GENERAL ASSEMBLY

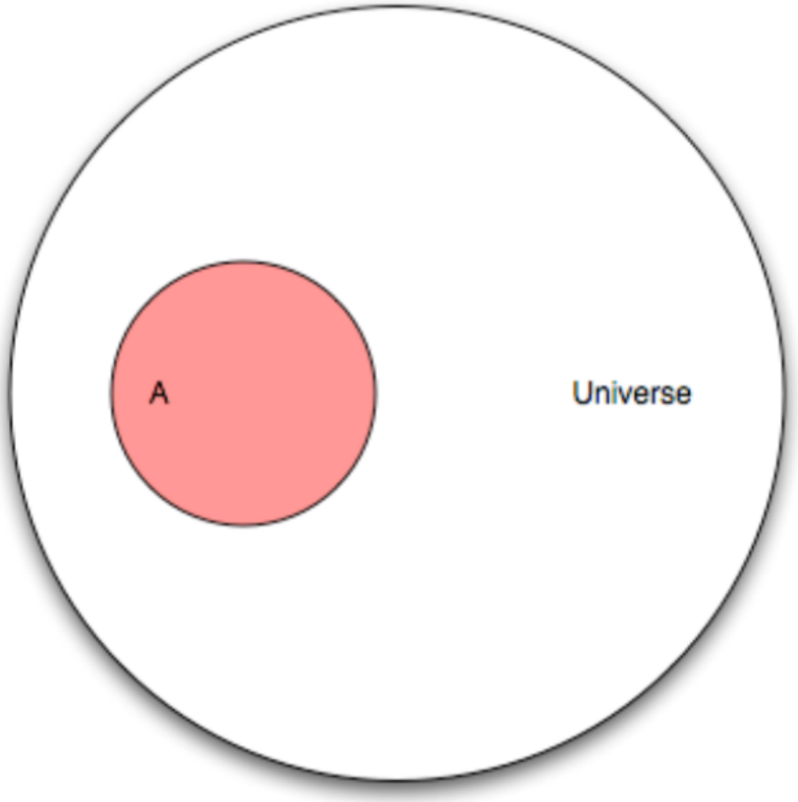
BAYES WRAP-UP AND NAÏVE BAYES

Joseph Nelson, Data Science Immersive

AGENDA

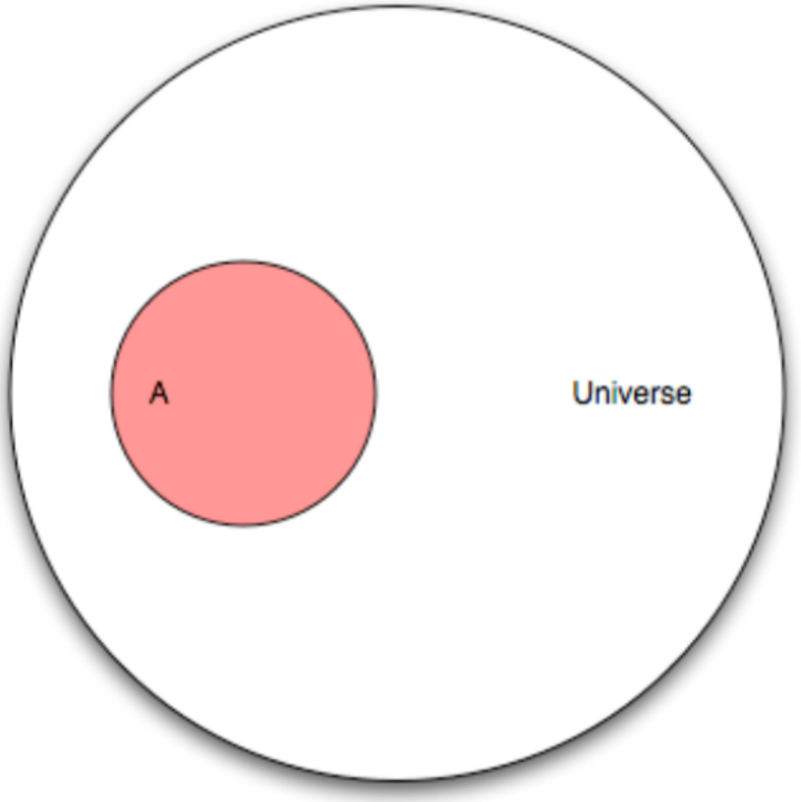
- Review Bayes theorem
- Formalize Naïve Bayes
- Applying Naïve Bayes in Sklearn

BAYES REVIEW



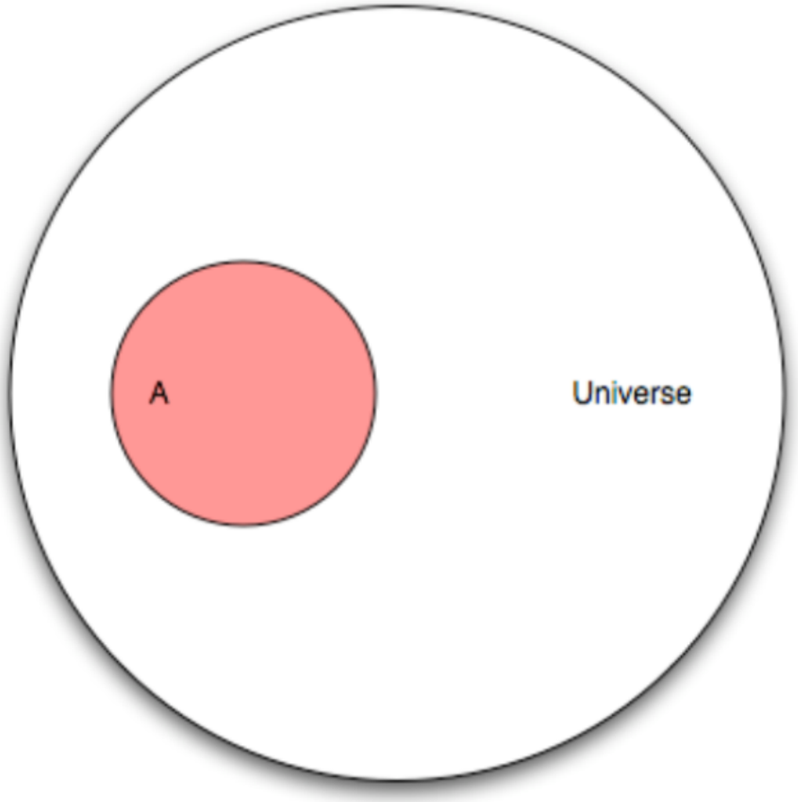
- You flip a coin
- This diagram represents the universe of all possible outcomes (aka events). The universe is known as our sample space.
- What are the mutually exclusive events that make up the sample space for a coin flip? (Yes, this is easy)

