


# **INTRO to DATA SCIENCE**

## **LECTURE 5: CROSS VALIDATION AND NAÏVE BAYESIAN CLASSIFICATION**

 **DataTau** [new](#) | [comments](#) | [leaders](#) | [submit](#)

1. [▲ Comparison of ML libraries+algorithms \(SVM, RF, sklearn, etc\)](#) (github.com)  
4 points by elyase 11 hours ago | [discuss](#)
2. [▲ Jupyter Ascending: IPython 4.0 released!](#) (jupyter.org)  
8 points by tfturing 18 hours ago | [3 comments](#)
3. [▲ How a Kalman filter works, in pictures](#) (bzarg.com)  
9 points by elev8 21 hours ago | [1 comment](#)
4. [▲ Interview with DrivenData "Countable Care" winner Gilberto Titericz Jr.](#) (drivendata.org)  
2 points by isms 8 hours ago | [discuss](#)
5. [▲ Optimizing Chicago's Services with Data Science](#) (dominodatalab.com)  
9 points by AnnaAnisin 1 day ago | [discuss](#)
6. [▲ Named Entity Recognition: Examining the Stanford NER Tagger](#) (insightdatascience.com)  
4 points by woofi 1 day ago | [discuss](#)
7. [▲ Composing Music With Recurrent Neural Networks](#) (hexahedria.com)  
12 points by elyase 3 days ago | [discuss](#)
8. [▲ I was wrong about statistics](#) (win-vector.com)  
4 points by tfturing 2 days ago | [1 comment](#)
9. [▲ Understanding Bayes: Visualization of the Bayes Factor](#) (alexanderetz.com)  
9 points by EtzA 3 days ago | [discuss](#)
10. [▲ Measuring Word Significance using Distributed Representations of Words](#) (arxiv.org)  
2 points by benjaminwilson 1 day ago | [1 comment](#)
11. [▲ Analyzing Flight Data: A Gentle Introduction to Spark's GraphX](#) (sparktutorials.net)  
6 points by anabranh 3 days ago | [4 comments](#)
12. [▲ Data Science Blogs: A curated list](#) (github.com)  
23 points by coglog 6 days ago | [2 comments](#)
13. [▲ Frequentism and Bayesianism in Model Selection](#) (github.io)  
11 points by ebellm 5 days ago | [discuss](#)
14. [▲ Mapping NYC Taxi Data with Python](#) (danielforsyth.me)  
10 points by danielforsyth 5 days ago | [discuss](#)
15. [▲ 3 steps to trust outcome of regression](#) (goo.gl)  
8 points by mhfiroz 6 days ago | [discuss](#)
16. [▲ Tufte plots in R](#) (motioninsocial.com)  
14 points by tagyoureit 8 days ago | [discuss](#)
17. [▲ Getting Started with Spark DataFrames](#) (ipython.org)  
12 points by syrios12 7 days ago | [1 comment](#)

## RECAP

---

### LAST TIME:

I. WHAT IS MACHINE LEARNING?

II. MACHINE LEARNING PROBLEMS

III. CLASSIFICATION

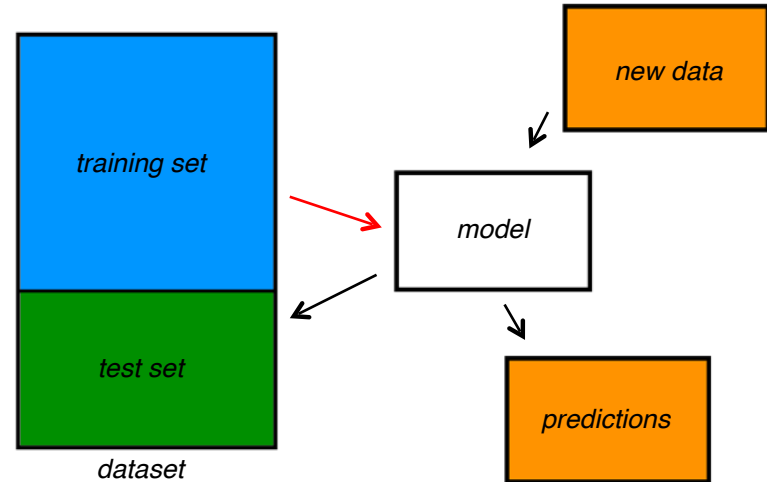
IV. BUILDING EFFECTIVE CLASSIFIERS

V. K-NEAREST NEIGHBORS

### EXERCISES:

VI. LAB: KNN CLASSIFICATION IN PYTHON

QUESTIONS?



