



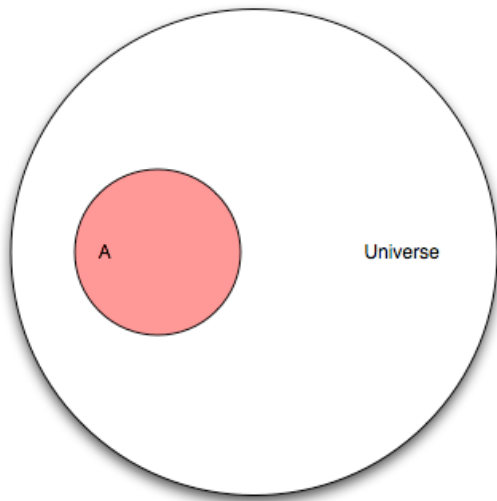
Naïve Bayes

Jay Urbain, PhD

Topics

- Probability and Bayes Theorem
- Naïve Bayes Theorem
- Extension to Bayesian Networks and Probabilistic Graphical Models

Probability

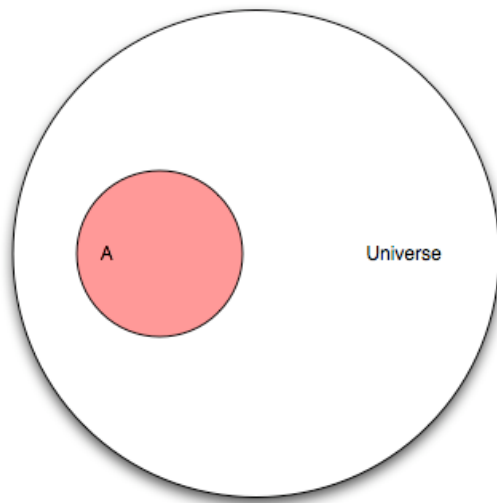


You're flipping a coin. This diagram represents the “universe” of all possible outcomes, also known as events. This universe is known as the sample space.

Q: What are the mutually exclusive events that make up the sample space for a coin flip?

A: Heads and tails

Probability



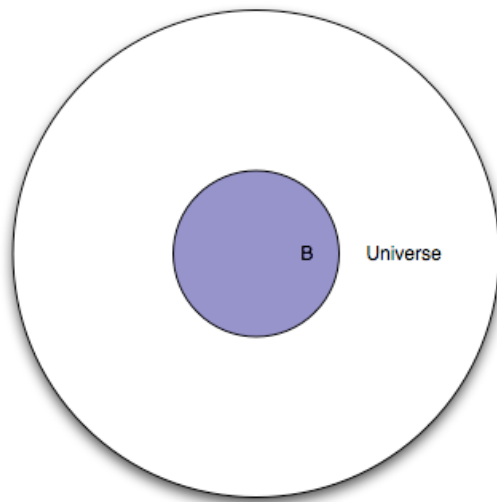
Let's now assume that our universe involves a research study on humans. Event "A" is people in that study who have cancer.

Q: If our study has 100 people and "A" has 25 people, what is the probability of A?

A: $P(A) = 25/100$

Q: What is the max probability of any event?

Probability

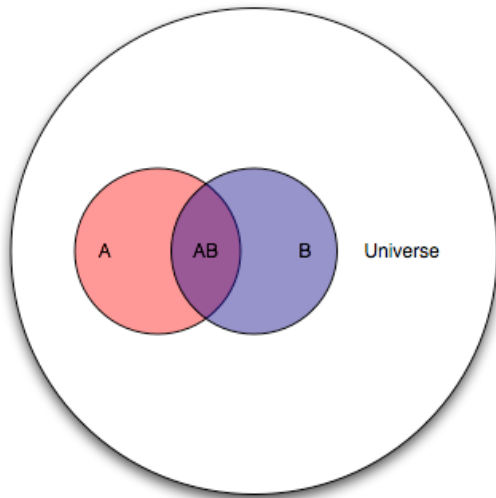


This represents the same set of people, except everyone in the study is given a test. Event “B” is everyone in the study for whom the test is positive.

Q: What portion of the diagram represents the subset of people with a negative test?

A: The white area between the smaller circle and the larger circle.

Probability



Because “A” and “B” are events from the same study, we can show them together.

Q: How would you describe the “cancer status” and “test status” of people in each area of the diagram?