BAYES REVIEW



- You flip a coin
- This diagram represents the universe of all possible outcomes (aka events). The universe is known as our sample space.
- What are the mutually exclusive events that make up the sample space for a coin flip? (Yes, this is easy)
- Heads and tails

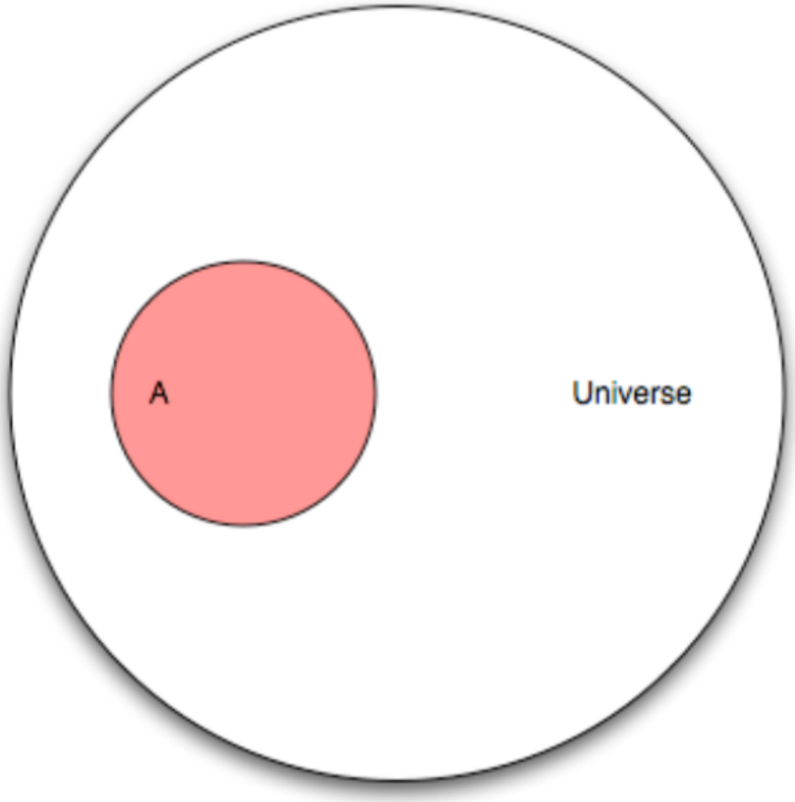
BAYES REVIEW



- Alright, our events are a bit more serious: our universe is a research study on humans. Event A is people in the study who have cancer.

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

BAYES REVIEW

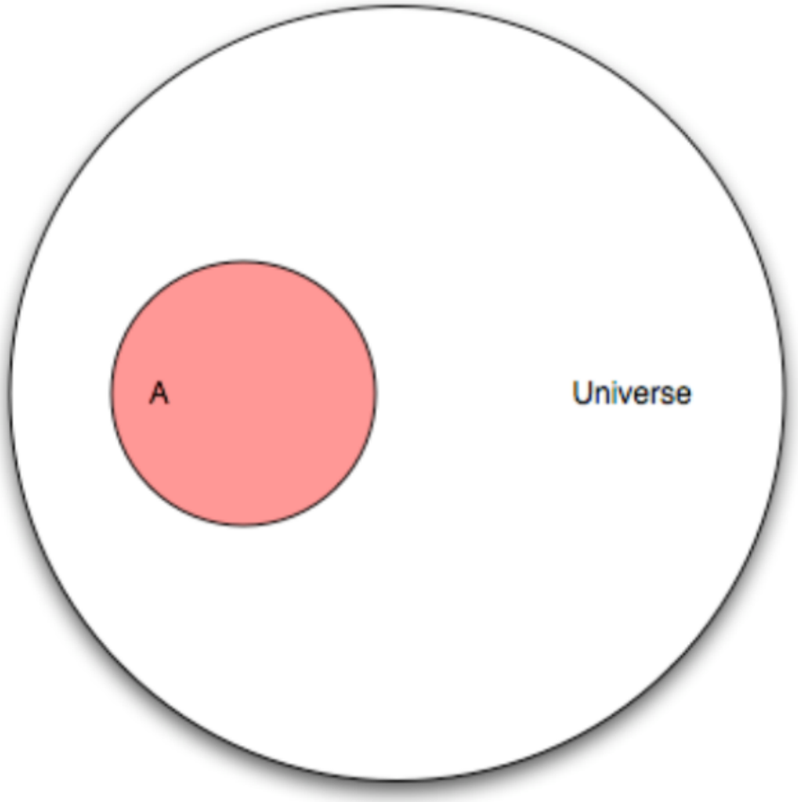


- Alright, our events are a bit more serious: our universe is a research study on humans. Event A is people in the study who have cancer.

- If our study has 100 people and A has 25 people, what is the probability of A?
- $P(A) = 25/100 = 0.25$

- What is the max probability of any event?