---

**INTRO TO DATA SCIENCE**

---

# **QUESTIONS?**

**WHAT WAS THE MOST INTERESTING THING YOU LEARNT?**

**WHAT WAS THE HARDEST TO GRASP?**

# HOW'S THE HOMEWORK GOING?









**I. CROSS VALIDATION**

**II. INTRO TO PROBABILITY**

**III. NAÏVE BAYESIAN CLASSIFICATION**

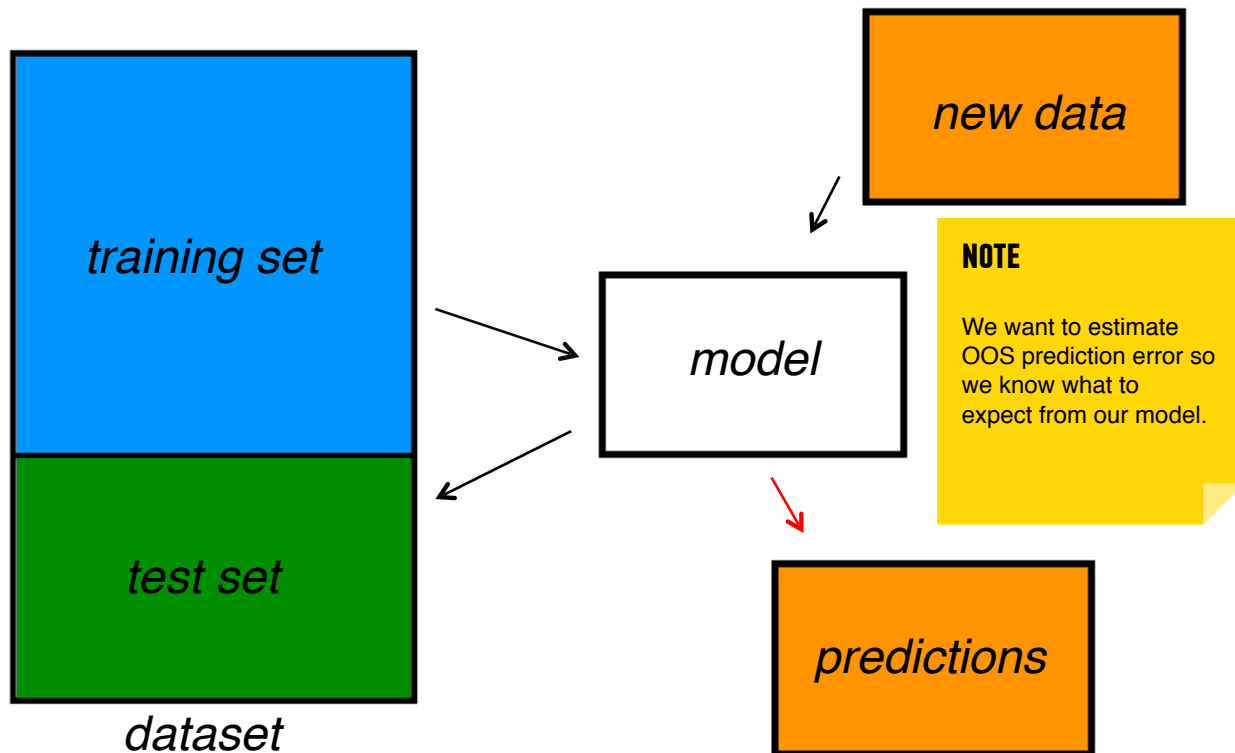
**EXERCISES:**

**IV. NAÏVE BAYES CLASSIFICATION IN PYTHON**

# **CROSS VALIDATION**

*Q: What types of prediction error will we run into?*

- 1) *training error*
- 2) *generalization error*
- 3) *OOS error*



*We can do better than that.... as we will see in the next class....*

*We can do better than that.... as we will see in the next class....*

*Let's recap briefly...*

*Suppose we do the train/test split.*

*Q: How well does generalization error predict OOS accuracy?*

*Suppose we do the train/test split.*

*Q: How well does generalization error predict OOS accuracy?*

*Thought experiment:*

*Suppose we had done a different train/test split.*

*Q: Would the generalization error remain the same?*

*A: Of course not!*

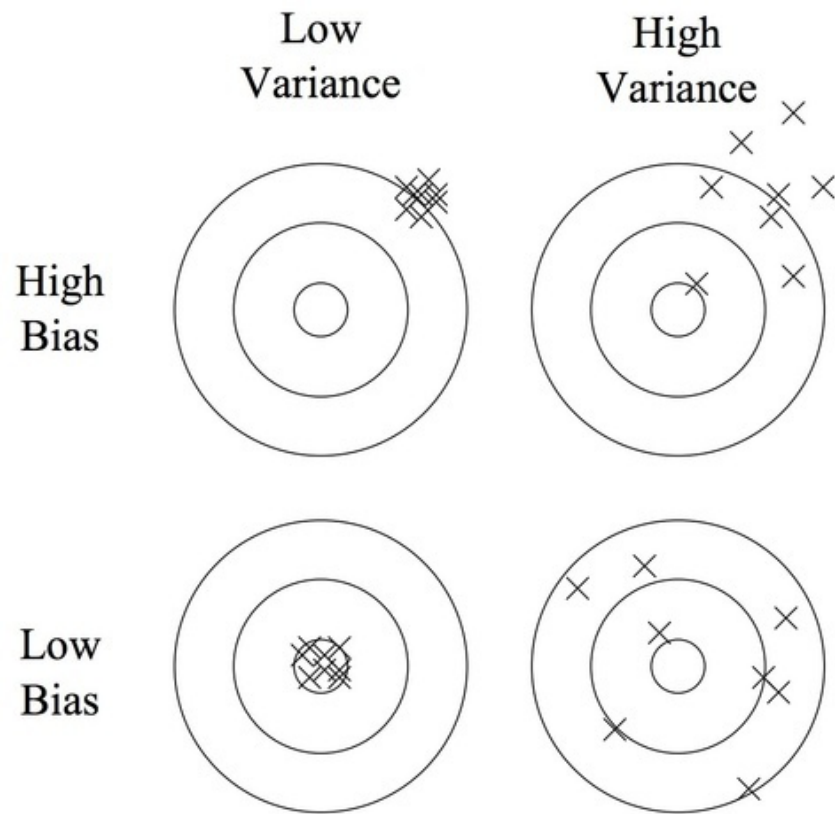
*A: On its own, not very well.*

**NOTE**

The generalization error gives a *high-variance estimate* of OOS accuracy.

