GENERAL ASSEMBLY

DATA SCIENCE

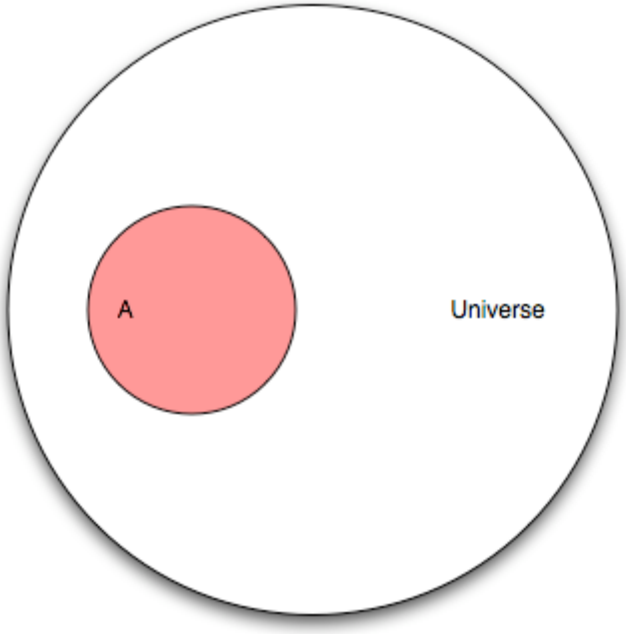
NAIVE BAYES CLASSIFICATION AND ROC/AUC CURVES

I. PROBABILITY AND BAYES' THEOREM

II. NAÏVE BAYES CLASSIFICATION

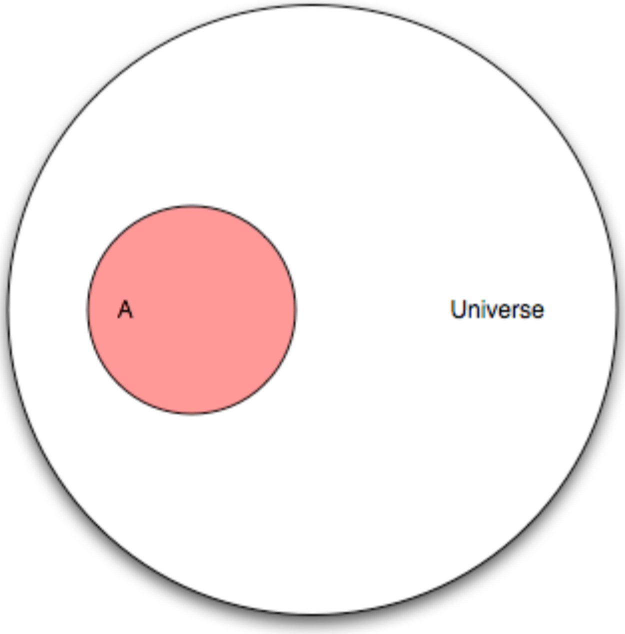
III. ROC AUC CURVES

I. PROBABILITY AND BAYES' THEOREM



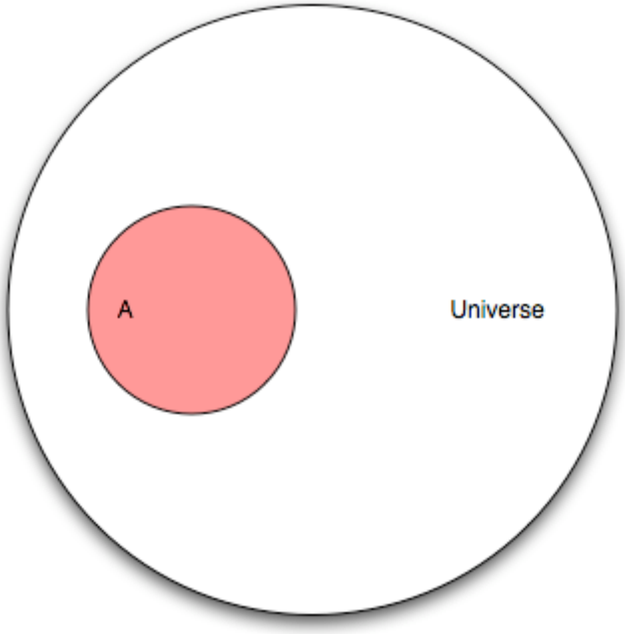
Let's pretend you are flipping a coin. This diagram represents the "universe" of all possible outcomes, also known as events. This universe is known as the sample space.

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



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.