BAYES REVIEW

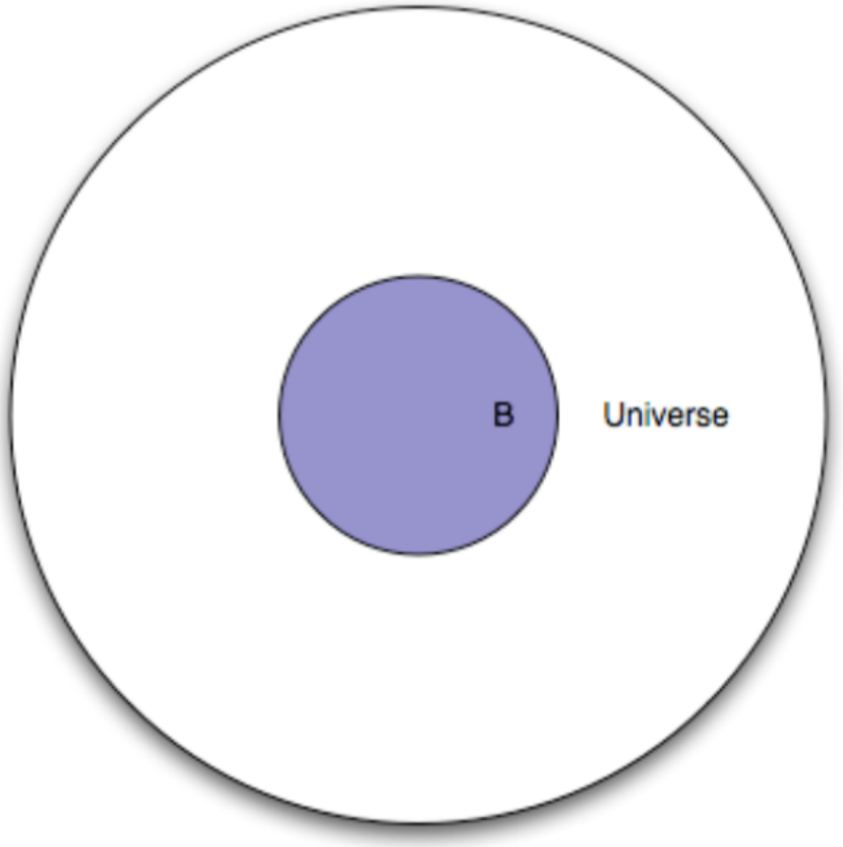


- Alright, our events are a bit more serious: our universe is a research study on humans. Event A is people in the study who have cancer.

- If our study has 100 people and A has 25 people, what is the probability of A?
- $P(A) = 25/100 = 0.25$

- What is the max probability of any event?
- 1

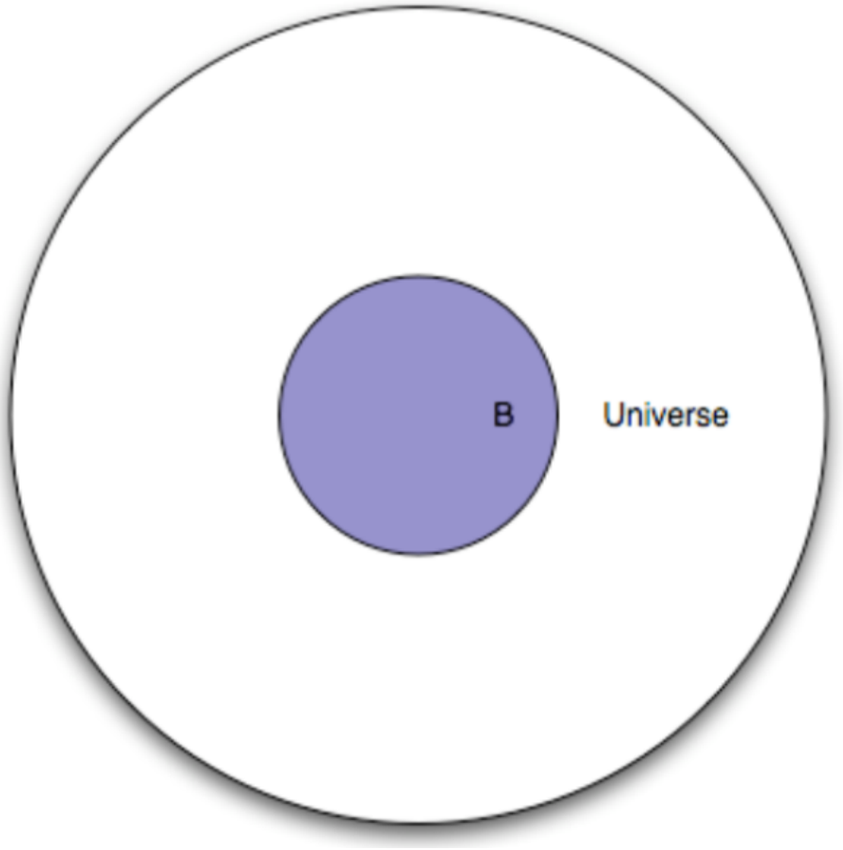
BAYES REVIEW



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

- Visually, what portion of this diagram represents people with a negative test?