# BIAS-VARIANCE

---





*Something is still missing!*

*Q: How can we do better?*

*Thought experiment:*

*Different train/test splits will give us different generalization errors.*

*Something is still missing!*

*Q: How can we do better?*

*Thought experiment:*

*Different train/test splits will give us different generalization errors.*

*Q: What if we did a bunch of these and took the average?*

*Something is still missing!*

*Q: How can we do better?*

*Thought experiment:*

*Different train/test splits will give us different generalization errors.*

*Q: What if we did a bunch of these and took the average?*

*A: Now you're talking!*

*Something is still missing!*

*Q: How can we do better?*

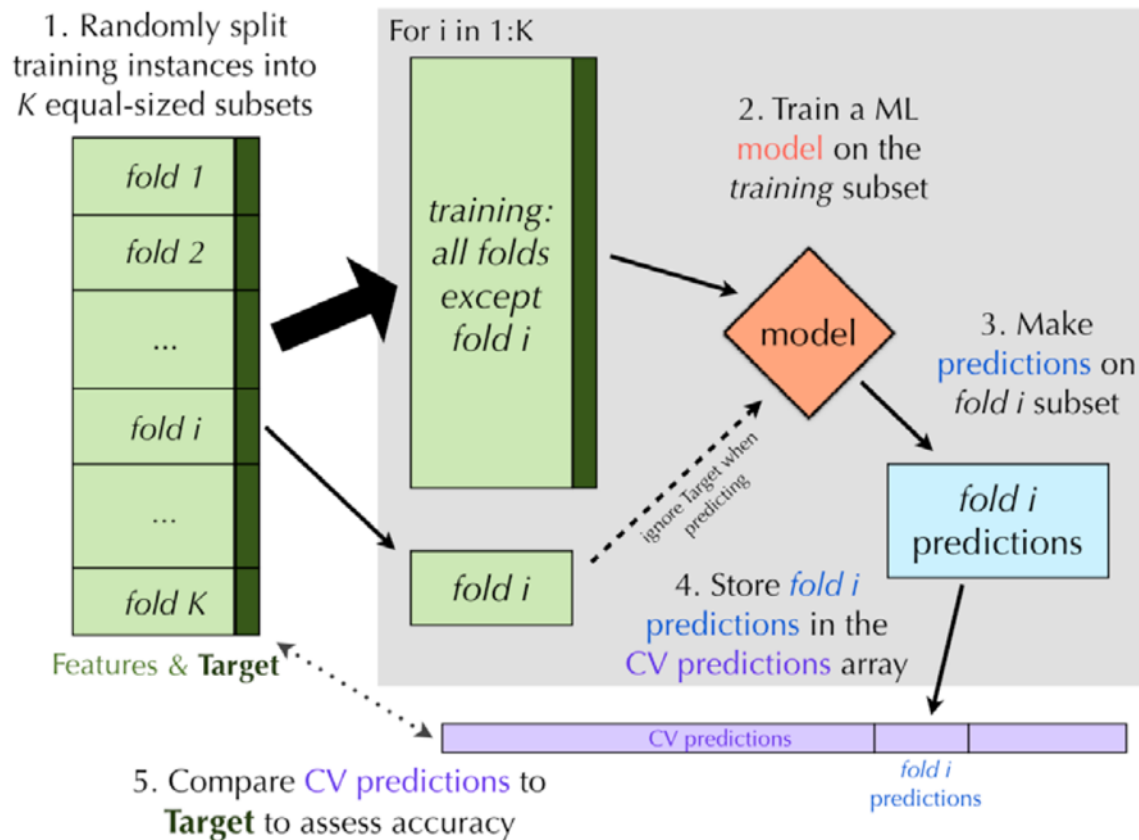
*Thought experiment:*

*Different train/test splits will give us different generalization errors.*

*Q: What if we did a bunch of these and took the average?*