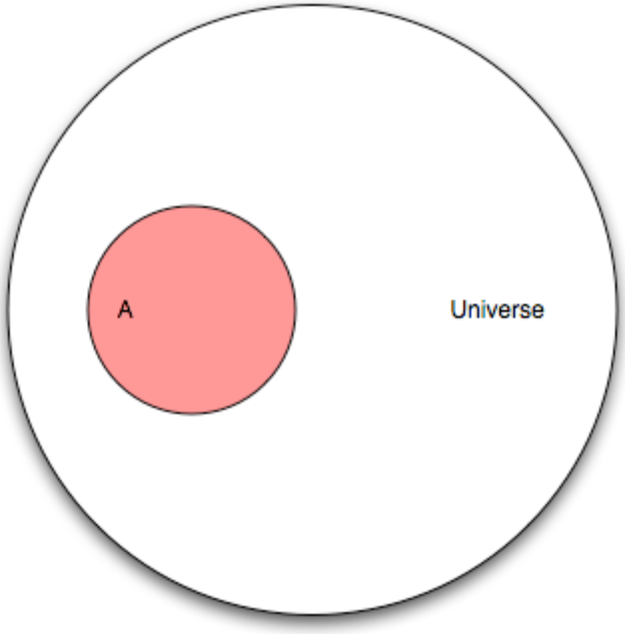
Q: What are the mutually exclusive events that make up the sample space for a coin flip?
A: Heads and tails



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?

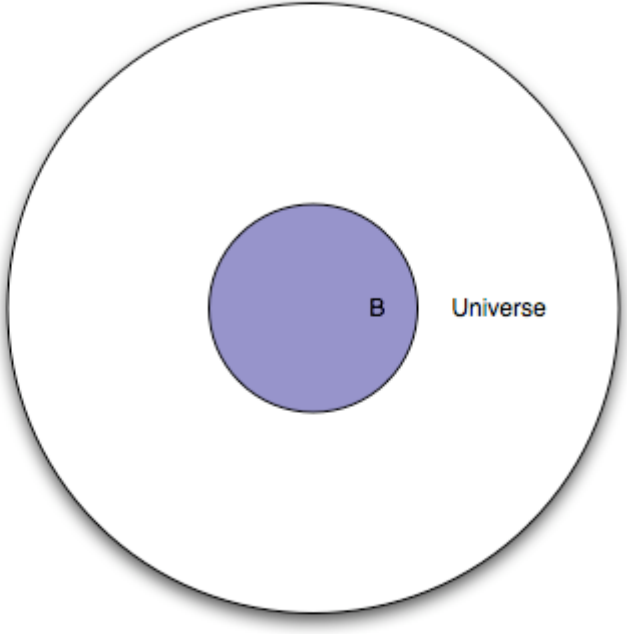
Q: What is the max probability of any event?



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

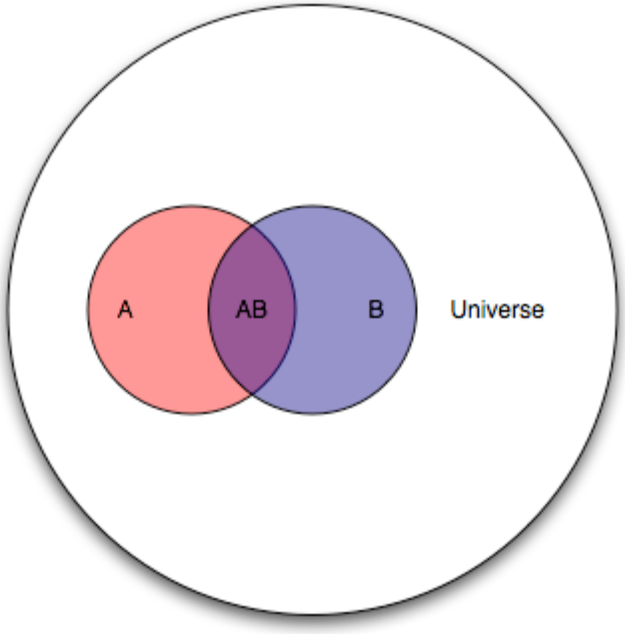
Q: If our study has 100 people and "A" has 25 people, what is the probability of A?
A: P(A) = 25/100

