Naive Bayes Classification

GA DAT3

Agenda

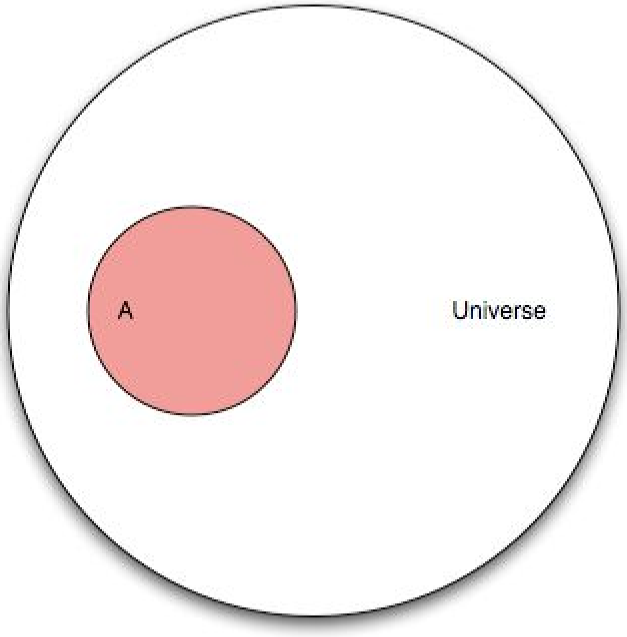
1. PROBABILITY AND BAYES’ THEOREM
2. NAÏVE BAYES CLASSIFICATION

Probability and Bayes Theorem

Probability

Let’s pretend you are 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?

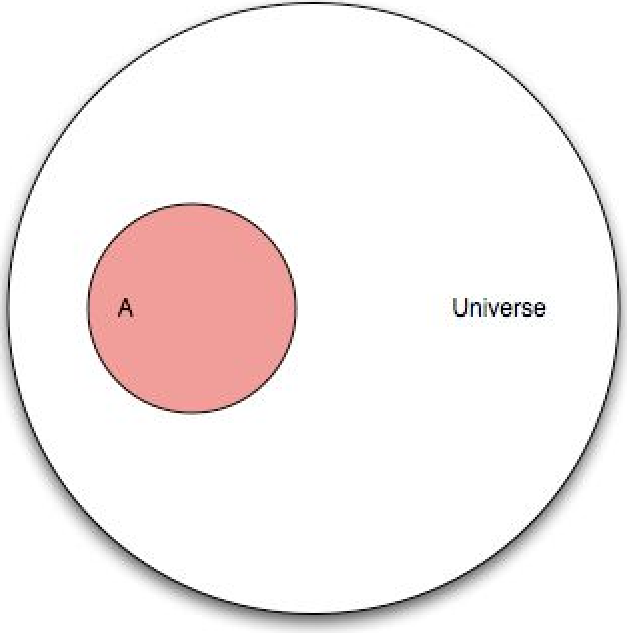


Probability

Let’s pretend you are 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

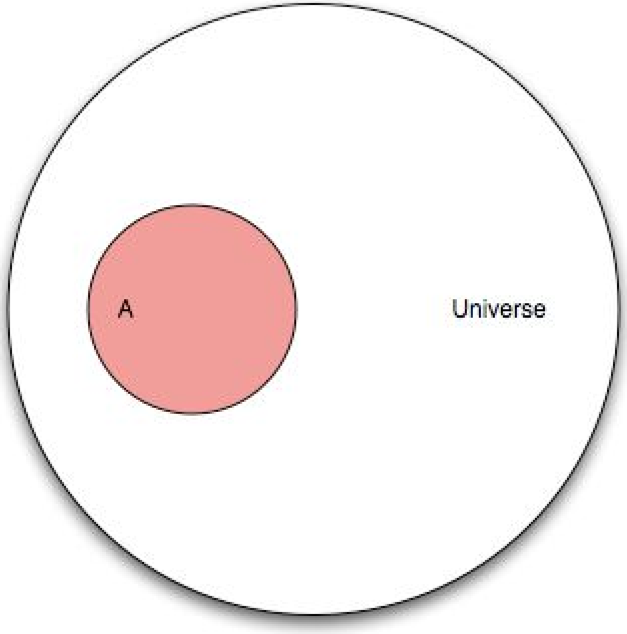


Probability

Let’s now pretend 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?

