Machine Learning

Probability and Bayesian Networks

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

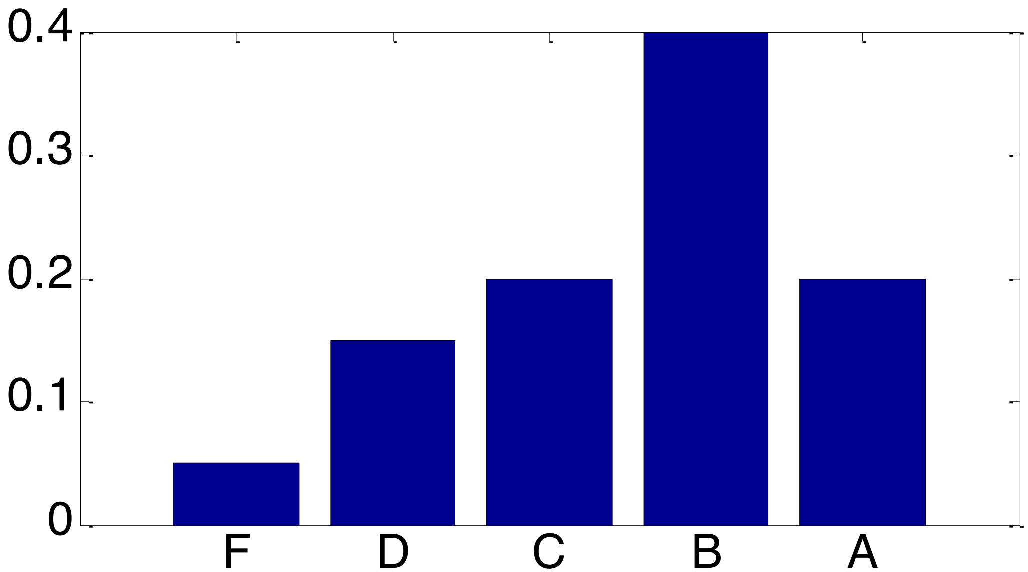
* Discrete random variable X represents some experiment.

* $P(X)$ 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

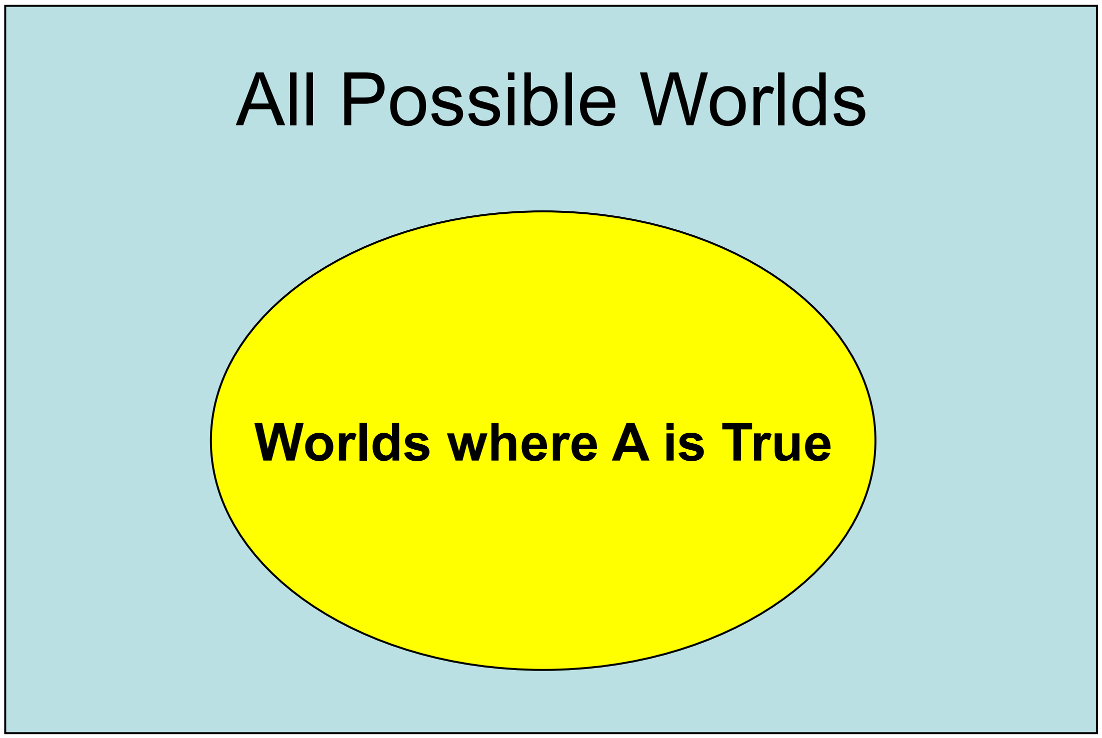
- Boolean random variable: A random variable that has only two possible outcomes

