GENERAL ASSEMBLY

NAIVE BAYES CLASSIFICATION

Brian Chung

LAST TIME - REGULARIZATION

Regularization (through LASSO and Ridge) allow us to reduce the variance of predictions in OLS solutions (in return for greater bias) through the **hyperparameter lambda**, thus achieving a better **overall** model fit

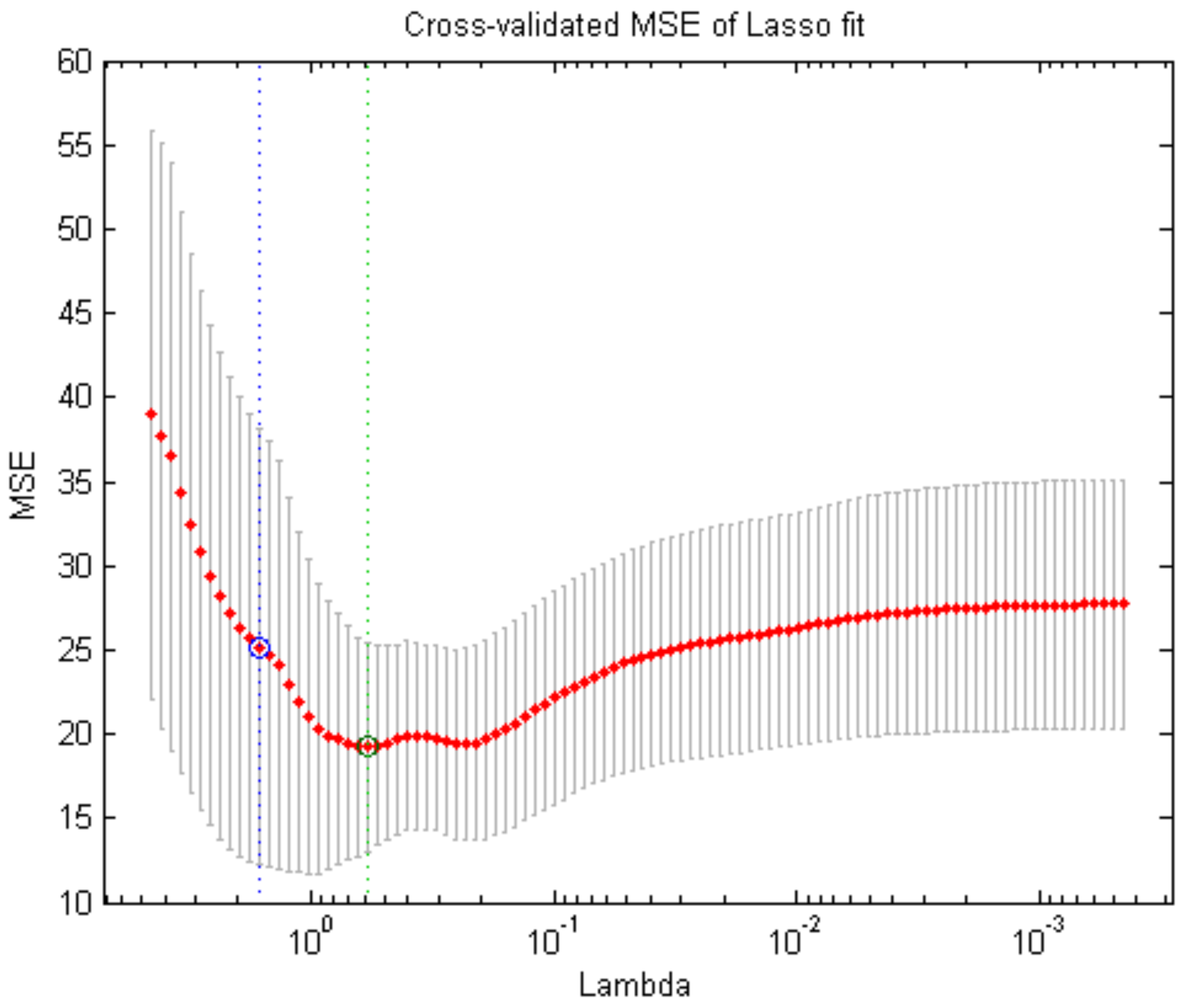
OLS: $$\min J(\beta) = \min ||(Y - X\beta)||^2$$

LASSO $$\min J(\beta) = \min(||(Y - X\beta)||^2 + \lambda||\beta||_1)$$

Ridge Regression $$\min J(\beta) = \min(||(Y - X\beta)||^2 + \lambda||\beta||_2^2)$$

HYPERPARAMETER SELECTION REVISITED

Algorithm
- Split data into train and test sets
- On train data:
  For lambda range (i.e. .0001 to 1000)
    Generate avg of cross validated MSE with that particular lambda on the K folds
- Choose the simplest lambda that results in lowest error
- Another choice is the 1 std error rule. Choose the simplest lambda that is within 1 SE of the lowest error lambda



NAIVE BAYES AGENDA

I. Intro to Probability
II. Bayes' Theorem
III. Naive Bayes Classification
IV. Spam Filter Lab

NAIVE BAYES CLASSIFICATION

I. INTRO TO PROBABILITY

INTRO TO PROBABILITY

Q: What is a **probability?**

INTRO TO PROBABILITY

Q: What is a **probability?**

A: A real number between 0 and 1 that characterizes the likelihood that some event will occur

Q: What is a **probability?**

A: A real number between 0 and 1 that characterizes the likelihood that some event will occur

The probability of event *A* occurring is denoted by *P(A)*

INTRO TO PROBABILITY

Q: What is a **probability?**

A: A real number between 0 and 1 that characterizes the likelihood that some event will occur

The probability of event *A* occurring is denoted by *P(A)*

“The probability of rain tomorrow is 67%” <—> P(A) = .67

INTRO TO PROBABILITY

Q: What is the set of all possible events called?

Q: What is the set of all possible events called?

A: This set is called the **sample space Ω.** Event A is a member of the sample space, as is every other event.

INTRO TO PROBABILITY

Q: What is the set of all possible events called?

A: This set is called the **sample space Ω.** Event *A* is a member of the sample space, as is every other event.

Ω = { rain tomorrow, not rain tomorrow }

Q: What is the set of all possible events called?

A: This set is called the **sample space Ω.** Event *A* is a member of the sample space, as is every other event.

The total probability of the sample space $P(\Omega) = 1$.

INTRO TO PROBABILITY

Q: What is the set of all possible events called?