*A: Now you're talking!*

*A: Cross-validation!*



Dataset	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	<u>Accuracy</u>
1	Test	Train	Train	Train	Train	$k_1 \%$
2	Train	Test	Train	Train	Train	$k_2 \%$
3	Train	Train	Test	Train	Train	$k_3 \%$
4	Train	Train	Train	Test	Train	$k_4 \%$
5	Train	Train	Train	Train	Test	$k_5 \%$

$$5\text{-Fold Generalization Error} = (k_1 + k_2 + k_3 + k_4 + k_5) / 5$$



*MORE ACCURATE*

Cross validation is a **more accurate** estimate of Out Of Sample (OOS) prediction error.



**MORE EFFICIENT**

Each record is used for both training and testing



10-fold CV is 10x more computationally expensive than a single train/test split

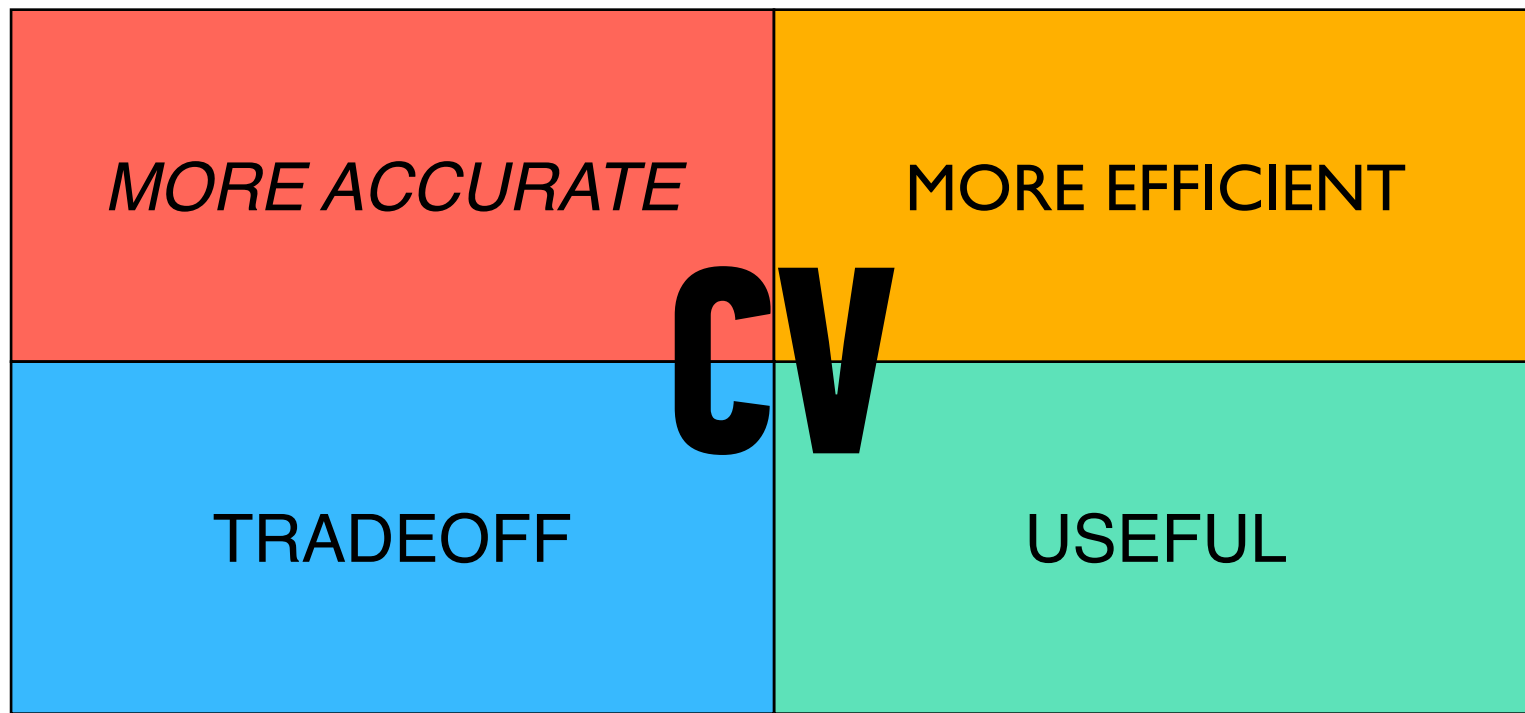


TRADEOFF

Cross validation average score can be used for  
model selection



USEFUL



*Can you think of some situations where running a cross validation could be problematic?*