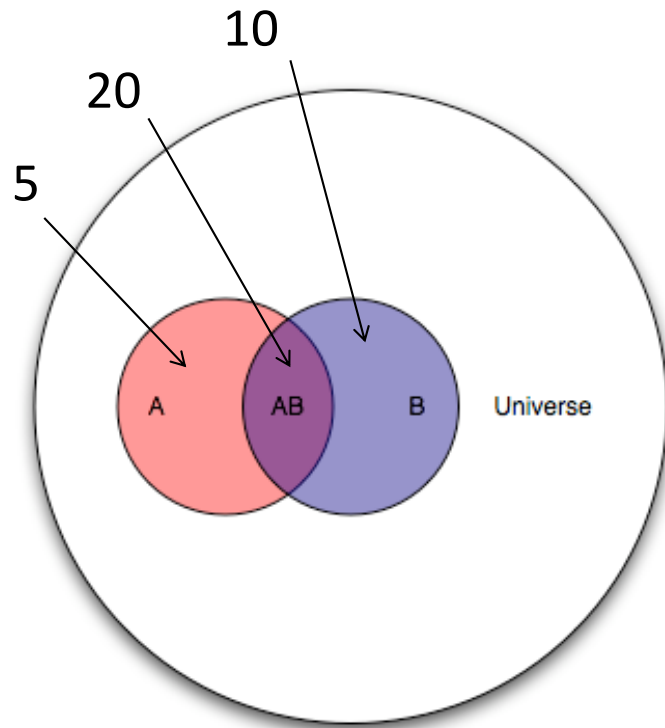
A: Pink: cancer, negative test

Purple: cancer, positive test

Blue: no cancer, positive test

White: no cancer, negative test

Probability

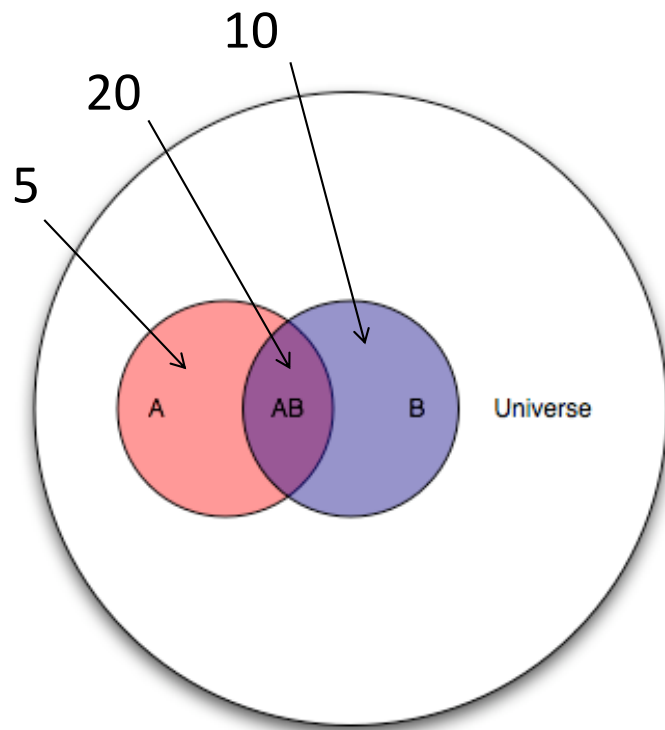


The purple section is known as the intersection of A and B, denoted as $P(A,B)$.

Thinking of this test as a classifier for predicting cancer, draw the confusion matrix.

n=100	Predicted: NO	Predicted: YES
Actual: NO	65	10
Actual: YES	5	20

Probability



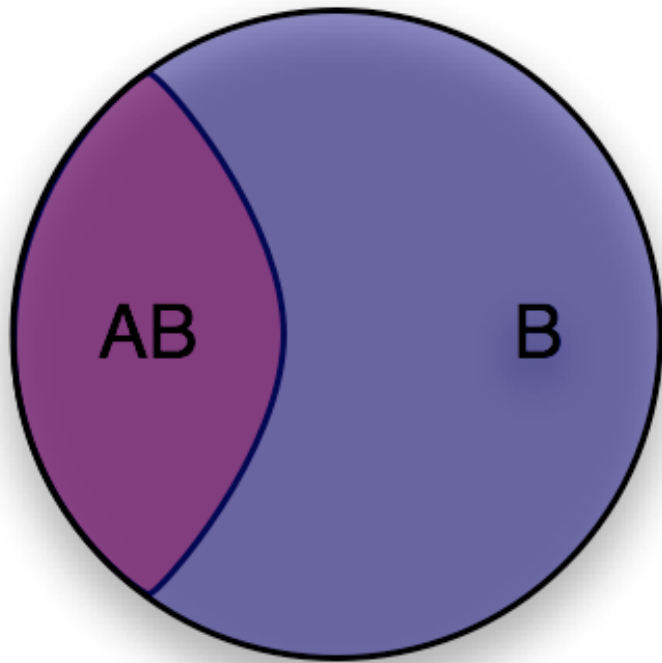
Q: Let's pick an arbitrary person from this study. If you were told their test result was positive, what is the probability they actually have cancer?

