60201	Own	1/6
1	90210	Rent	0
1	90210	Own	0
2	60201	Rent	0
2	60201	Own	1/6
2	90210	Rent	0
2	90210	Own	1/3
10	60201	Rent	1/6
10	60201	Own	0
10	90210	Rent	0
10	90210	Own	0

Assume Complete Independence

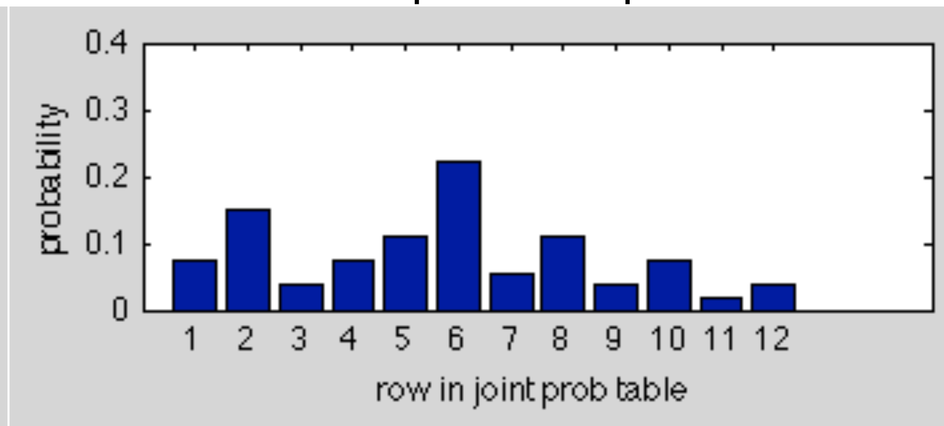
Size of Household	Zip Code	Rent vs. Own	Probability
1	60201	Rent	16/216
1	60201	Own	32/216
1	90210	Rent	8/216
1	90210	Own	16/216
2	60201	Rent	24/216
2	60201	Own	48/216
2	90210	Rent	12/216
2	90210	Own	24/216
10	60201	Rent	8/216
10	60201	Own	16/216
10	90210	Rent	4/216
10	90210	Own	8/216

Let’s compare the distributions

Assume No Independence



Assume Complete Independence



- Very different distributions learned from the same data
- How would you decide which method to use?
  - If you know it is illegal for 10 people to share a house in Beverly Hills? (two variable values are mutually exclusive)
  - If you can’t get more than 15 census questionnaires filled out? (sparse data)
- Can we assume something between complete independence and “everything is connected”?
  - We’ll discuss this later…also, take Doug Downey’s 395 class.

What kind of learning is this?

- Categorization by feedback:
  * Supervised
  * Semi-supervised
  * Reinforcement
  * Unsupervised

- Categorization by WHAT you're learning:
  * Classifier (e.g. decision tree)
  * Regressor (e.g. linear regression)
  * Probability Distribution Estimator (what we just did)