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



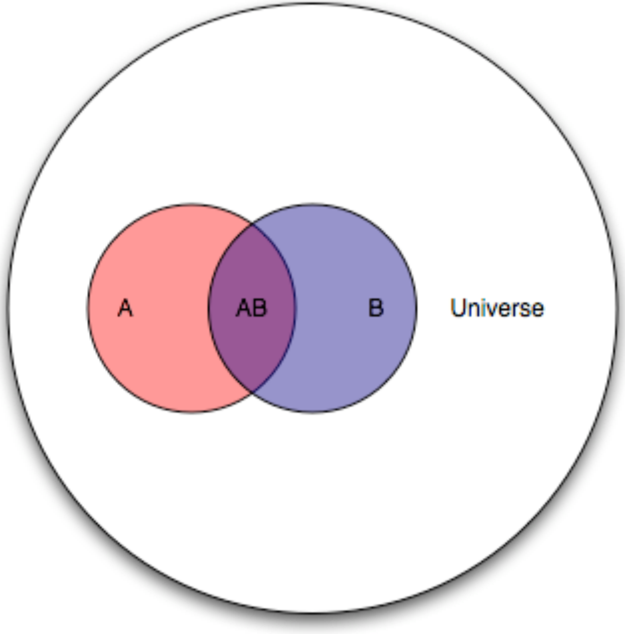
This represents the same set of people, except everyone in the study is given a test. Event “B” is everyone in the study for whom the test is positive.

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?



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

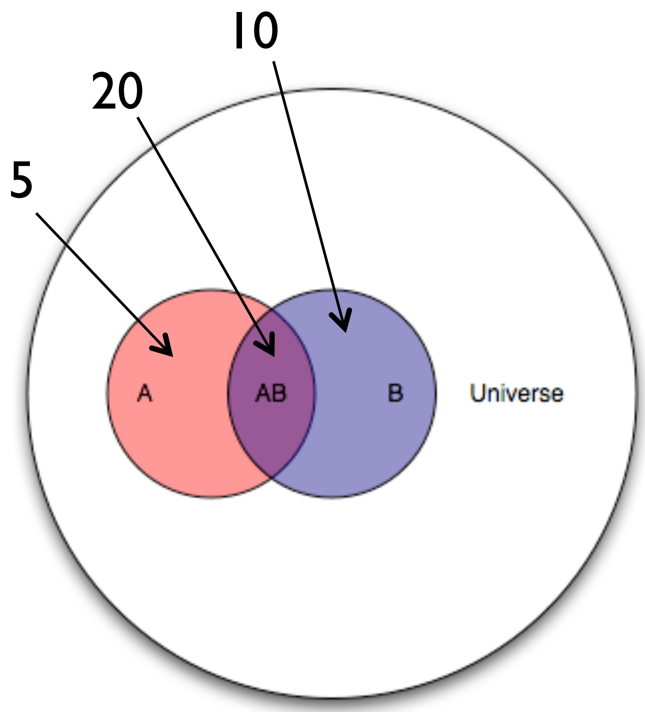
Q: How would you describe the “cancer status” and “test status” of people in each area of the diagram?