A: 20/30

This is the conditional probability of A given B, denoted as $P(A|B)$.

$$P(A|B) = P(AB) / P(B) = (20/100) / (30/100)$$

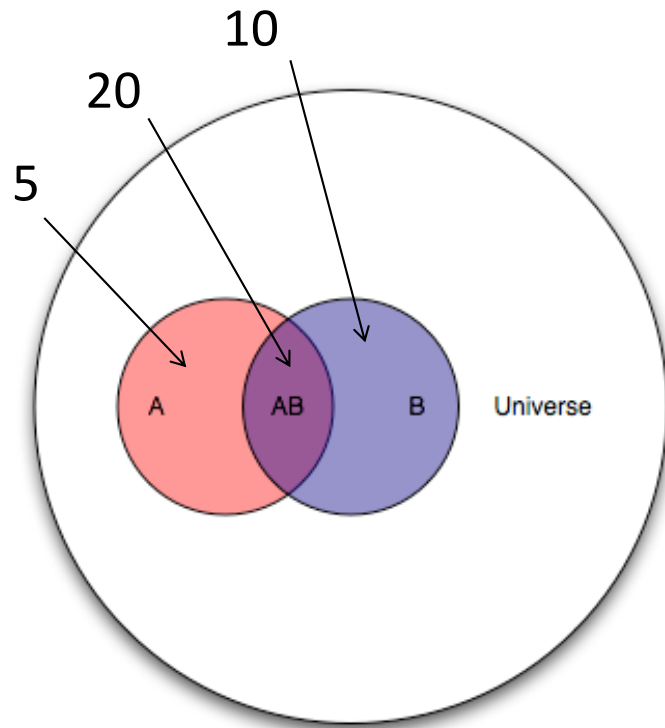
Probability



You can think of conditional probability as “changing the relevant universe.” $P(A|B)$ is a way of saying “Given that my entire universe is now B, what is the probability of A?”

This is also known as transforming the sample space.

Probability



Q: Let's pick another arbitrary person from this study. If you were told they have cancer, what is the probability they had a positive test result?

A: $P(B|A) = P(AB) / P(A) = 20/25$

Bayes Theorem

Deriving Bayes' theorem:

We know: $P(A|B) = P(AB) / P(B)$ and $P(B|A) = P(AB) / P(A)$

Thus: $P(AB) = P(A|B) * P(B) = P(B|A) * P(A)$

Rearrange to get Bayes' theorem: $P(A|B) = P(B|A) * P(A) / P(B)$

Bayesian Inference

Suppose we have a dataset with features x_1, \dots, x_n and a class label C . What can we say about classification using Bayes' theorem?

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

Bayes' theorem can help us to determine the probability of a record belonging to a class, *given* the data we observe.

Bayesian Inference

This term is the prior probability of c . It represents the probability of a record belonging to class c before the data is taken into account.

$$P(\text{class } C | \{x_i\}) = \frac{P(\{x_i\} | \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$



Prior

Bayesian Inference

This term is the likelihood function. It represents the joint probability of observing features $\{x_i\}$ given that that record belongs to class C .

$$P(\text{class } C | \{x_i\}) = \frac{P(\{x_i\} | \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$



Likelihood

Bayesian Inference

This term is the normalization constant. It doesn't depend on C , and is generally ignored.

$$P(\text{class } C | \{x_i\}) = \frac{P(\{x_i\} | \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$



Marginal probability of X

Bayesian Inference

This term is the posterior probability of C . It represents the probability of a record belonging to class C after the data is taken into account.

$$P(\text{class } C | \{x_i\}) = \frac{P(\{x_i\} | \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

Posterior



The idea of Bayesian inference, then, is to update our beliefs about the distribution of C using the data (“evidence”) at our disposal.

Bayes Classification

Q: What piece of the puzzle we've seen so far looks like it could intractably difficult in practice?

A: Estimating the full likelihood function.

$$P(\{x_i\} | C) = P(\{x_1, x_2, \dots, x_n\} | C)$$

Observing this exactly would require us to have enough data for every possible combination of features to make a reasonable estimate.

Bayes Classification

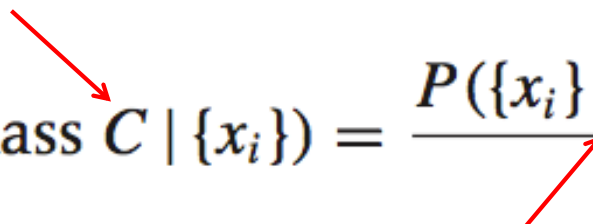
Q: So what can we do about it?

A: Make a simplifying assumption. In particular, we assume that the features x_i are conditionally independent from each other (given the class):

$$P(\{x_i\} | C) = P(\{x_1, x_2, \dots, x_n\} | C) \approx P(x_1 | C) * P(x_2 | C) * \dots * P(x_n | C)$$

This “naïve” assumption simplifies the likelihood function to make it tractable.

Bayes Classification


$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

In summary, the training phase of the model involves computing the likelihood function, which is the conditional probability of each feature given each class.

The prediction phase of the model involves computing the posterior probability of each class given the observed features, and choosing the class with the highest probability.

Bayesian Network