Q: What is the max probability of any event?



Probability

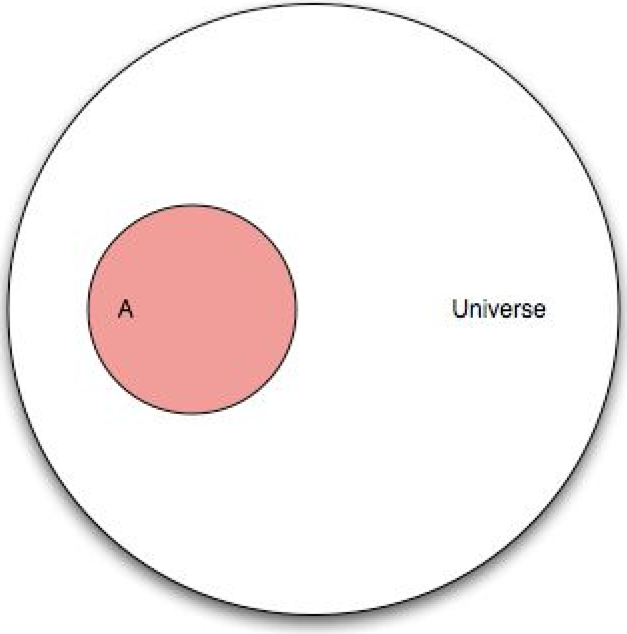
Let’s now pretend that our universe involves a research study on humans. Event “A” is people in that study who have cancer.

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?

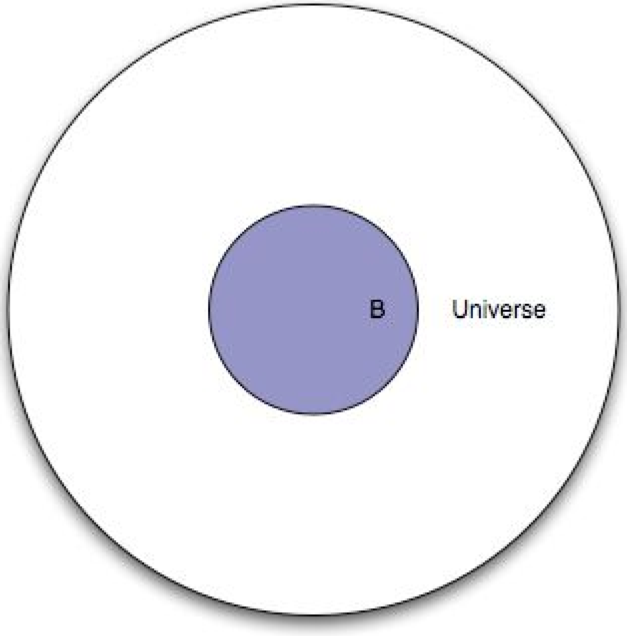
A: 1



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.

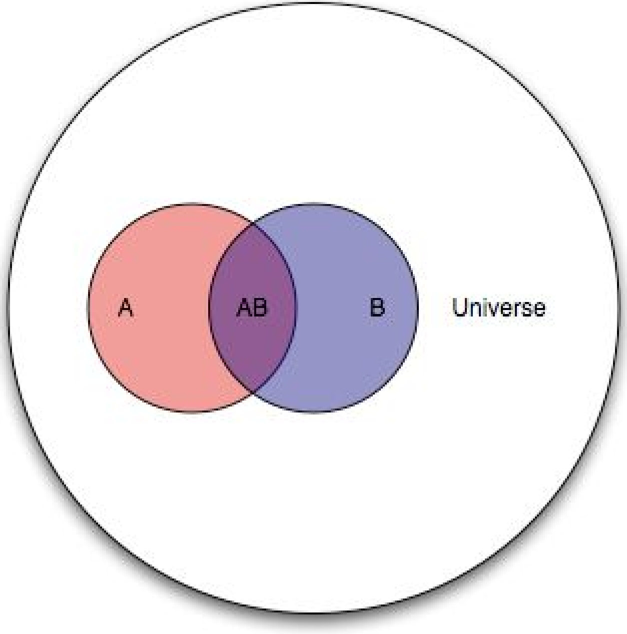
Q: What portion of the diagram represents the subset of people with a negative test?
A: The white area between the smaller circle and the larger circle.