  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

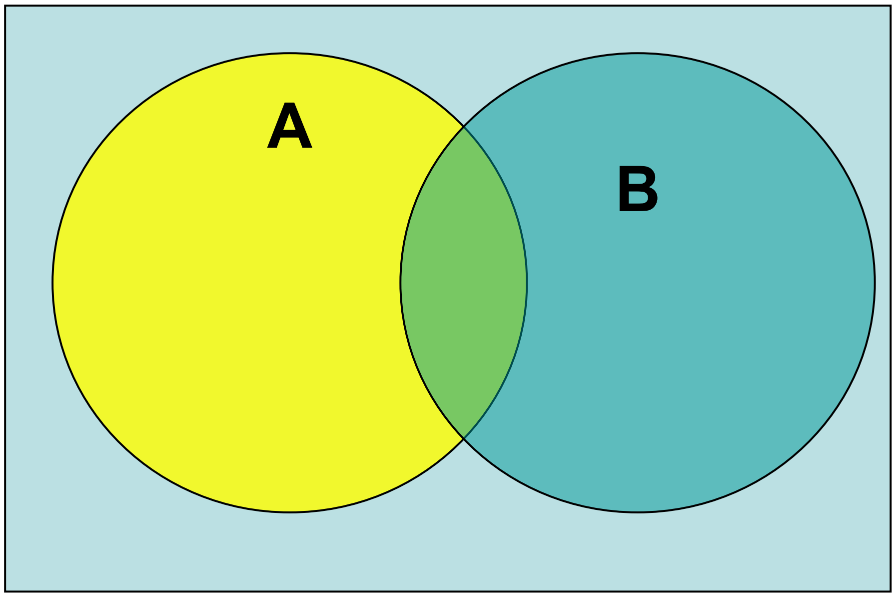


$$0 \le P(A) \le 1$$

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

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

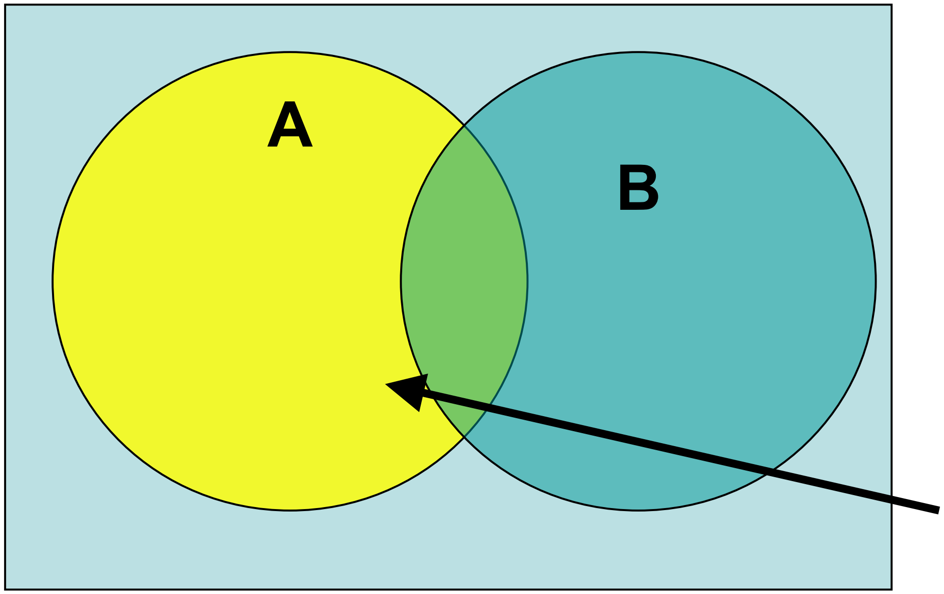
- Bayes Rule

  (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



NOT Independent

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

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

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***

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

$$P(A) = 0.05 + 0.2 + 0.25 + 0.05 = 0.55$$

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

Why not use the Joint Distribution?

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

- 20 yes-no questions = a million values

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

…back to Independence

- The probability I have an apple in my lunch bag is independent of the probability of a blizzard in Japan.
- This is DOMAIN Knowledge, typically supplied by the problem designer
- Independence implies…

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

Let's show that.

assuming independence...

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

plus the chain rule...

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

imply...

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

which means...

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

Some Definitions

- ***Prior probability of h, P(h):***
  - The background knowledge we have about the chance that *h* is a correct hypothesis (before having observed the data).

- ***Prior probability of D, P(D):***
  - the probability that training data *D* will be observed given no knowledge about which hypothesis *h* holds.

- ***Conditional Probability of D, P(D|h):***
  - the probability of observing data *D* given that hypothesis *h* holds.

- ***Posterior probability of h, P(h|D):***
  - the probability that *h* is true, given the observed training data *D*.
  - the quantity that Machine Learning researchers are interested in.