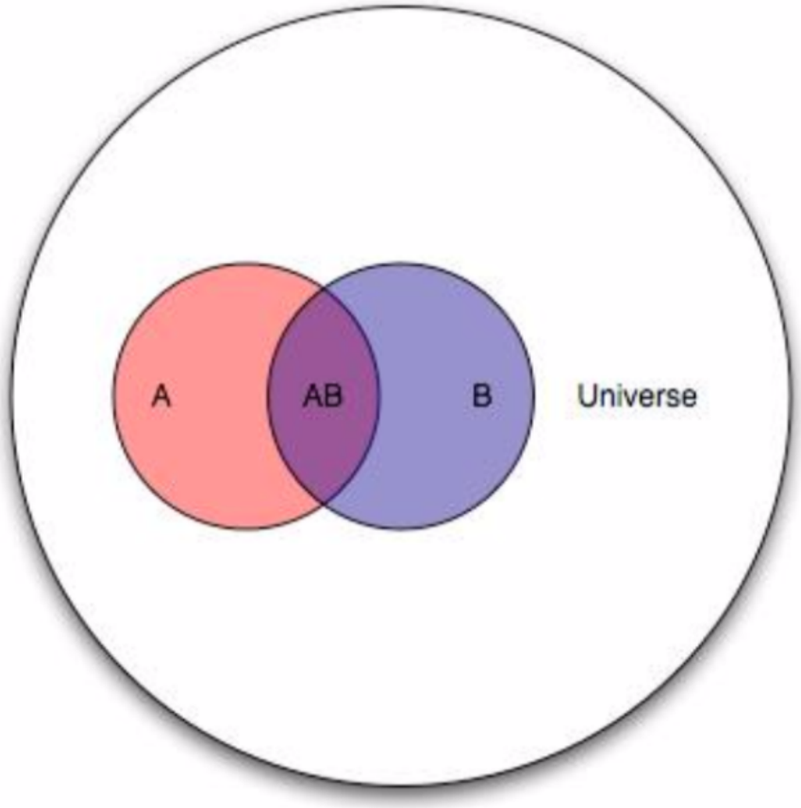
BAYES REVIEW



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

- Visually, what portion of this diagram represents people with a negative test?
- The white area between the smaller and larger circle

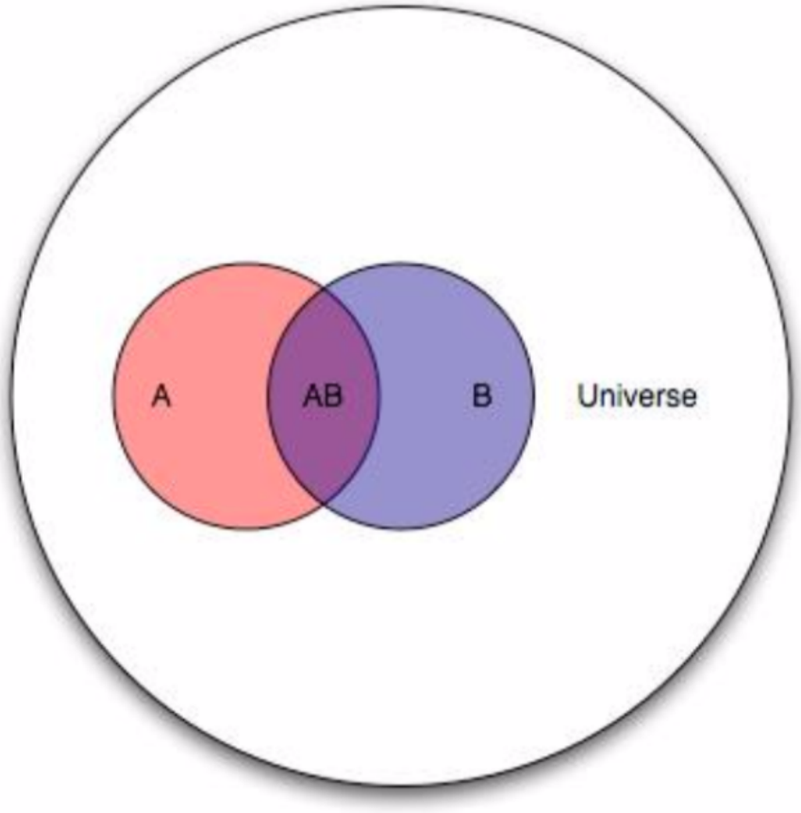
BAYES REVIEW



- 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 portion of the diagram (by color)?
- Pink:
- Purple:
- Blue:
- White:

