GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

SESSION 13: NAIVE BAYESIAN CLASSIFICATION

Rob Hall

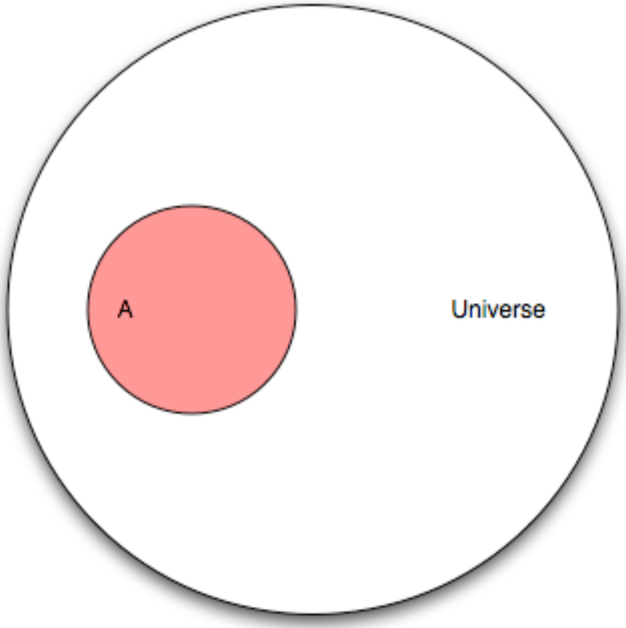
DAT13 SF // April 20, 2015

AGENDA

I. PROBABILITY & BAYES’ THEOREM
II. NAÏVE BAYESIAN CLASSIFICATION

EXERCISES:
III. LAB: NAIVE BAYES CLASSIFICATION IN PYTHON

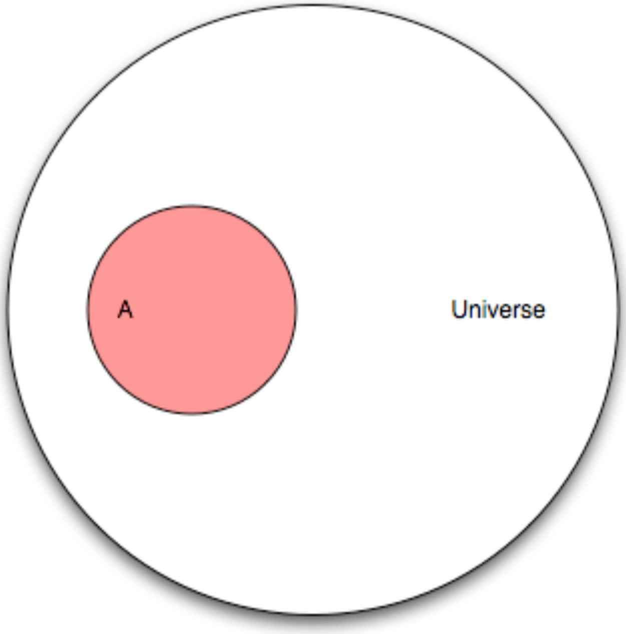
I. PROBABILITY AND BAYES' THEOREM



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

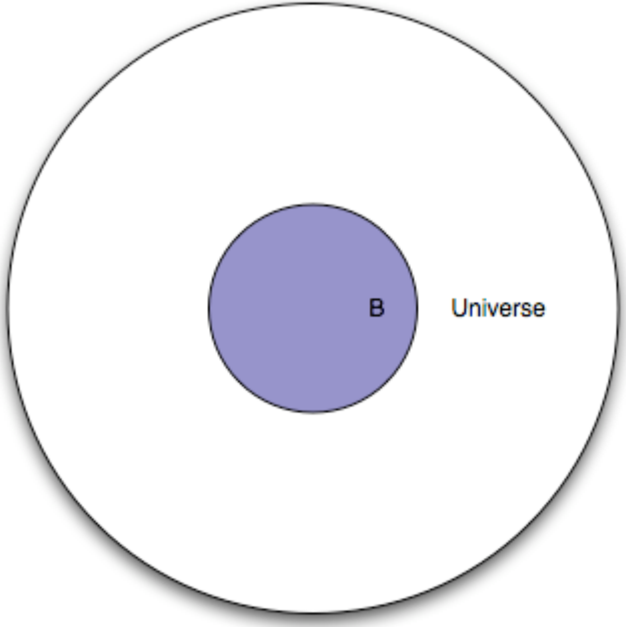
PROBABILITY