**THINK - PAIR - SHARE**

---

**INTRO TO DATA SCIENCE**

---

**QUESTIONS?**

# **LAB: CROSS VALIDATION**

---

**INTRO TO DATA SCIENCE**

---

# **INTRO TO PROBABILITY**



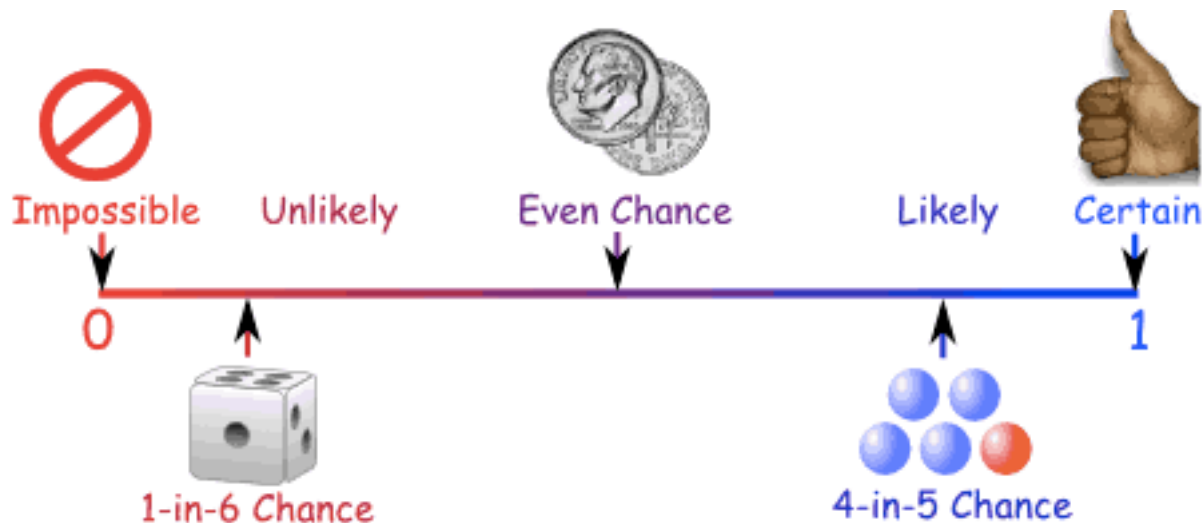
*BEWARE: EQUATIONS AHEAD !*



*Q: What is a probability?*

YOU TELL ME

*The probability  $p(A)$  for some event  $A$  is number between **0** and **1** that characterizes the **likelihood** the event  $A$  will occur.*



$\Omega$ 

the **SAMPLE SPACE** is the  
set of all possible events



EXAMPLES?