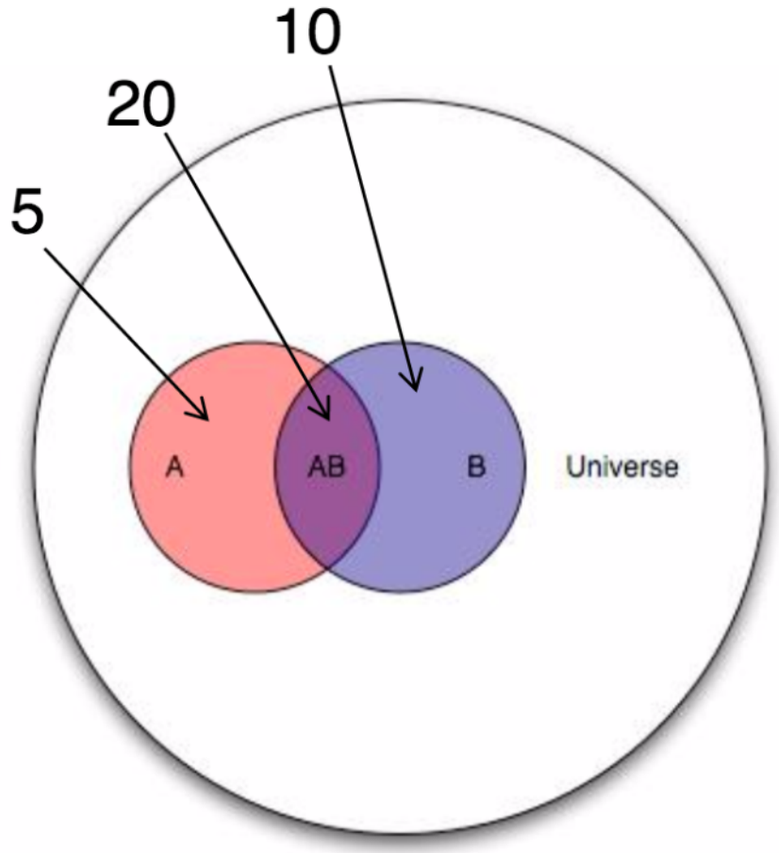
BAYES REVIEW



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

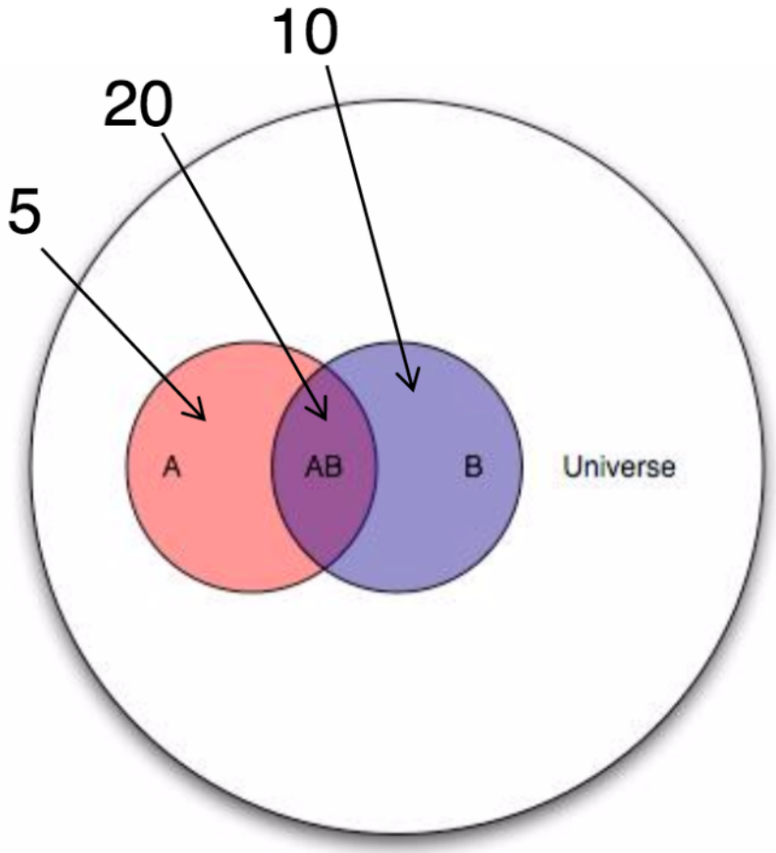
- Q: How would you describe the “cancer status” and “test status” of people in each portion of the diagram (by color)?
- Pink: cancer, negative test
- Purple: cancer, positive test
- Blue: no cancer, positive test
- White: no cancer, negative test

BAYES REVIEW



- The purple section is the intersection of A and B – P(AB)
- Thinking of this as a classifier result, draw a confusion matrix
- (I did say review, didn't I?!)

BAYES REVIEW

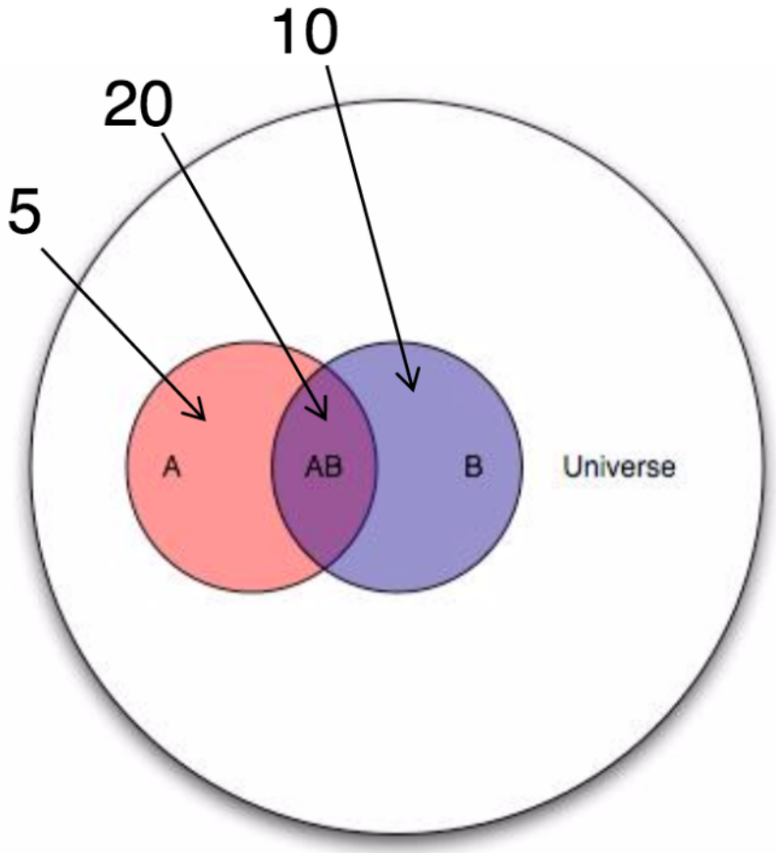


- The purple section is the intersection of A and B – P(AB)

- Thinking of this as a classifier result, draw a confusion matrix
- (I did say review, didn't I?!)

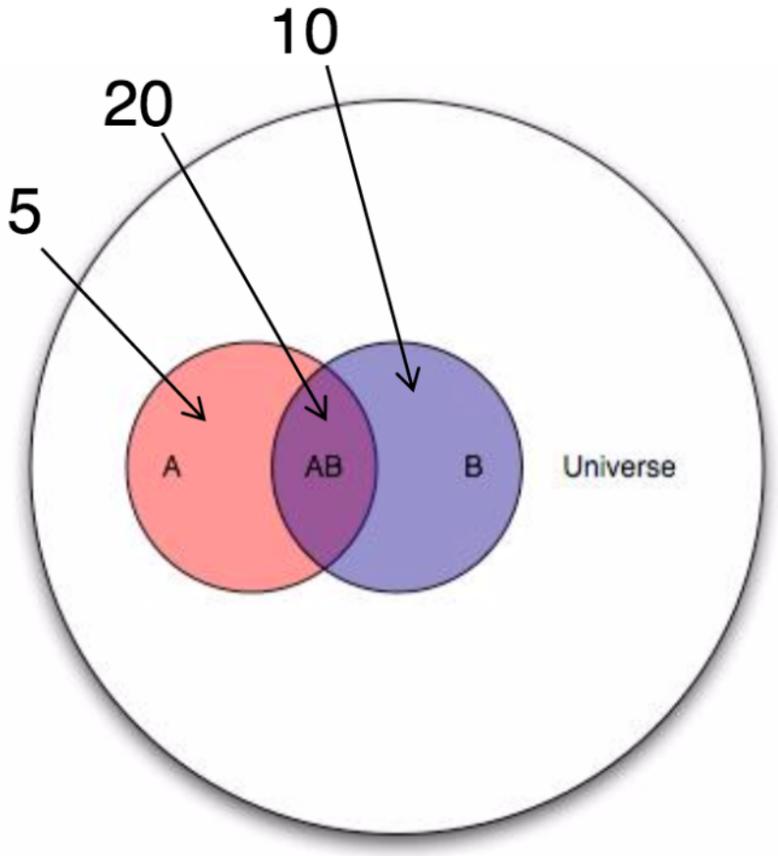
n=100	**Predicted: NO**	**Predicted: YES**
Actual: NO	65	10
Actual: YES	5	20

BAYES REVIEW



- Let's pick an arbitrary person from the study. If we knew their test result was positive, what is the probability they actually have cancer?