Probability

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

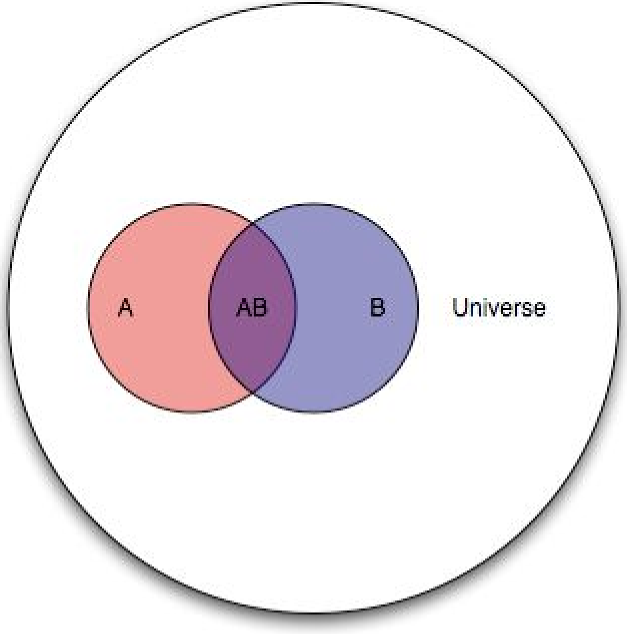
Q: How would you describe the “cancer status” and “test status” of people in each area of the diagram?



Probability

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

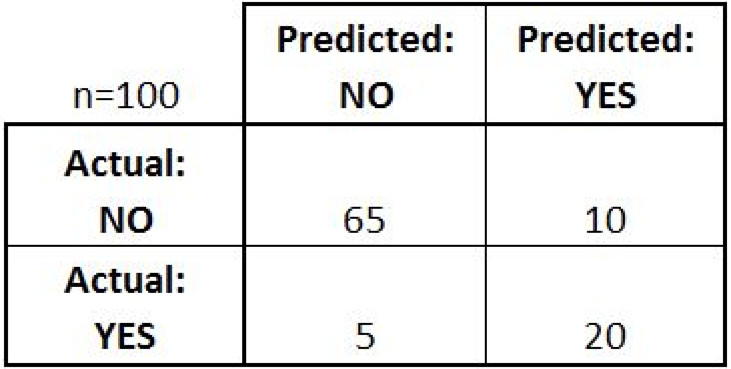
Q: How would you describe the “cancer status” and “test status” of people in each area of the diagram?
A: Pink: cancer, negative test
Purple: cancer, positive test
Blue: no cancer, positive test
White: no cancer, negative test