$\Omega$

the **SAMPLE SPACE** is the  
set of all possible events



$$p(\Omega) = ?$$

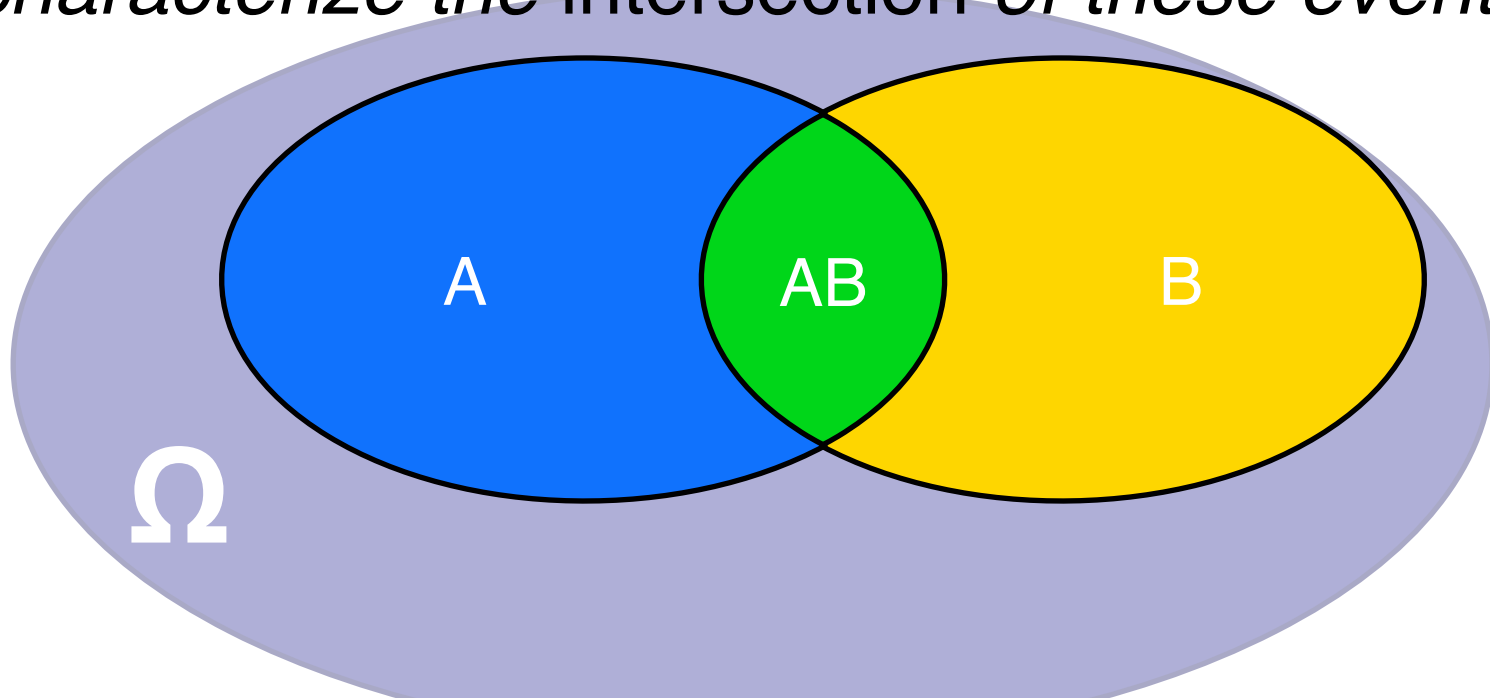
$\Omega$ 

the **SAMPLE SPACE** is the  
set of all possible events

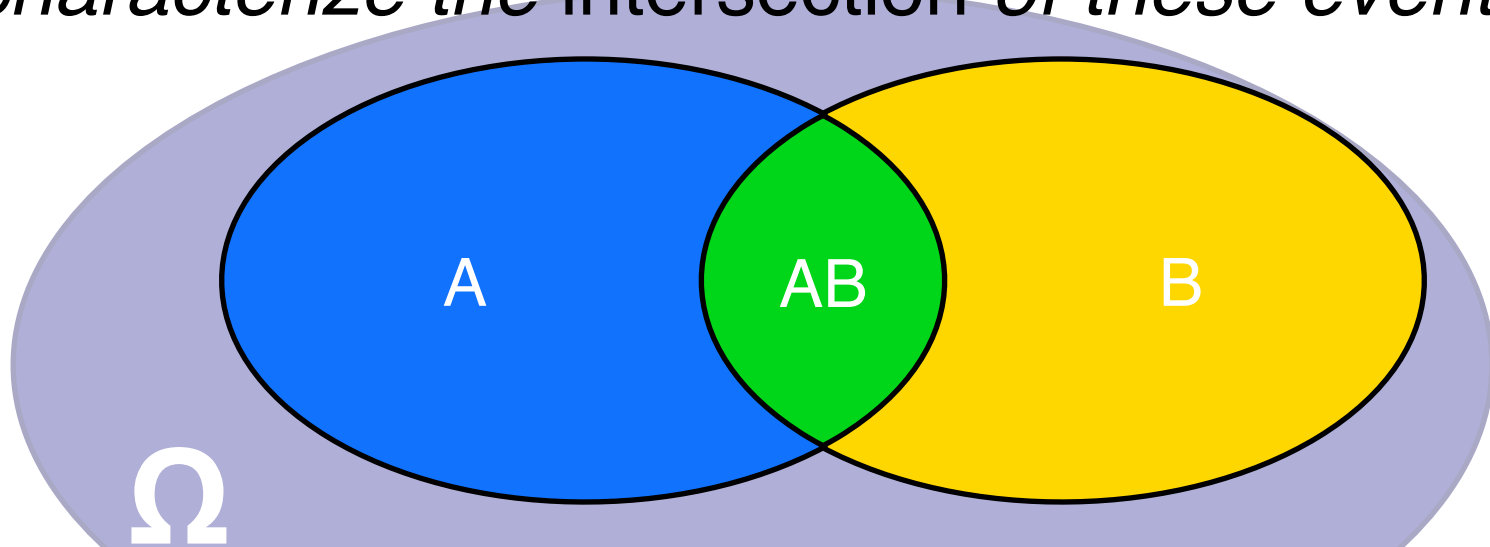


$$p(\Omega) = 1$$

*Q: Consider two events  $A$  &  $B$ . How can we characterize the intersection of these events?*