BAYES REVIEW

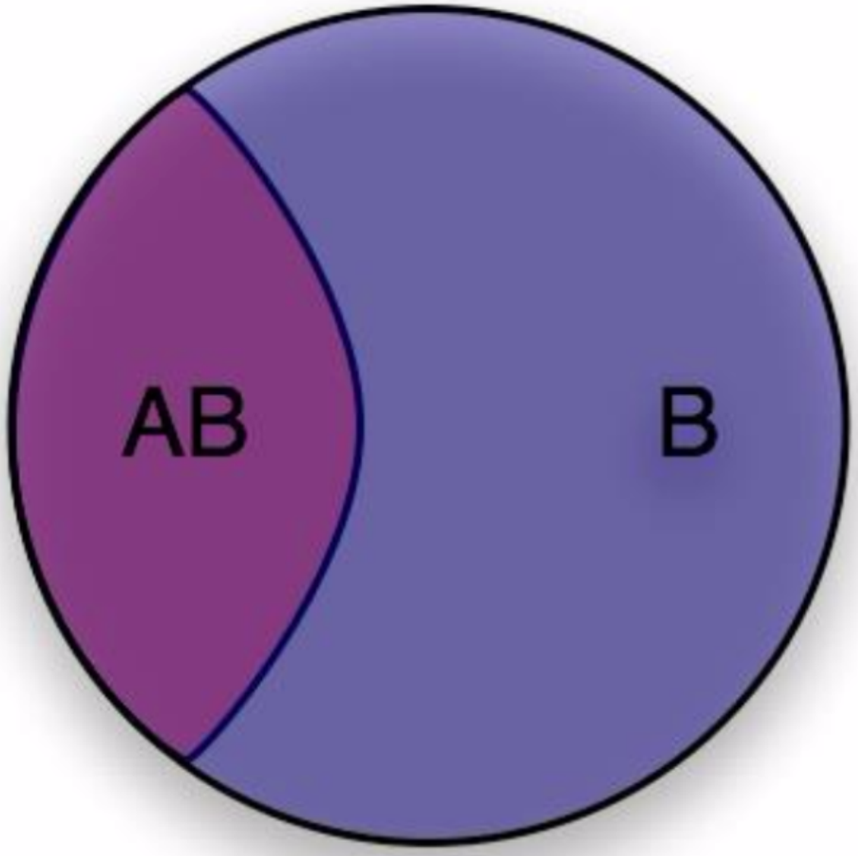


- Let's pick an arbitrary person from the study. If we knew their test result was positive, what is the probability they actually have cancer?
- 20/30

- This is the conditional probability of A given B, $P(A|B)$

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

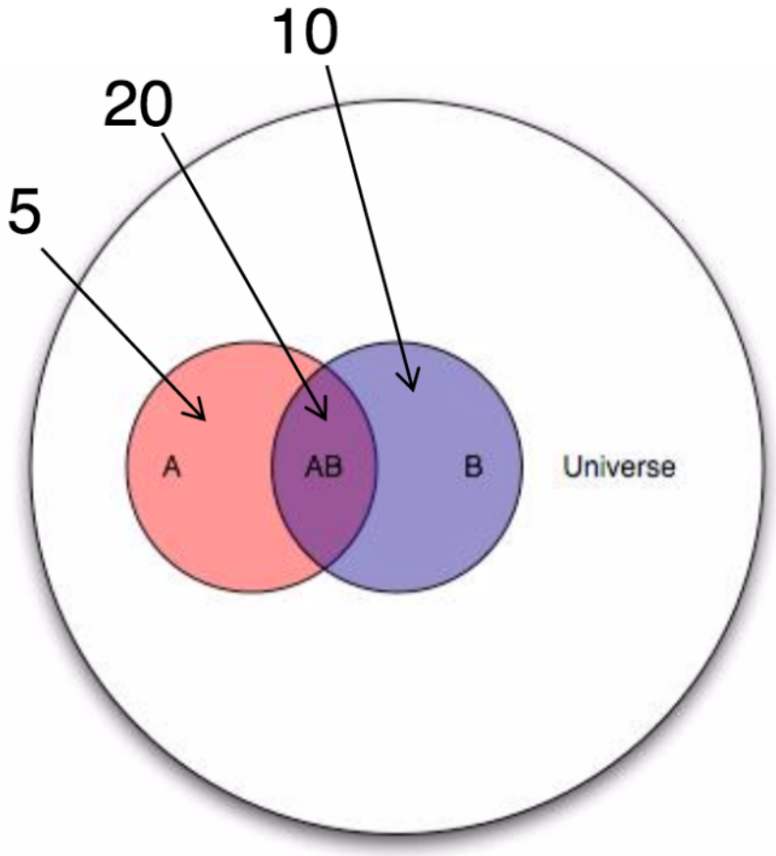
BAYES REVIEW



- Thus, we think of the conditional probability of changing the relevant universe from which we're interested. P(A|B) is a way of saying "given that my entire universe is now B, what is the probability of A?"