A: This set is called the **sample space Ω.** Event *A* is a member of the sample space, as is every other event.

The total probability of the sample space $P(\Omega) = 1$.

$\mathbf{\Omega}$ = { A: rain tomorrow, A': not rain tomorrow }

P(A) = .67

INTRO TO PROBABILITY

Q: What is the set of all possible events called?
A: This set is called the **sample space Ω.** Event *A* is a member of the sample space, as is every other event.

The total probability of the sample space $P(\Omega) = 1$.

$\boldsymbol{\Omega}$ = { A: rain tomorrow, A’: not rain tomorrow }
P(A) = .67
P(A’) = ???

INTRO TO PROBABILITY

Q: What is the set of all possible events called?

A: This set is called the **sample space Ω.** Event *A* is a member of the sample space, as is every other event.

The total probability of the sample space $P(\Omega) = 1$.

$\boldsymbol{\Omega}$ = { A: rain tomorrow, A': not rain tomorrow }

P(A) = .67

P(A') = 1 - .67 = **.33**

INTRO TO PROBABILITY

Q: What is the set of all possible events called?

A: This set is called the **sample space Ω.** Event *A* is a member of the sample space, as is every other event.

The **rule of subtraction** says that the probability of event A will occur is equal to 1 minus that probability that A will **not** occur

$$P(A) = 1 - P(A')$$

Q: Consider two events A and B. How can we characterize the intersection of these events both occurring?

Q: Consider two events *A* and *B*. How can we characterize the intersection of these events both occurring?

A: Through the **joint probability** of *A* and *B*, written *P(AB)*

Q: Consider two events A and B. How can we characterize the intersection of these events both occurring?

A: Through the **joint probability** of A and B, written $P(AB)$

Q: Consider two events A and B. How can we characterize the intersection of these events both occurring?

A: Through the **joint probability** of A and B, written $P(AB)$

Another notation for the joint probability is $P(A \cap B)$

" Probability of A intersection B"

INTRO TO PROBABILITY

Q: Consider two events *A* and *B*. How can we characterize the intersection of these events both occurring?

A: Through the **joint probability** of *A* and *B*, written *P(AB)*

Another notation for the joint probability is $P(A \cap B)$

" Probability of A intersection B"

"The probability of the blue line being delayed AND CTA buses running late is .95"

P(blue line delayed, CTA buses running late) = .95

Q: What about either A and/or B happening?

Q: What about either A and/or B happening?
A: This is called the union of A and B. This is denoted by P(A U B)

“Probability of A union B”

INTRO TO PROBABILITY

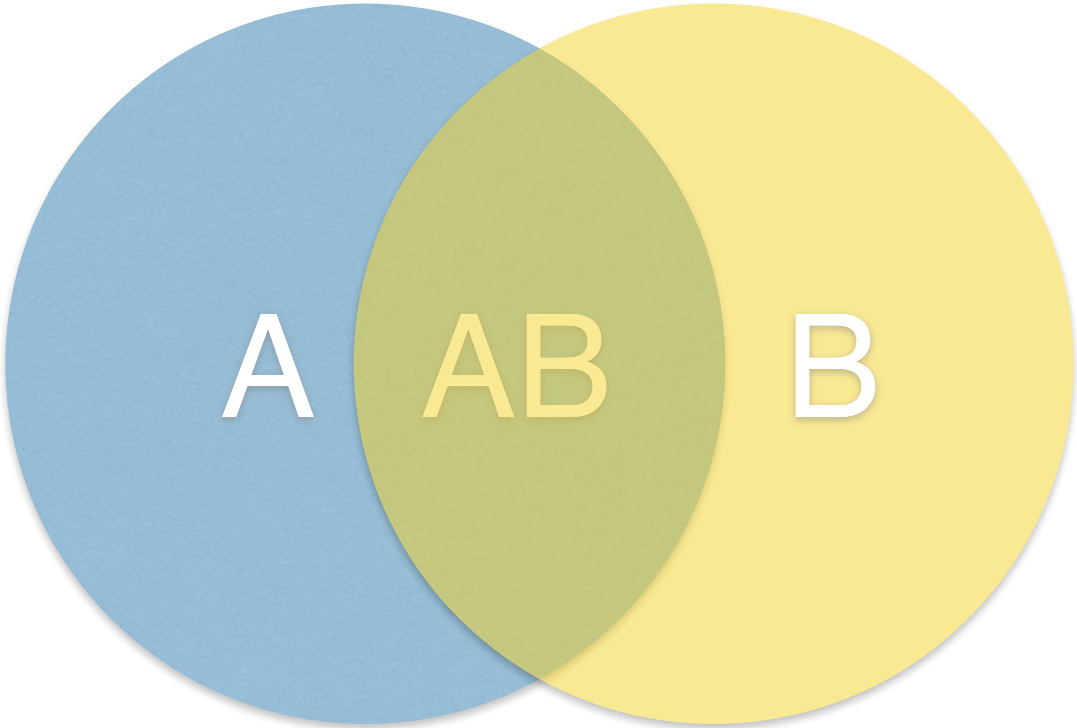
Q: What about either A and/or B happening?
A: This is called the union of A and B. This is denoted by P(A U B)

$$P(A \cup B) = P(A) + P(B) - P(AB)$$

Why?

Q: What about either A and/or B happening?
A: This is called the union of A and B. This is denoted by P(A U B)

P(A U B) = P(A) + P(B) - P(AB)



Imagine probabilities are represented by the area of the circles.

The events are A and B.

The total probability would be area of A + area of B. But this would double count that area in-between of P(AB)!

INTRO TO PROBABILITY

Q: What about either A and/or B happening?
A: This is called the union of A and B. This is denoted by P(A U B)

$$P(A \cup B) = P(A) + P(B) - P(AB)$$

P(A) = Probability of getting sick today = .10
P(B) = Probability of stepping in a puddle = .40
P(AB) = Probability of getting sick AND stepping in a puddle = .2
$P(A \cup B) = P(A) + P(B) - P(AB) = .1 + .4 - .2 = .3$

Q: Suppose event B has occurred. How do we represent the probability of A given this information about B?

INTRO TO PROBABILITY

Q: Suppose event *B* has occurred. How do we represent the probability of *A* given this information about *B*?

A: The intersection of *A* and *B* occurring, divided by the region of *B.*

Q: Suppose event *B* has occurred. How do we represent the probability of *A* given this information about *B*?