A: Pink: cancer, negative test

Purple: cancer, positive test

Blue: no cancer, positive test

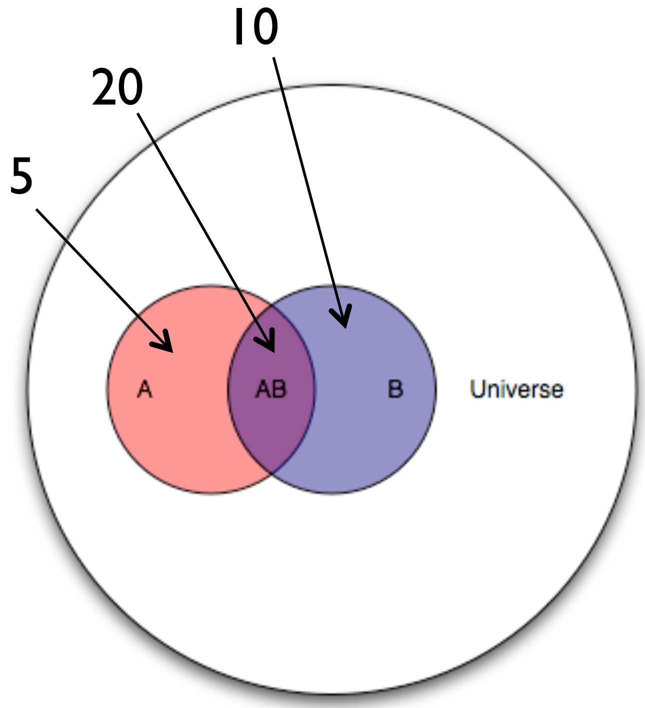
White: no cancer, negative test



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

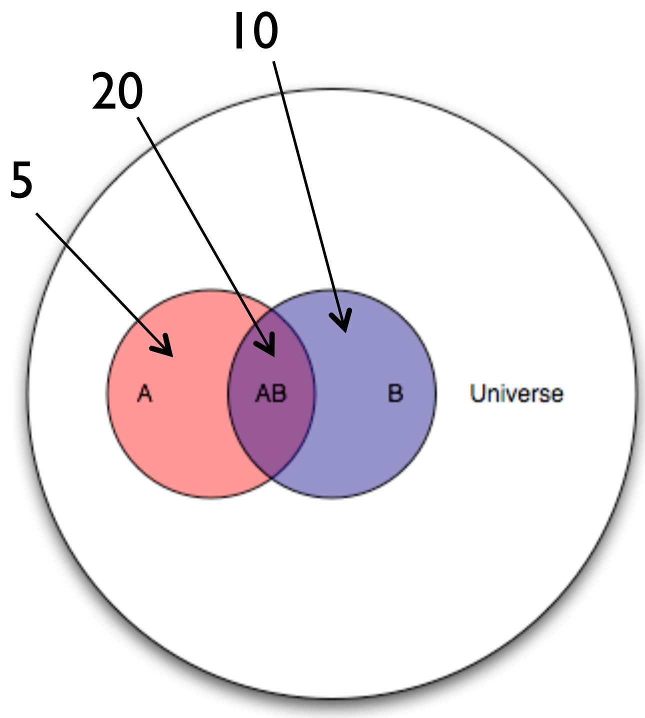
Thinking of this test as a classifier for predicting cancer, draw the confusion matrix.

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?

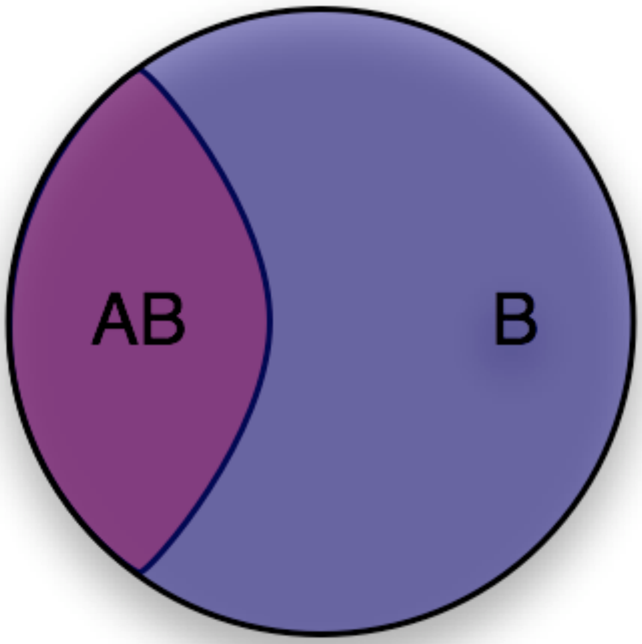
n=100	Predicted: NO	Predicted: YES
Actual: NO	65	10
Actual: YES	5	20



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

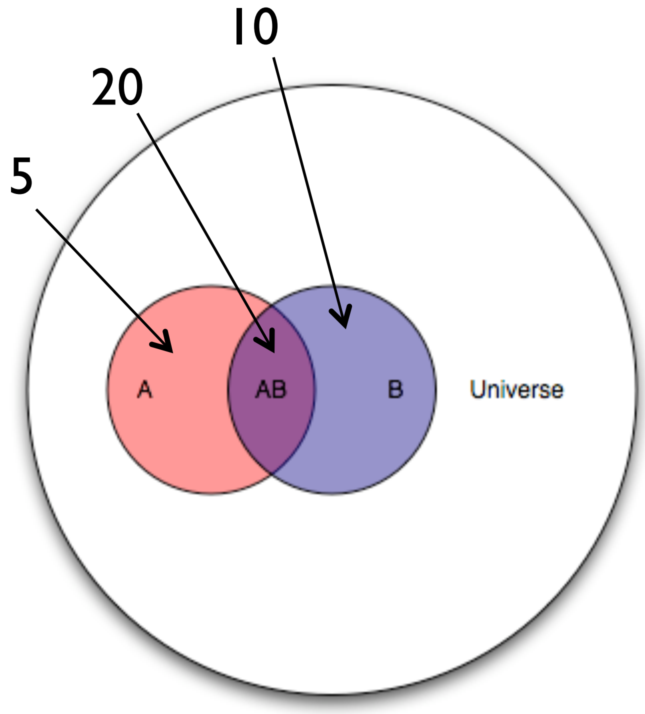
This is the conditional probability of A given B, **denoted as P(A|B).**