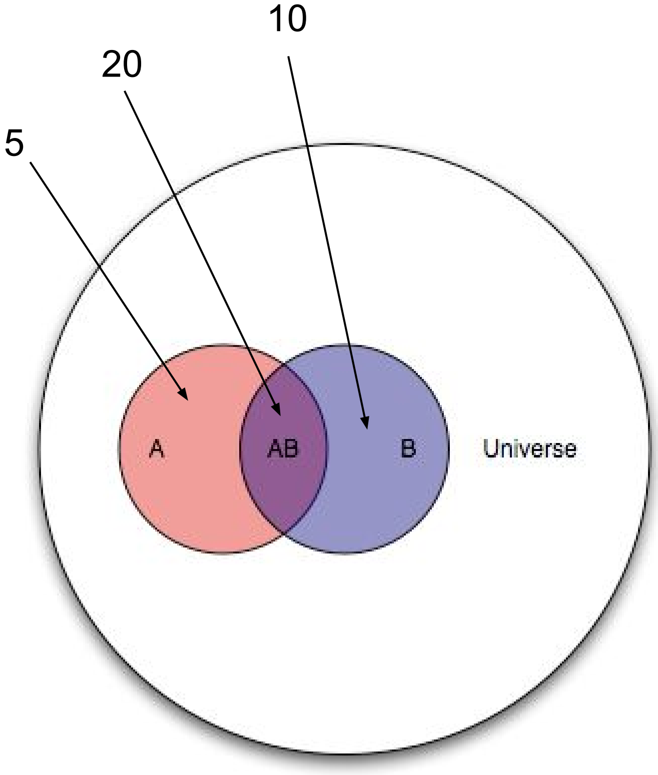
Probability

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

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

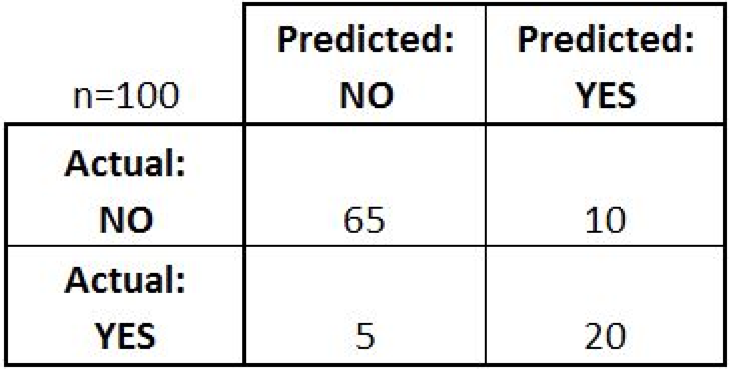


n=100	Predicted: NO	Predicted: YES
Actual: NO	65	10
Actual: YES	5	20

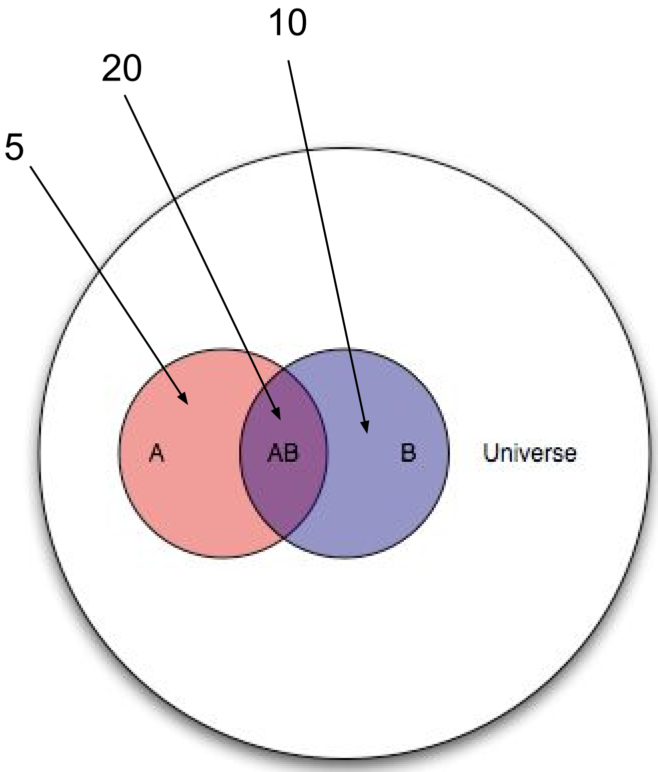


Probability

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?



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



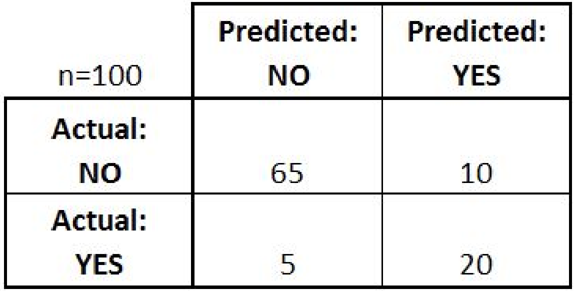
Probability

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?