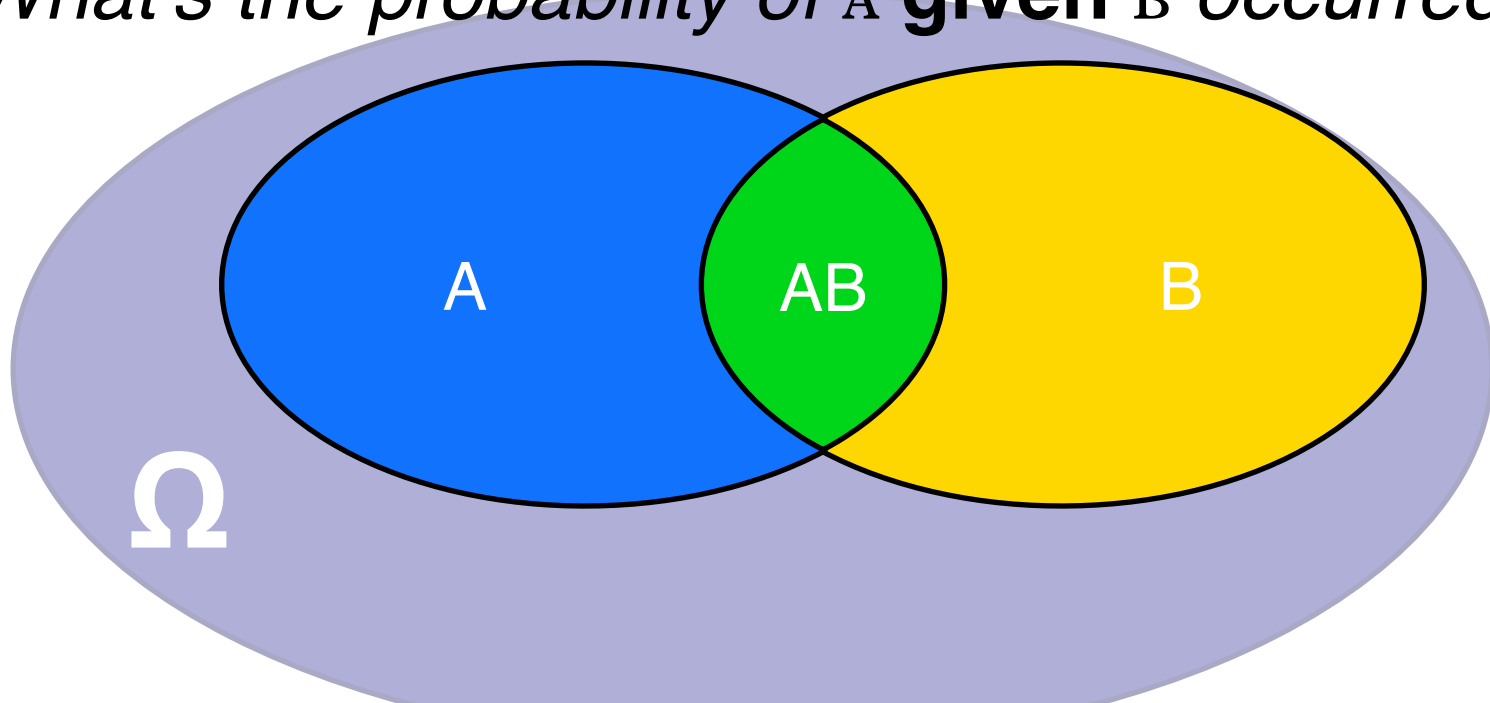


*Q: Consider two events  $A$  &  $B$ . How can we characterize the intersection of these events?*



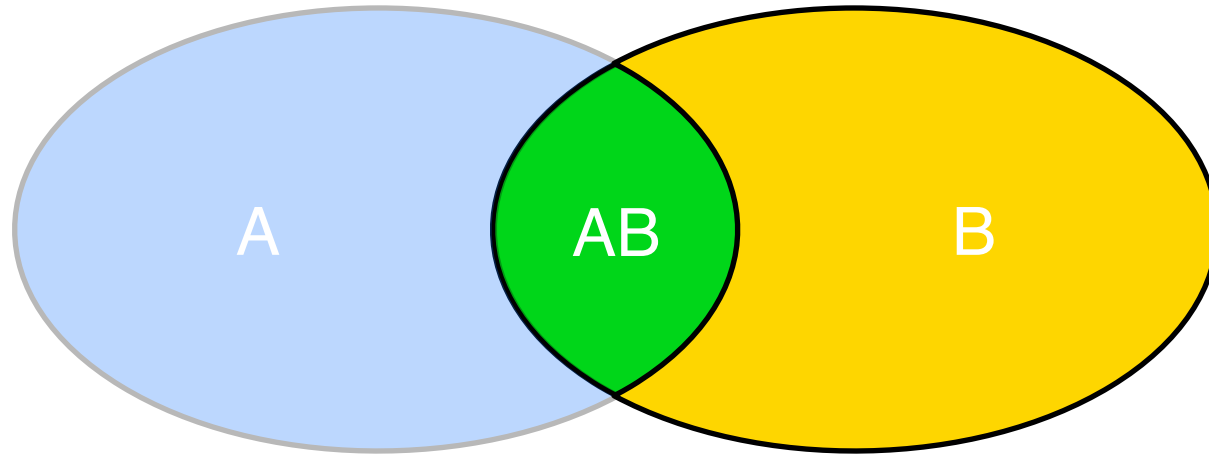
joint probability of  $A$  and  $B$ , written  $P(AB)$ .