A: The intersection of *A* and *B* occurring, divided by the region of *B.*

This is called a **conditional probability**

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

"The probability of A given B is the joint probability of AB divided by the probability of B"

INTRO TO PROBABILITY

Q: Suppose event *B* has occurred. How do we represent the probability of *A* given this information about *B*?
A: The intersection of *A* and *B* occurring, divided by the region of *B.*

This is called a **conditional probability**

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

With some shuffling, we can also represent the joint probability as:

P(AB) = P(A|B) * P(B) or P(BA) = P(B|A) * P(A)

Q: What does it mean for two events to be **independent**?

Q: What does it mean for two events to be **independent**?
A: It means that information about one does not affect the probability of the other.

We encode this information as:

$P(A|B) = P(A)$ (IF A is independent of B)

"Probability of A given B is just the Probability of A. The knowledge of what B is does not affect our beliefs about A"

"There is a 38% chance it will rain tomorrow. In addition, normally I have a probability of 17% of drinking diet coke.

Given that I drink a diet coke tomorrow, what is the probability it will not rain?

INTRO TO PROBABILITY

"There is a 38% chance it will rain tomorrow. In addition, normally I have a probability of 17% of drinking diet coke.

Given that I drink a diet coke tomorrow, what is the probability it will not rain?

A: Probability of raining

B: Probability of drinking a diet coke

INTRO TO PROBABILITY

“There is a 38% chance it will rain tomorrow. In addition, normally I have a probability of 17% of drinking diet coke.

Given that I drink a diet coke tomorrow, what is the probability it will not rain?

A: Probability of raining = .38

B: Probability of drinking a diet coke = .17

INTRO TO PROBABILITY

“There is a 38% chance it will rain tomorrow. In addition, normally I have a probability of 17% of drinking diet coke.

Given that I drink a diet coke tomorrow, what is the probability it will not rain?

A: Event space of rain, not rain

B: Event space of drank a coke, did not drink a coke

P(A = not rain I B = drank a coke)

INTRO TO PROBABILITY

“There is a 38% chance it will rain tomorrow. In addition, normally I have a probability of 17% of drinking diet coke.

Given that I drink a diet coke tomorrow, what is the probability it will not rain?

A: Event space of rain, not rain

B: Event space of drank a coke, did not drink a coke

P(A = not rain | B = drank a coke) = P(A = not rain)

INTRO TO PROBABILITY

“There is a 38% chance it will rain tomorrow. In addition, normally I have a probability of 17% of drinking diet coke.

Given that I drink a diet coke tomorrow, what is the probability it will not rain?

A: Event space of rain, not rain

B: Event space of drank a coke, did not drink a coke

P(A = not rain | B = drank a coke) = P(A = not rain) = 1 - .38 = **.62**

Summary of Rules so far:

1. $P(A) = 1 - P(A')$
2. $\sum P(\text{ all events in } A\ ) = P(\Omega) = 1$
3. $P(AB) = P(A|B) * P(B)$ [or $P(BA) = P(B|A) * P(A)$
4. $P(AUB) = P(A) + P(B) - P(AB)$
5. $P(A|B) = P(A)$ **iff A is independent of B**

TEST YOUR KNOWLEDGE!

1. An urn contains 6 black marbles and 4 red marbles. Two marbles are drawn WITH replacement from the urn. What is the probability that both of the marbles are black?

2. An urn contains 6 red marbles and 4 black marbles. Two marbles are drawn *without* replacement from the urn. What is the probability that both of the marbles are black? (Hint: represent B as the event that the first marble is black; let A represent the event that the second marble is black; and use rule 3 from above)

3. A card is randomly drawn from a deck of ordinary playing cards. You win $50 dollars if the card is a heart or an Ace. What is the probability that you will win the game? (Hint: Use rule 4)

I. INTRO TO PROBABILITY
II. BAYES' THEOREM

BAYES RULE

Remember earlier:

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

BAYES RULE

Remember earlier:

$P(AB) = P(A|B) * P(B)$	*Joint Probability*
$P(BA) = P(B|A) * P(A)$	*From substitution*

BAYES RULE

Remember earlier:

$P(AB) = P(A|B) * P(B)$	*Joint Probability*
$P(BA) = P(B|A) * P(A)$	*From substitution*
$P(AB) = P(BA)$	*Event AB is same as event BA*

BAYES RULE

Remember earlier:

$P(AB) = P(A \mid B) * P(B)$	*Joint Probability*
$P(BA) = P(B \mid A) * P(A)$	*From substitution*
$P(AB) = P(BA)$	*Event AB is same as event BA*
$P(A \mid B)*P(B) = P(B \mid A)*P(A)$	*Substitute the first two eq.*

BAYES RULE

Remember earlier:

$P(AB) = P(A|B) * P(B)$	*Joint Probability*
$P(BA) = P(B|A) * P(A)$	*From substitution*
$P(AB) = P(BA)$	*Event AB is same as event BA*
$P(A|B)*P(B) = P(B|A)*P(A)$	*Substitute the first two eq.*

Finally:

$$\boldsymbol{P(A|B) = P(B|A) * P(A) / P(B)}$$

BAYES RULE

Bayes Theorem (/Rule/Law): P(A|B) = P(B|A) * P(A) / P(B)

Bayes’ theorem is the ‘equation’ that launched a thousand scientists

“P(dog|eyes,jowl,ears,sound) =
*P(eyes,jowl,ears,sound|dog) * P(dog) / P(eyes,jowl,ears,sound)”*

“P(stock go up | marketconditions,) =
*P(marketconditions | stock up go) * P(stock go up) / P(market conditions)”*

BAYES RULE

Bayes Theorem (/Rule/Law): $P(A|B) = P(B|A) * P(A) / P(B)$

Bayes' theorem is the 'equation' that launched a thousand scientists

- It provides a "wormhole" between two different "interpretations" of probability (More on this later)

BAYES RULE

Bayes Theorem (/Rule/Law): P(A|B) = P(B|A) * P(A) / P(B)

Bayes' theorem is the 'equation' that launched a thousand scientists

- It provides a "wormhole" between two different "interpretations" of probability
- It is a powerful computational tool

BAYES RULE

Bayes Theorem (/Rule/Law): $P(A|B) = P(B|A) * P(A) / P(B)$

Bayes' theorem is the 'equation' that launched a thousand scientists

- It provides a "wormhole" between two different "interpretations" of probability (More on this later)
- It is a powerful computational tool
- Gives us insights into otherwise low-frequency events