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

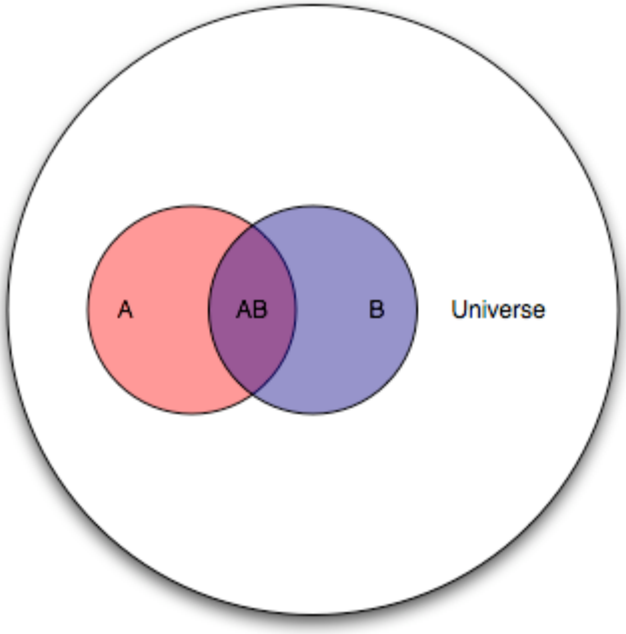
Q: If our study has 100 people and "A" has 25 people, what is the **probability** *of A?*
A: P(A) = 25/100

Q: What is the max probability of any event?
A: 1



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.



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

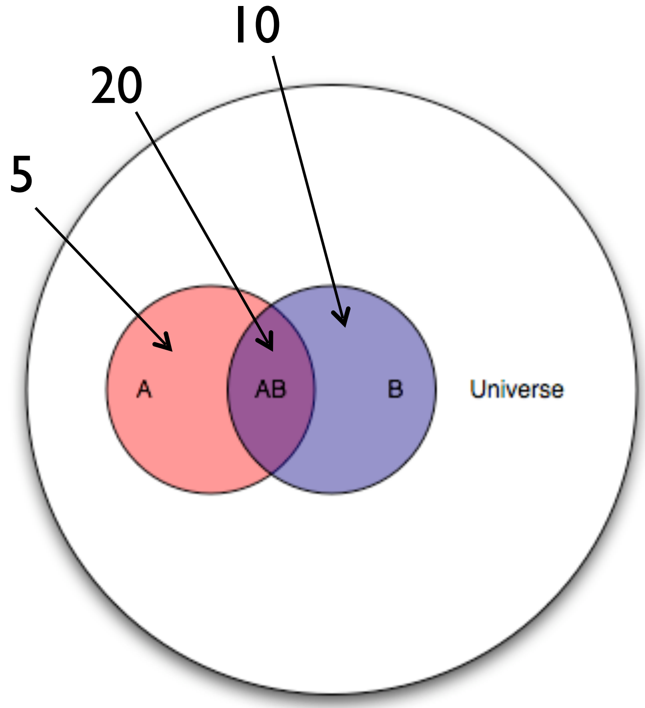
Q: How would you describe the "cancer status" and "test status" of people in each area of the diagram?