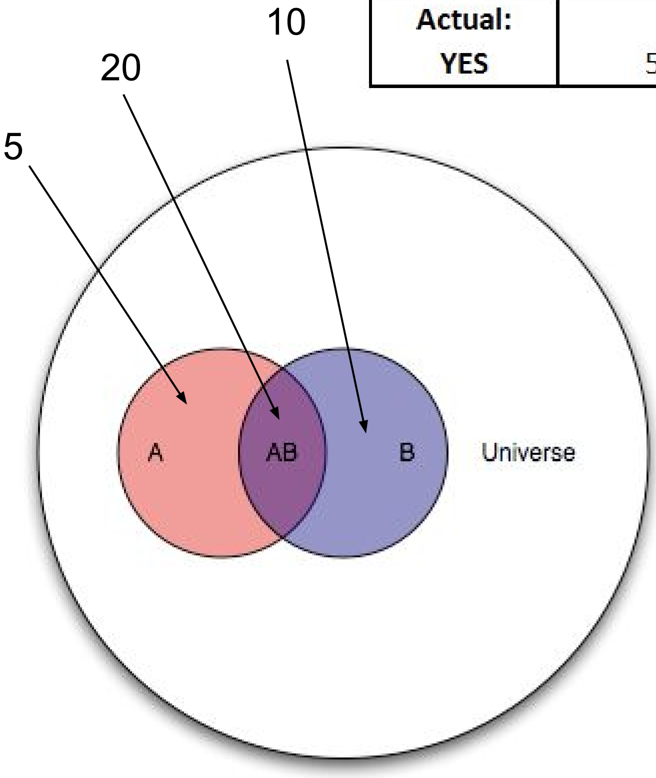
A: 20/30

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)



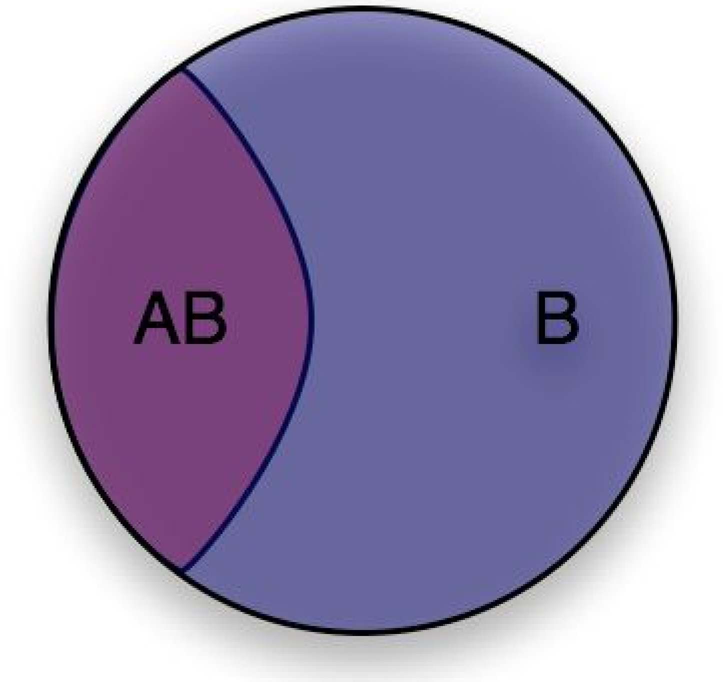
n=100	Predicted: NO	Predicted: YES
Actual: NO	65	10
Actual: YES	5	20



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?"

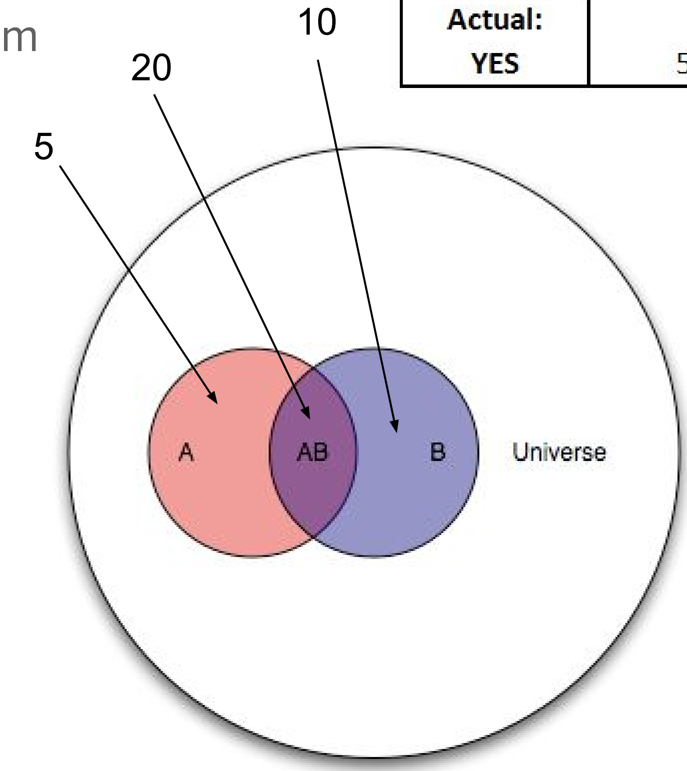
This is also known as transforming the sample space.