BAYES RULE

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

Every term in this relationship and each plays a distinct role in probability calculations

BAYES RULE

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

The term **P(A)** is often called the **prior**. This term indicates an initial belief in A.

$$P(spam \mid words) = \frac{P(words \mid spam)P(spam)}{P(words)}$$

BAYES RULE

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

The term **P(B|A) is** called the **likelihood** function. It is a probability indicating how likely it is to see event B given event A.

$$P(spam \mid words) = \frac{P(words \mid spam)P(spam)}{P(words)}$$

BAYES RULE

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

The term **P(A|B)** is called the **posterior probability**. It indicates a "revised" estimate of A given the event of B.

$$P(spam \mid words) = \frac{P(words \mid spam)P(spam)}{P(words)}$$

BAYES RULE

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

P(B) is called the **normalization constant**. It is often ignored for comparison purposes.

$$P(spam \mid words) = \frac{P(words \mid spam)P(spam)}{P(words)}$$

BAYES RULE

As example, we can use Bayesian statistics to estimate parameters (**β**) of our models as well given data (**X)**.

$$P(\beta \mid X) = \frac{P(X \mid \beta)P(\beta)}{P(X)}$$

BAYES RULE

We can use Bayesian statistics to estimate parameters (**β**) of our models as well given data (**X)**.

$$P(\beta \mid X) = \frac{P(X \mid \beta)P(\beta)}{P(X)}$$

Coefficients of Regression
Class Labels of Samples
Student proficiency

BAYES RULE

We can use Bayesian statistics to estimate parameters (**β**) of our models as well given data (**X**).

$$P(\beta \mid X) = \frac{P(X \mid \beta)P(\beta)}{P(X)}$$

Note that you get a probability distribution, not just a point estimate!

BAYES RULE

We can use Bayesian statistics to estimate parameters (**β**) of our models as well given data (**X)**.

$$P(\beta \mid X) = \frac{P(X \mid \beta) P(\beta)}{P(X)}$$

Data points in Euclidean space
List of labeled samples
Student responses

BAYES RULE

Starting with a prior belief of the parameters β…

$$P(\beta \mid X) = \frac{P(X \mid \beta) P(\beta)}{P(X)}$$

What are reasonable coefficients?
What are common class labels?
How are student scores generally distributed?

BAYES RULE

…and updating them as new data comes in

$$P(\beta \mid X) = \frac{P(X \mid \beta)P(\beta)}{P(X)}$$

Given the new data, are the coefficients likely?
What is the probability of the data having this set of class labels?
How likely are these student scores?

BAYES RULE

The **Maximum likelihood estimator (MLE)** finds the parameters that make the data most likely without prior belief

$$P(\beta \mid X) = \frac{P(X \mid \beta) P(\beta)}{P(X)}$$

BAYES RULE

The normalization constant is generally ignored (and can be computed anyway if need be)

$$P(\beta \mid X) = \frac{P(X \mid \beta)P(\beta)}{P(X)}$$

How likely is this data anyway?

BAYES RULE

The heart of Bayesian inference is that we take our initial beliefs in **β**, and update those beliefs through the likelihood function P(X|β)

$$P(\beta \mid X) = \frac{P(X \mid \beta) P(\beta)}{P(X)}$$

BAYES RULE

The **Maximum a posteriori (MAP)** estimate finds the parameters of beta that are most likely, given our prior and the data

$$P(\beta \mid X) = \frac{P(X \mid \beta)P(\beta)}{P(X)}$$

BAYES RULE

The **Maximum a posteriori (MAP)** estimate finds the parameters of beta that are most likely, given our prior and the data

$$P(\beta \mid X) = \frac{P(X \mid \beta) P(\beta)}{P(X)}$$

(Maximum likelihood Estimator finds parameters of beta most likely ONLY given the data)

BAYES RULE

The **Maximum a posteriori (MAP)** estimate finds the parameters of beta that are most likely, given our prior and the data

$$\beta_{MAP} = \arg\max_{\beta} P(X \mid \beta) P(\beta)$$

BAYES RULE

Unlike the algorithms we learned so far when we tried to model beta based on X, we invert this relationship!

$$P(\beta \mid X) = \frac{P(X \mid \beta)P(\beta)}{P(X)}$$

$$\beta_{MAP} = \arg\max_{\beta} P(X \mid \beta)P(\beta)$$

A TALE OF TWO STATISTICIANS

As mentioned earlier, there are two interpretations of probability which can be described as follows…

A TALE OF TWO STATISTICIANS

The two interpretations of probability can be described as follows:

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

A TALE OF TWO STATISTICIANS

The two interpretations of probability can be described as follows:

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

"

i.e. the probability of a biased coin turning up heads is the number of times it came up heads divided by the number of tosses over infinity trials

"

A TALE OF TWO STATISTICIANS

The two interpretations of probability 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.

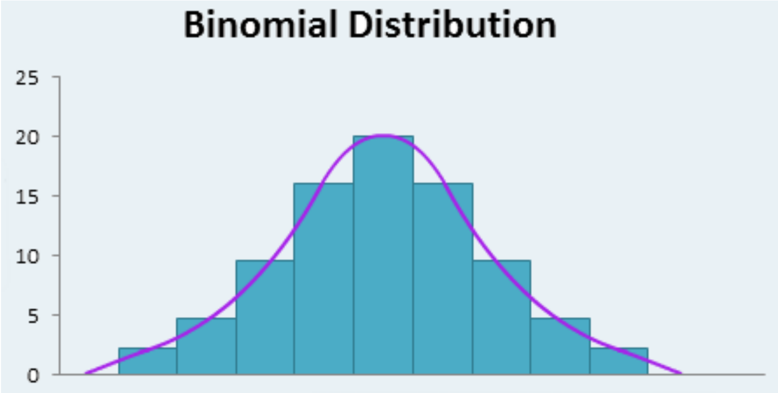
A TALE OF TWO STATISTICIANS

The two interpretations of probability 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.

"i.e. the probability of a biased coin turning up heads is based on a prior distribution, updated by sampled events"



A TALE OF TWO STATISTICIANS

Method	Predictions
Frequentist interpretation	point estimates
Bayesian interpretation	distributions

A TALE OF TWO STATISTICIANS

Bayes’ theorem provides this transition between bayesian interpretations and frequentist interpretations of probability

If this sounds crazy to you, don’t worry — we won’t focus too much on theoretical details here

A TALE OF TWO STATISTICIANS

However, I do want to show you **why Bayesian methods are cool**.