A: Pink: cancer, negative test

Purple: cancer, positive test

Blue: no cancer, positive test

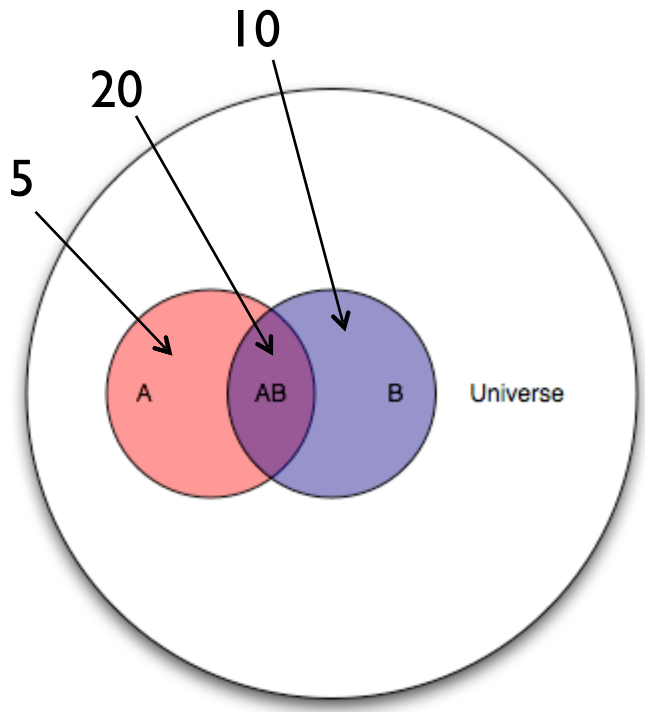
White: no cancer, negative test



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

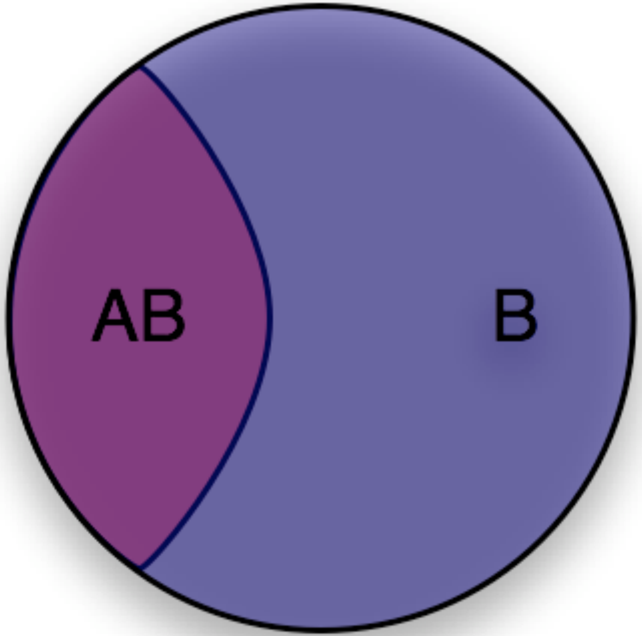


Q: Let's pick an arbitrary person from this study. If you were told their test result was positive, what is the probability they actually have cancer?

A: 20/30

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

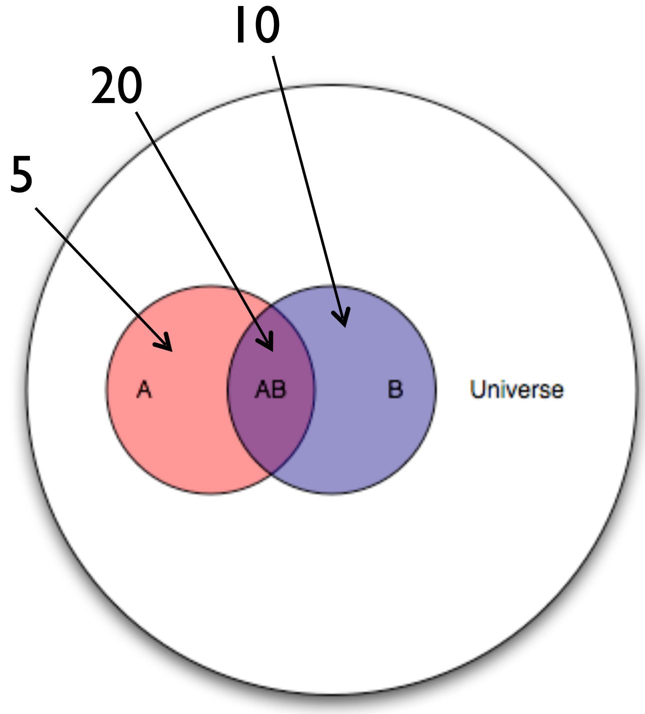
$P(A|B) = P(AB) / P(B) = (20/100) / (30/100)$



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

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

Briefly, the two interpretations can be described as follows:

Briefly, the two interpretations can be described as follows:

The frequentist interpretation regards an event's probability as its limiting frequency across a very large number of trials.

INTERPRETATIONS OF PROBABILITY

Briefly, the two interpretations can be described as follows:

The frequentist interpretation regards an event's probability as its limiting frequency across a very large number of trials.

The Bayesian interpretation regards an event's probability as a "degree of belief," which can apply even to events that have not yet occurred.

II. NAÏVE BAYESIAN CLASSIFICATION

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

Suppose we have a dataset with features $x_1, \ldots, x_n$ 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.

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

SOME TERMINOLOGY

Each term in this relationship has a name, and each plays a distinct role in any Bayesian calculation (including ours).

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

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

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

The ***likelihood*** *of seeing that evidence if your hypothesis is correct.*

THE LIKELIHOOD FUNCTION

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

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

We can observe the value of the likelihood function from the training data.

THE PRIOR

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.*

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

The ***prior***

THE PRIOR

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.*

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

The value of the prior is also observed from the data.

This term is the **normalizing constant.** *It doesn't depend on* C*, and is generally ignored until the end of the computation.*

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

The probability of the data under any hypothesis.

THE NORMALIZATION CONSTANT

This term is the **normalizing constant.** *It doesn't depend on* C*, and is generally ignored until the end of the computation.*

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

The normalizing constant doesn't tell us much.

THE POSTERIOR

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.*

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

In other words, the probability of the hypothesis ***after*** *seeing the evidence.*

THE POSTERIOR

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 goal of any Bayesian computation is to find ("learn") the posterior distribution of a particular variable.

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

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

Then we can use the posterior for prediction.

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

Remember the likelihood function?

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

Remember the likelihood function?

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

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

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.

Q: So what can we do about it?

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:

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:

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

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:

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

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

NAÏVE 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.*

III. LAB: NAIVE BAYESIAN CLASSIFICATION