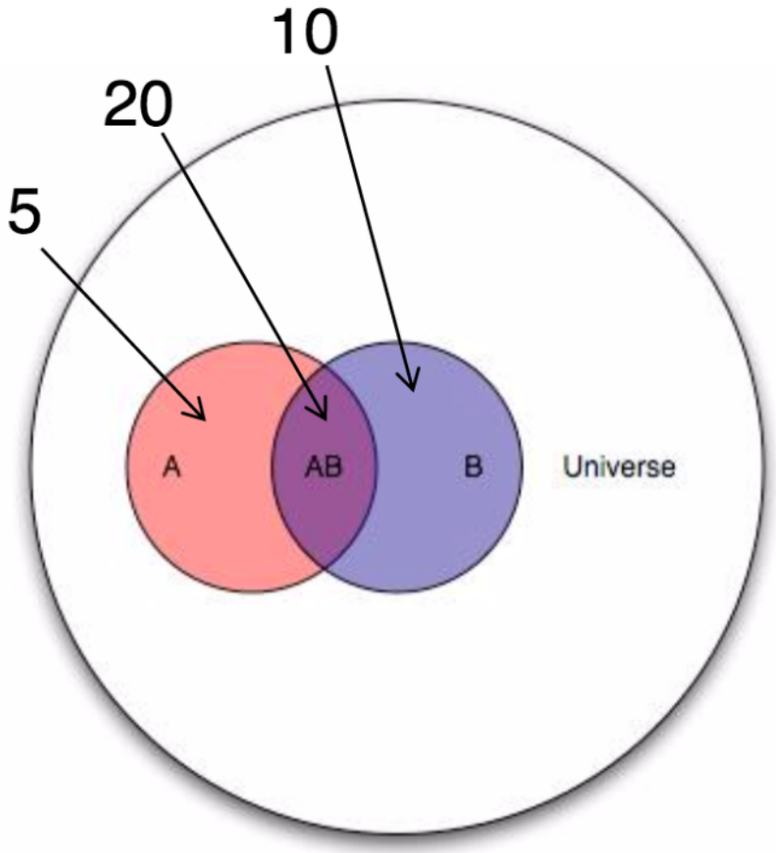
- This is **transforming the sample space**

BAYES REVIEW



- We’re picking another arbitrary person from our study. If you were told they have cancer, what is the probability that they had a positive test result?

BAYES REVIEW



- We're picking another arbitrary person from our study. If you were told they have cancer, what is the probability that they had a positive test result?
- $P(B|A) = P(AB)/P(A) = 20/25$

BAYES REVIEW

- We know $P(A|B) = P(AB)/P(B)$ and $P(B|A) = P(AB)/P(A)$
- Thus $P(AB) = P(A|B) * P(B) = P(B|A) * P(A)$
- Also can be written:
- $P(A|B) = P(B|A)*P(A)/P(B)$

- We've found Thomas.

COMPREHENSION CHECK

- 1% of women at age forty who participate in a routine screening have breast cancer. 80% of women with breast cancer will get positive mammograms. 9.6% of women without breast cancer will also get positive mammograms.

- A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer?

BAYES REVIEW

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

- A is “has cancer” and B is “positive test.” What is $P(A|B)$?

- $P(B|A) = 0.80$
- $P(A) = 0.01$
- $P(B) = 0.103$
- $P(B|A) = 0.90 * 0.01 / 0.103 = 7.8\%$

NAÏVE BAYES

- Now that we have a nice review of Bayes theorem

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

- We can again put this in the context of our models and data

$$P(model \mid data) = \frac{P(data \mid model)P(model)}{P(data)}$$

NAÏVE BAYES

- I'm going to introduce Naïve Bayes in an applied context

- Let's say we are trying to predict spam emails

$$P(S \mid W) = \frac{P(W \mid S)P(S)}{P(W)} = \frac{P(W \mid S)P(S)}{P(W \mid S)P(S) + P(W \mid H)P(H)}$$

- S: Spam
- W: Words
- H: Ham

- What does P(S|W) refer to?

NAÏVE BAYES

- I'm going to introduce Naïve Bayes in an applied context

- Let's say we are trying to predict spam emails

$$P(S \mid W) = \frac{P(W \mid S)P(S)}{P(W)} = \frac{P(W \mid S)P(S)}{P(W \mid S)P(S) + P(W \mid H)P(H)}$$

- S: Spam
- W: Words
- H: Ham

- What does P(S|W) refer to?
- Probability that a message is spam given word W occurs in it

NAÏVE BAYES

- I'm going to introduce Naïve Bayes in an applied context

- Let's say we are trying to predict spam emails

$$P(S \mid W) = \frac{P(W \mid S)P(S)}{P(W)} = \frac{P(W \mid S)P(S)}{P(W \mid S)P(S) + P(W \mid H)P(H)}$$

- P(S|W) : Probability that a message is spam given word W occurs in it
- P(W|S):
- P(W|H):

NAÏVE BAYES

- I'm going to introduce Naïve Bayes in an applied context

- Let's say we are trying to predict spam emails

$$P(S \mid W) = \frac{P(W \mid S)P(S)}{P(W)} = \frac{P(W \mid S)P(S)}{P(W \mid S)P(S) + P(W \mid H)P(H)}$$

- P(S|W) : Probability that a message is spam given word W occurs in it
- P(W|S): Probability that a word W occurs in a spam message
- P(W|H): Probability that a word W occurs in a ham message

NAÏVE BAYES

- Let's start by making some simplifying assumptions. Let's assume there's an equal chance a given message is spam / not spam

$$P(S \mid W) = \frac{P(W \mid S)}{P(W \mid S) + P(W \mid H)}$$

- What is a problem you recognize here? (Think about what W is)

NAÏVE BAYES

- Let's start by making some simplifying assumptions. Let's assume there's an equal chance a given message is spam / not spam

$$P(S \mid W) = \frac{P(W \mid S)}{P(W \mid S) + P(W \mid H)}$$

- What is a problem you recognize here? (Think about what W is)
- We have *multiple* words in a single message to consider. That complicates things quite a bit: we would have to consider the feature vector $X_1, X_2, \ldots X_n$