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

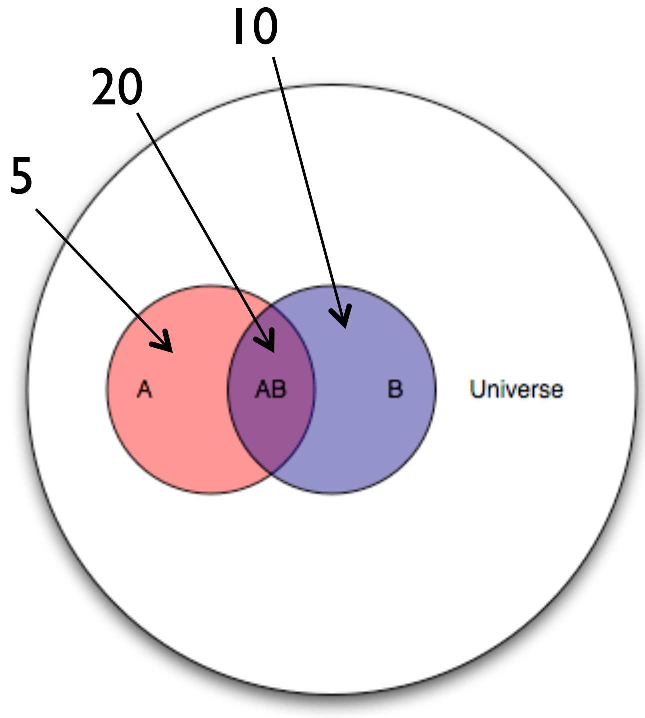


You can think of conditional probability as "changing the relevant universe." P(A|B) is a way of saying "Given that my entire universe is now B, what is the probability of A?"

This is also known as transforming the sample space.



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?



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)

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%

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

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 $P(B \mid \text{not } A) * P(\text{not } A)$

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

$= .010098$

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

Bayes Theorem: $= .99 * .0001 / .010098 = 0.00980$

II. NAÏVE BAYES 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?

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

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.

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.**

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

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

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

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

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

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.**

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

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.

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

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

Q: 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.

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

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. ROC AUC CURVES

Email Number	Score	True Label
5	0.99	Spam
8	0.82	Spam
2	0.60	Spam
1	0.60	Ham
7	0.48	Spam
3	0.22	Ham
4	0.10	Ham
6	0.02	Ham

Every email is assigned a "spamminess" score by our classification algorithm. To actually make our predictions, we choose a numeric cutoff for classifying as spam.

An ROC Curve will help us to visualize how well our classifier is doing without having to choose a cutoff!

Email Number	Score	True Label
5	0.99	Spam
8	0.82	Spam
2	0.60	Spam
1	0.60	Ham
7	0.48	Spam
3	0.22	Ham
4	0.10	Ham
6	0.02	Ham

The ROC plots the True Positive Rate (TRP) on the y-axis against the False Positive Rate (FPR) on the x-axis.

TPR: When actual value is spam, how often is prediction correct?

FPR: When actual value is ham, how often is prediction wrong?

Email Number	Score	True Label
5	0.99	Spam
8	0.82	Spam
2	0.60	Spam
1	0.60	Ham
7	0.48	Spam
3	0.22	Ham
4	0.10	Ham
6	0.02	Ham

TPR: When actual value is spam, how often is prediction correct?

FPR: When actual value is ham, how often is prediction wrong?

Cutoff	TPR (y)	FPR (x)	Cutoff	TPR (y)	FPR (x)
0			0.50		
0.05			0.65		
0.15			0.85		
0.25			1		

Email Number	Score	True Label
5	0.99	Spam
8	0.82	Spam
2	0.60	Spam
1	0.60	Ham
7	0.48	Spam
3	0.22	Ham
4	0.10	Ham
6	0.02	Ham

TPR: When actual value is spam, how often is prediction correct?

FPR: When actual value is ham, how often is prediction wrong?

Cutoff	TPR (y)	FPR (x)	Cutoff	TPR (y)	FPR (x)
0	1	1	**0.50**	0.75	0.25
0.05	1	0.75	**0.65**	0.5	0
0.15	1	0.5	**0.85**	0.25	0
0.25	1	0.25	**1**	0	0

Email Number	Score	True Label
5	0.99	Spam
8	0.82	Spam
2	0.60	Spam
1	0.60	Ham
7	0.48	Spam
3	0.22	Ham
4	0.10	Ham
6	0.02	Ham