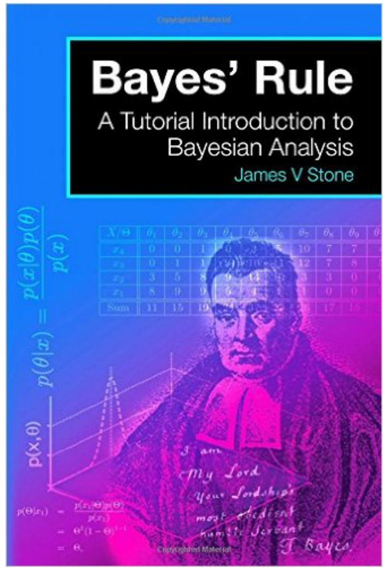
This next part will be through an Python notebook, but don’t worry about running through this on your own or understanding fully.

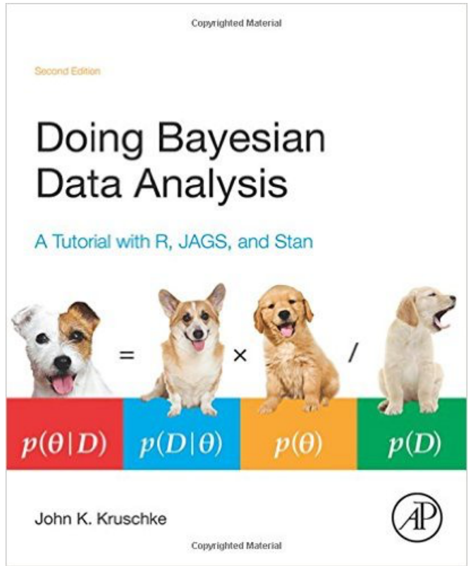
Sit back, relax, and enjoy the show!

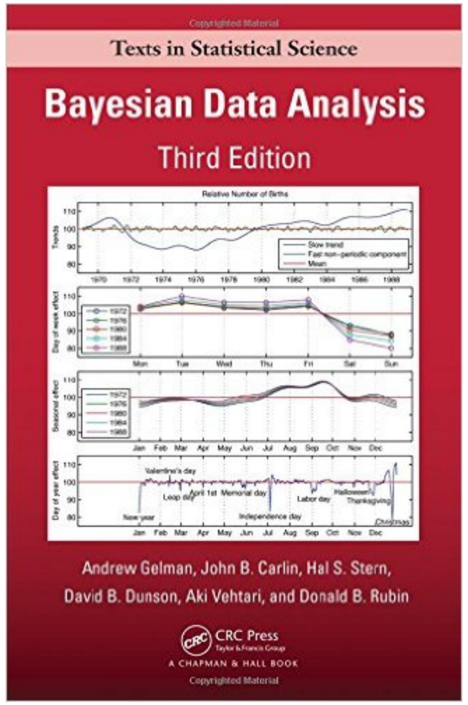
A TALE OF TWO STATISTICIANS

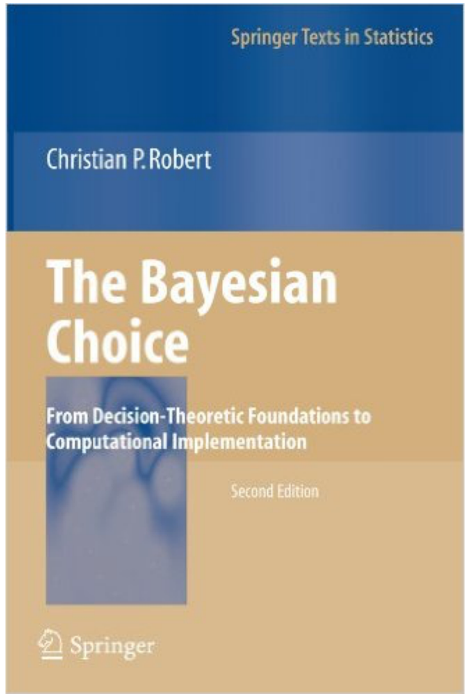
If this sounds crazy to you, don't worry — we won't focus too much on theoretical details here

If it sounds interesting, you can start on a path toward awesome models and cutting edge tools