Discrete Random Variables

- What if we have three hypotheses?
- How do we predict the most likely value for a new example?

h_1 : Looks matter

h_2 : Money matters

h_3 : Ideas matter

We want a prediction: yes or no? → Obama Elected President

Maximum A Posteriori (MAP)

- **<u>Goal:</u>** To find the most probable hypothesis *h* from a set of candidate hypotheses *H* given the observed data *D*.
- ***MAP Hypothesis, h_{MAP}***

$$h_{map} = \underset{h \in H}{\arg\max}(P(h \mid D))$$

$$= \underset{h \in H}{\arg\max}\left(\frac{P(D \mid h)P(h)}{P(D)}\right)$$

$$= \underset{h \in H}{\arg\max}(P(D \mid h)P(h))$$

Maximum A Posteriori (MAP)

- Find most probable hypothesis

$$h_{map} = \underset{h \in H}{\arg\max}(P(D \mid h)P(h))$$

- Use the predictions of that hypothesis

h_1 : Looks matter

h_2 : Money matters

h_3 : Ideas matter

…. do we really want to ignore the other hypotheses?

Imagine 8 hypotheses. Seven of them say “yes” and have a probability of 0.1 each. One says “no” and has a probability of 0.3. Who do you believe?

Maximum Likelihood (ML)

- ***ML hypothesis*** is a special case of the MAP hypothesis where all hypotheses are, to begin with, equally likely

$$h_{map} = \arg\max_{h \in H}(P(D \mid h)P(h))$$

Assume...

$$P(h) = \frac{1}{|H|} \quad \forall h \in H$$

Then...

$$h_{ml} = \arg\max_{h \in H}(P(D \mid h))$$

Bayes Optimal Classifier

- An advantage of Bayesian Decision Theory
  - it gives us a lower bound on the classification error that can be obtained for a given problem.

- **Bayes Optimal Classification:** The most probable classification of a new instance is obtained by combining the predictions of all hypotheses, weighted by their posterior probabilities:

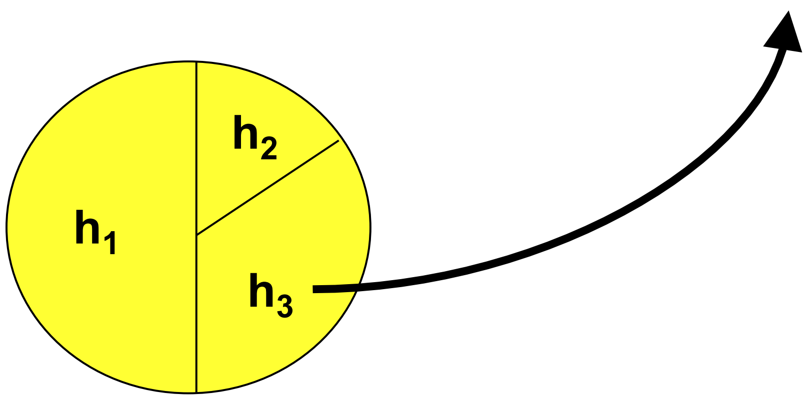
$$\arg\max_{v \in V} \sum_{h \in H} P(v \mid h) P(h \mid D)$$

…where *V* is the set of all the values a classification can take and *v* is one possible such classification.

Gibbs Classifier

- Bayes optimal classification can be too hard to compute
- Instead, randomly pick a single hypothesis (according to the probability distribution of the hypotheses)
- use this hypothesis to classify new cases

$$\arg\max_{v \in V} P(v \mid h) P(h \mid D)$$



Naïve Bayes Classifier

- Cases described by a conjunction of attribute values
  - These attributes are our "independent" hypotheses
- The target function has a finite set of values, *V*

$$v_{MAP} = \underset{v_j \in V}{\arg\max}\, P(v_j \mid a_1 \wedge a_2 ... \wedge a_n)$$

- Could be solved using the joint distribution table
- What if we have 50,000 attributes?
  - Attribute *j* is a Boolean signaling presence or absence of the *jth* word from the dictionary in my latest email.

Naïve Bayes Classifier

$$v_{MAP} = \underset{v_j \in V}{\arg\max}\, P(v_j \mid a_1 \wedge a_2 ... \wedge a_n)$$

$$= \underset{v_j \in V}{\arg\max} \frac{P(a_1 \wedge a_2 ... \wedge a_n \mid v_j) P(v_j)}{P(a_1 \wedge a_2 ... \wedge a_n)}$$

$$= \underset{v_j \in V}{\arg\max}\, P(a_1 \wedge a_2 ... \wedge a_n \mid v_j) P(v_j)$$

Naïve Bayes Continued

$$v_{MAP} = \underset{v_j \in V}{\arg\max}\, P(a_1 \wedge a_2 \ldots \wedge a_n \mid v_j) P(v_j)$$

independence step

$$v_{NB} = \underset{v_j \in V}{\arg\max}\, P(a_1 \mid v_j) P(a_2 \mid v_j) \ldots\ldots P(a_n \mid v_j) P(v_j)$$

$$= \underset{v_j \in V}{\arg\max}\, P(v_j) \prod_i P(a_i \mid v_j)$$

Instead of one table of size 2^{50000} we have 50,000 tables of size 2

Bayesian Belief Networks

- ***Bayes Optimal Classifier***
  - Often too costly to apply (uses full joint probability)

- ***Naïve Bayes Classifier***
  - Assumes conditional independence to lower costs
  - This assumption often overly restrictive

- ***Bayesian belief networks***
  - provide an ***intermediate*** approach
  - allows conditional independence assumptions that apply to ***subsets*** of the variable.