Probability

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

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

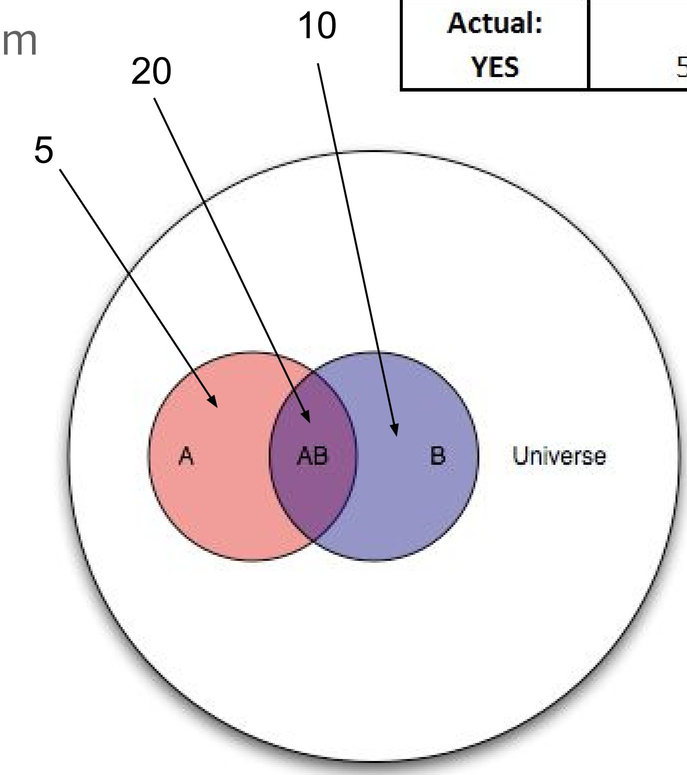


Probability

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

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

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



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)

Bayes’ Theorem Example

Suppose you might have a rare life-threatening disease and so you get tested.

The disease’s test is 99% sensitive and 99% specific (if you have it, the test is correct 99% of the time and same if you don’t have it). This disease occurs in 1 in every 10,000 people.

Q. Your test is positive. What is the probability that you have the disease?

Bayes' Theorem Example

Q. Your test is positive. What is the probability that you have the disease?

A. 1%

Let A be the event that you have the disease

B be the event that your test was positive

$P(B|A) = .99$ (**sensitivity**) $P(B|\text{ not }A) = .01$(1 - sensitivity) this is our **false positive**

Bayes' Theorem Example

Q. Your test is positive. What is the probability that you have the disease?

A. 1%

Let A be the event that you have the disease

B be the event that your test was positive

$P(B|A) = .99$ (**sensitivity**) $P(B| \text{ not } A) = .01(1 - \text{sensitivity})$ this is our **false positive**

$P(B) = P(\text{the test was positive}) = P(B \mid A) * P(A)$ OR $P(B \mid \text{not } A) * P(\text{not } A)$

$P(B) = .99 * .0001 + .01 * .9999 = .010098$

$$P(A|B) = P(B|A)P(A) / P(B)$$

$$= .99 * .0001 / .010098 = 0.00980$$

Naive Bayes’ Classification

Bayesian Inference

Suppose we have a dataset with features x1, …, xn 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

source: Data Analysis with Open Source Tools, by Philipp K. Janert. O'Reilly Media, 2011.

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

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

Bayesian Inference

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

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

Bayesian Inference

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

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

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.

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

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

Naive 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.

$$\mathbf{P(\{xi\}|C) = P(\{x1, x2, \ldots, xn\})|C)}$$

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

Naive Bayes' Classification

Q: So what can we do about it?

hence, 'naive'

A: Make a simplifying assumption. In particular, we assume that the features xi are conditionally independent from each other:

$$P(\{xi\}|C) = P(\{x1, x2, ..., xn\}|C) \approx P(x1|C) * P(x2|C) * ... * P(xn|C)$$

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

Bayesian Inference

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

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.

Q??