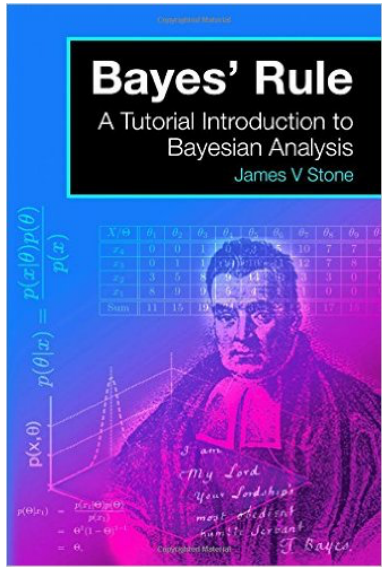


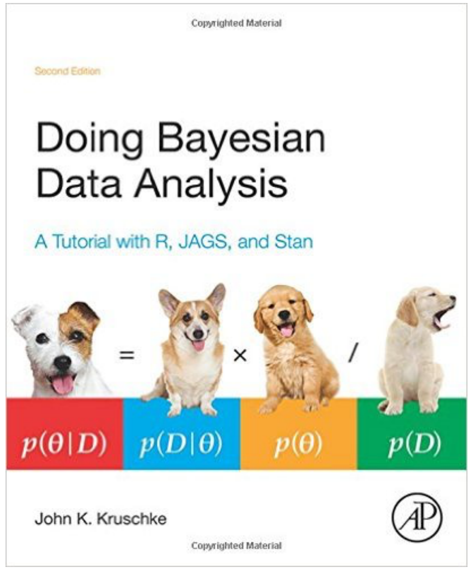


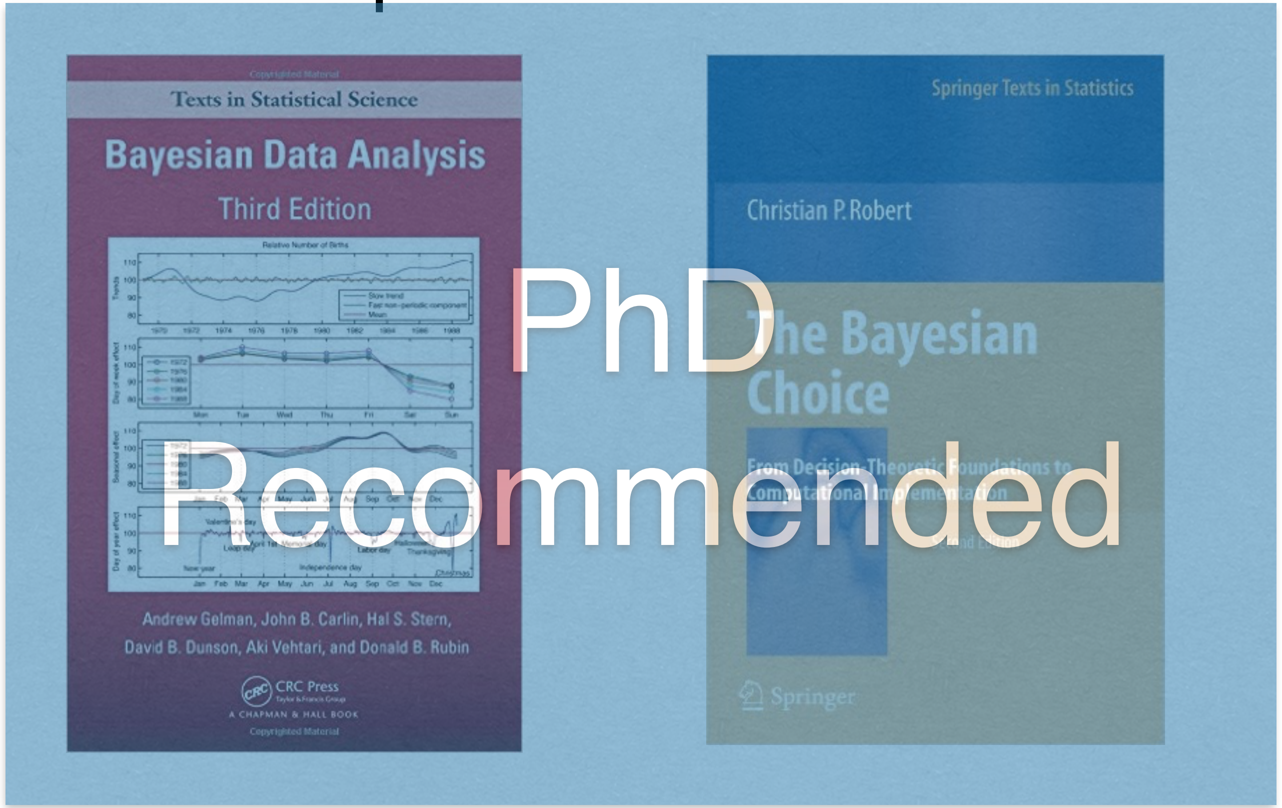
A TALE OF TWO STATISTICIANS

If this sounds crazy to you, don't worry — we won't focus too much on theoretical details here

If it sounds interesting, you can start on a path toward awesome models and cutting edge tools







I. INTRO TO PROBABILITY
II. BAYES' THEOREM
III. NAIVE BAYES

NAIVE BAYES CLASSIFICATION

We can even use Bayes' Theorem for **classification**. Suppose we have a dataset with features $\mathbf{x_1}$, $\mathbf{x_2}$, $\mathbf{x_2}$, $\mathbf{\ldots}$, $\mathbf{x_n}$ and class labels **C**. What can we say about classification using Bayes' theorem?

NAIVE BAYES CLASSIFICATION

We can even use Bayes' Theorem for **classification**. Suppose we have a dataset with features $\mathbf{x_1}$**,** $\mathbf{x_2}$**,** $\mathbf{x_2}$**, …,** $\mathbf{x_n}$ and class labels **C**. What can we say about classification using Bayes' theorem?

In this case, let's say our dataset is Email text, $\mathbf{x_1}$**,** $\mathbf{x_2}$**,** $\mathbf{x_2}$**, …,** $\mathbf{x_n}$ is the count of various words in emails, and the class labels within **C** is **spam** or **not spam**

NAIVE BAYES CLASSIFICATION

We can even use Bayes' Theorem for **classification**. Suppose we have a dataset with features $\mathbf{x_1}$, $\mathbf{x_2}$, $\mathbf{x_2}$, **…**, $\mathbf{x_n}$ and class labels **C**. What can we say about classification using Bayes' theorem?

spam == 1	x_1 = very	x_2 = drugs	x_3 = medicine	x_4 = novel
1	3	5	1	0
1	0	2	4	1
0	5	0	1	3
1	3	2	0	4

NAIVE BAYES CLASSIFICATION

We can even use Bayes' Theorem for **classification**. Suppose we have a dataset with features $\mathbf{x_1}$, $\mathbf{x_2}$, $\mathbf{x_2}$, **…**, $\mathbf{x_n}$ and class labels **C**. What can we say about classification using Bayes' theorem?

$$P(C = c \mid x_1 x_2 x_n) = \frac{P(x_1 x_2 x_n \mid C = c) P(C = c)}{P(x_1 x_2 x_n)}$$

What is the chance that the class is Spam/Not Spam given the presence/count of certain words?

NAIVE BAYES CLASSIFICATION

We can even use Bayes' Theorem for **classification**. Suppose we have a dataset with features $\mathbf{x_1}$, $\mathbf{x_2}$, $\mathbf{x_2}$, **…**, $\mathbf{x_n}$ and class labels **C**. What can we say about classification using Bayes' theorem?

$$P(C = c \mid x_1 x_2 x_n) = \frac{P(x_1 x_2 x_n \mid C = c) P(C = c)}{P(x_1 x_2 x_n)}$$

Prior belief on whether an email is spam or ham (not spam)

NAIVE BAYES CLASSIFICATION

We can even use Bayes' Theorem for **classification**. Suppose we have a dataset with features **x_1, x_2, x_2, …, x_n** and class labels **C**. What can we say about classification using Bayes' theorem?

$$P(C=c \mid x_1 x_2 x_n) = \frac{P(x_1 x_2 x_n \mid C=c)P(C=c)}{P(x_1 x_2 x_n)}$$