NAÏVE BAYES

- Let's start by making some simplifying assumptions. Let's assume there's an equal chance a given message is spam / not spam

$$P(\ S \mid W\ ) = \frac{P(\ W \mid S\ )}{P(\ W \mid S) + P(\ W \mid H)}$$

- What is a problem you recognize here? (Think about what W is)
- We have *multiple* words in a single message to consider. That complicates things quite a bit: we would have to consider the feature vector $X_1, X_2, \ldots X_n$

$$P(\ S \mid X_1,\ X_2,\ ...,\ X_n\ ) = \frac{P(\ X_1,\ X_2,\ ...,\ X_n \mid S\ )}{P(\ X_1,\ X_2,\ ...,\ X_n \mid S) + P(\ X_1,\ X_2,\ ...,\ X_n \mid H)}$$

NAÏVE BAYES

- We have *multiple* words in a single message to consider. That complicates things quite a bit: we would have to consider the feature vector $X_1, X_2, \ldots X_n$

$$P(S \mid X_1, X_2, \ldots, X_n) = \frac{P(X_1, X_2, \ldots, X_n \mid S)}{P(X_1, X_2, \ldots, X_n \mid S) + P(X_1, X_2, \ldots, X_n \mid H)}$$

- Because of this, we would have to consider the *joint* probabilities of each feature with one another to determine a message is ham or spam

$$P(S \mid X_1 = x1, X_2 = x2, \ldots, X_n = xn) = \frac{P(X_1 = x1, X_2 = x2, \ldots, X_n = xn \mid S)}{P(X_1 = x1, X_2 = x2, \ldots, X_n = xn \mid S) + P(X_1 = x1, X_2 = x2, \ldots, X_n = xn \mid H)}$$

NAÏVE BAYES

- How could we eliminate the problem of having to consider the joint probabilities between each of our estimates?

NAÏVE BAYES

- How could we eliminate the problem of having to consider the joint probabilities between each of our estimates?
- That's right – we're going to pretend that never happened and ASSUME it away (phew, that was easy)

- In other words, we're going to ASSUME feature independence, which means we do not have to consider the joint probabilities for each message, but instead just the probability of each feature vs its likelihood of being spam

- *(A physicist, chemist, and economist are stuck on an island with a can of corn…)*

NAÏVE BAYES

- In other words, we're going to ASSUME feature independence, which means we do not have to consider the joint probabilities for each message, but instead just the probability of each feature vs its likelihood of being spam

$$P(X_{1=x1}, X_{2=x2}, \ldots, X_{n=xn} \mid S) = P(X_{1=x1} \mid S) * P(X_{2=x2} \mid S) \ldots P(X_{n=xn} \mid S)$$

$$P(S \mid X_{1=x1}, X_{2=x2}, \ldots, X_{n=xn}) = \prod_{i=1}^{n} P(X_i = x_i \mid S) / C$$

- [Where C is some constant for our marginal probability of those data]

NAÏVE BAYES

- Ok, so intuitively, we're going to be looking at various email messages. For each of those words, we're going to examine the "spaminess" and "haminess" of those words based on any previous training set we've seen.

- Think about it like this: if I'm looking at a given message, and there's a lot of single words in there that correspond to spam messages, I'm going to probability call that email spam.

- In this case, the classification of spam is our target y. The X features are each of the individual words we're examining.

NAÏVE BAYES

- The fruit stand example…
- If I hand you a basket of fruit, and you pick any given fruit up, you have to tell me what type of fruit it is (based on your previous knowledge of fruits). However, you have to look at each feature in independence.
- 1) You pick something up that is 3” in diameter. What is it?

NAÏVE BAYES

- The fruit stand example…

- If I hand you a basket of fruit, and you pick any given fruit up, you have to tell me what type of fruit it is (based on your previous knowledge of fruits). However, you have to look at each feature in independence.

- 1) You pick something up that is 3” in diameter. What is it?
- 2) This fruit is also yellow. What is it?

NAÏVE BAYES

- The fruit stand example…

- If I hand you a basket of fruit, and you pick any given fruit up, you have to tell me what type of fruit it is (based on your previous knowledge of fruits). However, you have to look at each feature in independence.

- 1) You pick something up that is 3" in diameter. What is it?
- 2) This fruit is also yellow. What is it?

- Our model looks at every single one of those features in independence to determine its fruit classification.

NAÏVE BAYES

- But wait, a few more things…

- 1) How does it determine the y-class to estimate?
- 2) What are we assuming about our feature distributions?

NAÏVE BAYES

- 1.) How do we maximize our y probability estimate for some given data?
- So you're asking me…to produce a maximum resulting estimate...for some previous amount of data you've shown me...
- What does this sound like? What is the y-class in this example?

NAÏVE BAYES

- 1.) How do we maximize our y probability estimate for some given data?

- We're going to perform Maximum A Posteriori estimation to find the maximum of some y against some X data.

$$P(y \mid x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

$$\Downarrow$$

$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y),$$

- (Refer to MLE/MAP)

NAÏVE BAYES

- What are we assuming about our feature distributions?
- Ok, maybe you didn't think about this. In other words, I'm asking what did you assume about how your features were distributed? Were they just continuous values? Where they only zeroes or ones? Where they multinomial?
- This impacts how we fulfill our classification of a given sample. What distribution does X follow?

NAÏVE BAYES

- There are three likelihood functions from which we'll instantiate for our X

$$P(X_1 = x1, X_2 = x2, \ldots, X_n = xn \mid S)$$

- For a binary X event, we'll model with a binomial distribution
- For >2 discrete outcomes, we'll assume multinomial distribution
- For real valued features, we'll assume Gaussian

NAÏVE BAYES

- Final notes:

- Despite violating a clear assumption (words are independent from one another), Naive Bayes models have proven to be quite effective in text problems.
- Given the independence assumption between features, Naive Bayes models are quite fast to train
- While Naive Bayes can perform well on entire class estimates, the predicted probabilities of a given class (predict proba) are suspect, given the independence assumption and potential poor observations in a dataset