Likelihood of a set of words appearing given that the email is spam or ham

NAIVE BAYES CLASSIFICATION

The likelihood here is difficult to compute because of a lack of data. Suppose $\mathbf{x_1}$, $\mathbf{x_2}$, $\mathbf{x_2}$, …, $\mathbf{x_n}$ represents the counts of all the words in an email. How many emails do we have in our dataset to make a probability estimate where $\mathbf{x_1}$=apple, $\mathbf{x_2}$=xylophone, $\mathbf{x_3}$=tree, … $\mathbf{x_n}$=purple? **It is data intractable to make joint conditional estimates of the likelihood function**

$$P(x_1 x_2 x_n \mid C = c)$$

NAIVE BAYES CLASSIFICATION

So we make a simplifying assumption. We assume that the features $\mathbf{x_i}$ are conditionally independent from each other (i.e. Given that C = spam or ham, the presence of the word “bigger” is independent of the presence of the word “apple” and so on

$$P(x_1 x_2 x_n \mid C = c) = P(x_1 \mid C = c) * P(x_2 \mid C = c) * \ldots * P(x_n \mid C = c)$$

$$P(x_1 x_2 x_n \mid C = c) = \prod_i P(x_i \mid C = c)$$

NAIVE BAYES CLASSIFICATION

This assumption is where we get the "Naive" in Naive Bayes classification. We're naively assuming that given C = spam, the presence of "simple", "trick", "doctors", and "hate" are all independent each other. When in reality, if we see "simple" and "trick", the likelihood of seeing "doctors" is pretty high!

$$P(x_1 x_2 x_n \mid C = c) = \prod_i P(x_i \mid C = c)$$

NAIVE BAYES CLASSIFICATION

That said, this assumption works pretty well for text classification still. And it makes our problem so much more computationally tractable

$$P(x_1 x_2 x_n \mid C = c) = \prod_i P(x_i \mid C = c)$$

NAIVE BAYES CLASSIFICATION

Back to our original equation with the assumption built in (I've removed the C=c or C=spam/spam and x1x2..xn notation for ease)

$$P(C \mid x) = \frac{\prod_i P(x_i \mid C) P(C)}{P(x)}$$

NAIVE BAYES CLASSIFICATION

Using the **Maximum a posteriori (MAP)** rule from earlier, we iterate through each label in C (spam, ham), and choose the class label with the highest resulting posterior

$$C_{MAP} = \arg\max_{C} \prod_{i} P(x_i \mid C) P(C)$$

NAIVE BAYES CLASSIFICATION

This means we have to choose a distribution for P(x|C) (**event model**), and P(C) **(prior model)**

We will use a **multinomial distribution** for the event model & a **categorical distribution** for the prior model

$$C_{MAP} = \arg\max_{C} \prod_{i} \underbrace{P(x_i \mid C)}_{\text{event model}} \underbrace{P(C)}_{\text{prior model}}$$

NAIVE BAYES CLASSIFICATION

The MAP estimates for P(C) and P(w|C) over all your training documents:

$$P(c_j) = \frac{doc\_count(C == c_j)}{N_{doc}}$$

$$P(w_i \mid c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

NAIVE BAYES CLASSIFICATION

Estimating P(C) and P(w|C) over all your training documents:

$$P(c_j) = \frac{doc\_count(C == c_j)}{N_{doc}}$$

Fraction of documents having class == j

$$P(w_i \mid c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

NAIVE BAYES CLASSIFICATION

Estimating P(C) and P(w|C) over all your training documents:

$$P(c_j) = \frac{doc\_count(C == c_j)}{N_{doc}}$$

Fraction of documents having class == j

$$P(w_i \mid c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

Fraction of times word w_i appears among all words in documents of class c_j

NAIVE BAYES CLASSIFICATION

Great, so over our training documents, we've calculated P(C) and P(w|C) for all the words. And can classify new documents using the MAP estimate

Testing Documents

Training Documents

$$P(w_i \mid c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

$$P(c_j) = \frac{doc\_count(C == c_j)}{N_{doc}}$$

$$C_{MAP} = \arg\max_{C} \prod_{i} P(w_i \mid C) P(C)$$

Predicted Class Label

NAIVE BAYES CLASSIFICATION

Great! We can calculate P(C) and P(w|C) quite easily.

One problem before we can choose the most likely class for a given document…

NAIVE BAYES CLASSIFICATION

Suppose we have never seen the word "**plastic**" in our training samples

$$P(w_i \mid c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

$$P(\text{"}plastic\text{"} \mid spam) = \frac{count(\text{"}plastic\text{"}, spam)}{\sum_{w \in V} count(w, spam)}$$

$$P(\text{"}plastic\text{"} \mid spam) = 0$$

$$P(\text{"}plastic\text{"} \mid ham) = 0$$

NAIVE BAYES CLASSIFICATION

Suppose we have never seen the word "**plastic**" in our training samples

$$P(w_i \mid c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

$$P("plastic" \mid spam) = \frac{count("plastic", spam)}{\sum_{w \in V} count(w, spam)}$$

$$P("plastic" \mid spam) = 0$$

$$P("plastic" \mid ham) = 0$$

So which is it??

NAIVE BAYES CLASSIFICATION

The probabilities will be 0 for either class, and any documents that have words not encountered in the training documents will not be classified correctly. Solution to this is called a “Laplace Smoothing”

$$P(w_i \mid c_j) = \frac{count(w_i, c_j) + 1}{\left(\sum_{w \in V} count(w, c_j)\right) + |V|}$$

i.e. add 1 to the numerator, and add the total number of unique words, V, to the denominator

NAIVE BAYES CLASSIFICATION

The probabilities will be 0 for either class, and any documents that have words not encountered in the training documents will not be classified correctly. Solution to this is called a "Laplace Smoothing"

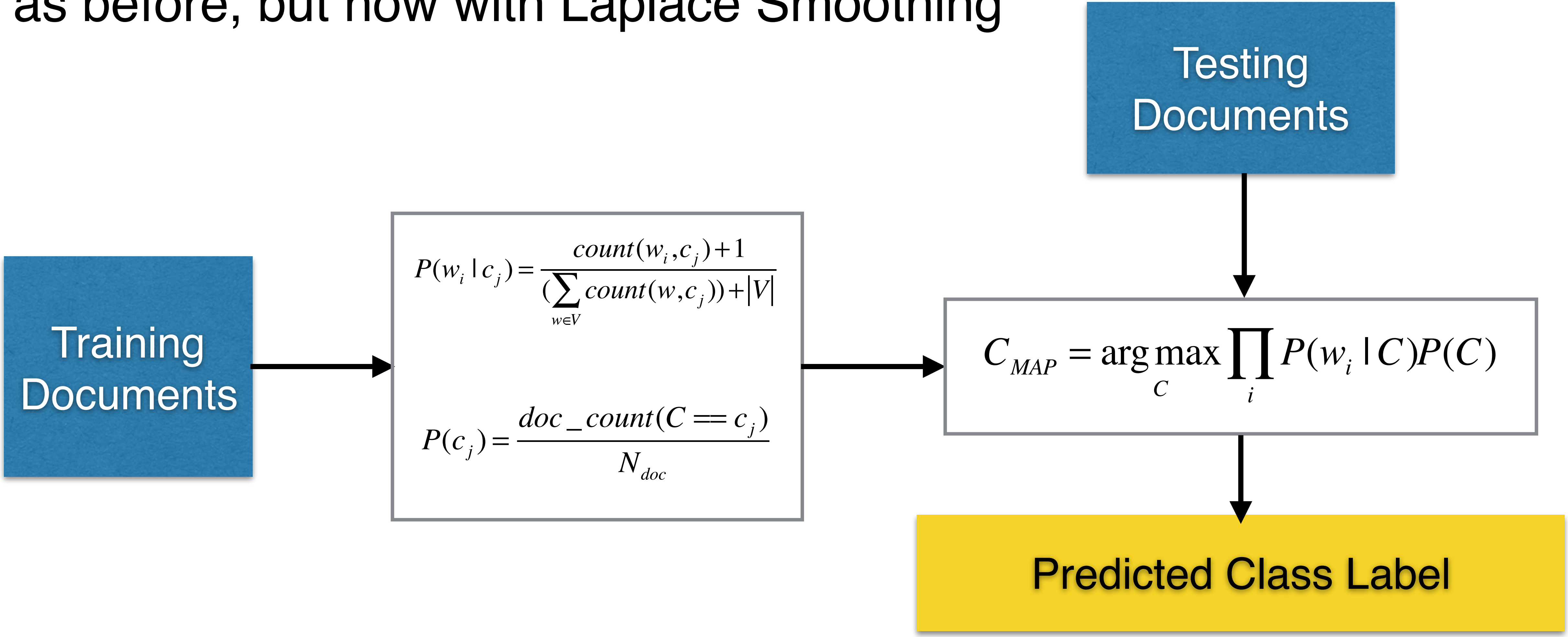
$$P(w_i \mid c_j) = \frac{count(w_i, c_j) + 1}{\left(\sum_{w \in V} count(w, c_j)\right) + |V|}$$

|V| is the number of unique words/tokens in your TRAINING set

i.e. add 1 to the numerator, and add the total number of unique words, V, to the denominator

NAIVE BAYES CLASSIFICATION

This is our new scheme for prediction. Same as before, but now with Laplace Smoothing



NAIVE BAYES ALGORITHM

* From training corpus (set of training documents), extract *Vocabulary (set of unique words)*

* Calculate $P(c_j)$ terms
  * For each c_j in C do:
    * $docs_j$ <- all docs with class == c_j

$$P(c_j) = \frac{|docs_j|}{\#total\_docs}$$

* Calculate $P(w_k|c_j)$ terms
  * $Text_j$ <- single doc containing all $docs_j$
    * $docs_j$ <- all docs with class == c_j
    * For each word w_k in Vocabulary
      * n_k <- # occurrences of w_k in $Text_j$

$$P(w_k | c_j) = \frac{n_k + 1}{n + |V|}$$

* For each test document
  * For each unique class, calculate: $\prod_k P(x_k | C_j) P(C_j)$
  * Assign the highest value class to this document

NAIVE BAYES WORKED EXAMPLE

$$C_{MAP} = \arg\max_{C} \prod_{i} P(w_i \mid C) P(C)$$

$$P(c) = \frac{N_c}{N}$$

$$P(w \mid c) = \frac{count(w,c)+1}{count(c)+|V|}$$

Set	Doc	Words	Class
Training	1	Drugs Cheap Drugs	s
	2	Drugs Drugs Xanax	s
	3	Drugs Ebay	s
	4	Review Drug Inventory	h
Test	5	Drugs Drugs Drugs Review Inventory	?

Priors:

$$P(spam) = \frac{3}{4}$$

$$P(ham) = \frac{1}{4}$$

Conditional Probabilities:

P(Drugs|spam)=(5+1)/(8+6)= 6/14
P(Review|spam)=(0+1)/(8+6)= 1/14
P(Inventory|spam)=(0+1)/(8+6)= 1/14

P(Drugs|ham)=(1+1)/(3+6)= 2/9
P(Review|ham)=(1+1)/(3+6)= 2/9
P(Inventory|ham)=(1+1)/(3+6)= 2/9

Posterior Probabilities:

P(spam | d5) ~ 3/4 * $(6/14)^3$ * 1/14 * 1/14 = **0.0003**

P(ham | d5) ~ (1/4) * $(2/9)^3$ * 2/9 * 2/9 = **0.0001**

NAIVE BAYES WORKED EXAMPLE

$$C_{MAP} = \arg\max_{C} \prod_{i} P(w_i \mid C) P(C)$$

$$P(c) = \frac{N_c}{N}$$

$$P(w \mid c) = \frac{count(w,c)+1}{count(c)+|V|}$$

Set	Doc	Words	Class
Training	1	Drugs Cheap Drugs	s
	2	Drugs Drugs Xanax	s
	3	Drugs Ebay	s
	4	Review Drug Inventory	h
Test	5	Drugs Drugs Drugs Review Inventory	?

Priors:

$$P(spam) = \frac{3}{4}$$

$$P(ham) = \frac{1}{4}$$

Conditional Probabilities:

P(Drugs|spam)=(5+1)/(8+6)= 6/14
P(Review|spam)=(0+1)/(8+6)= 1/14
P(Inventory|spam)=(0+1)/(8+6)= 1/14

P(Drugs|ham)=(1+1)/(3+6)= 2/9
P(Review|ham)=(1+1)/(3+6)= 2/9
P(Inventory|ham)=(1+1)/(3+6)= 2/9

Posterior Probabilities:

P(spam | d5) ~ log(3/4) + 3*log(6/14) + log(1/14) + log(1/14) = **-8.107**

P(ham | d5) ~ log(1/4) + 3*log(2/9) + log(2/9) + log(2/9) = **-8.906**

I. INTRO TO PROBABILITY
II. BAYES' THEOREM
III. NAIVE BAYES
IV. LAB

THAT'S IT!

- Exit Tickets: DAT1 - Lesson 7 - Naive Bayes
- Milestone 2 is due Jan 25

- Next week is Logistic Regression