*Suppose event  $B$  has occurred*  
*What's the probability of  $A$  **given**  $B$  occurred?*





*Suppose event  $B$  has occurred*  
*What's the probability of  $A$  **given**  $B$  occurred?*



**NOTE**

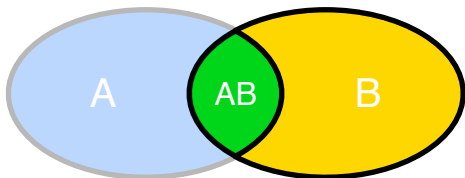
*This information about  $B$  transforms the sample space.*

*The intersection of  $A$  &  $B$  divided by region  $B$ .*

*Suppose event B has occurred*  
*What's the probability of A **given** B occurred?*

*This is called the **conditional probability***  
*of A given B*

*written*  $P(A|B) = P(AB) / P(B).$

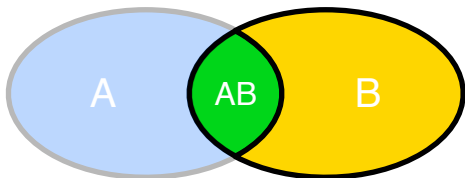


*Suppose event B has occurred*  
*What's the probability of A **given** B occurred?*

*This is called the **conditional probability***  
*of **A** given **B***

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

or  $P(A|B) P(B) = P(AB) \dots$

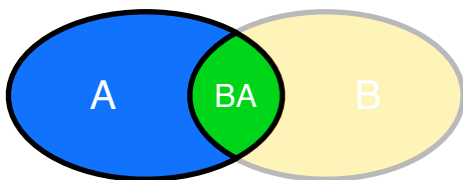


*Now let's ask the converse question:  
what is  $P(B|A)$  ?*

*This is called the **conditional probability**  
of **B** given **A***

*written*  $P(B|A) = P(BA) / P(A)$

or  $P(B|A) P(A) = P(BA) \dots$



*Let's recap*

*Let's recap*

$P(AB) = P(A|B) * P(B)$       *conditional probability of A given B*

*Let's recap*

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

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

*conditional probability of A given B  
by substitution*

*Let's recap*

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

*conditional probability of A given B*

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

*by substitution*

*But*  $P(AB) = P(BA)$

*since event AB = event BA*



*Let's recap*

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

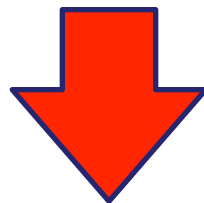
*conditional probability of A given B*

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

*by substitution*

*But*  $P(AB) = P(BA)$

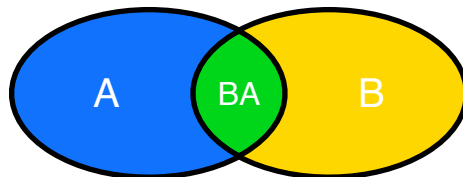
*since event AB = event BA*



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

*This result is called Bayes' theorem*

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



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

1% of women at age forty who participate in 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' THEOREM: EXAMPLE

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

A = breast cancer

B = positive mammogram

1% of women at age forty who participate in 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' THEOREM: EXAMPLE

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

A = breast cancer

B = positive mammogram

$$P(A) = 0.01$$

1% of women at age forty who participate in 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' THEOREM: EXAMPLE

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

A = breast cancer

B = positive mammogram

$$P(A) = 0.01$$

$$P(B|A) = 0.80$$

1% of women at age forty who participate in 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' THEOREM: EXAMPLE

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

A = breast cancer

B = positive mammogram

$$P(A) = 0.01$$

$$P(B|A) = 0.80$$

$$P(B|\sim A) = 0.096$$

1% of women at age forty who participate in 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' THEOREM: EXAMPLE

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

A = breast cancer

B = positive mammogram

1% of women at age forty who participate in 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?

$$P(A) = 0.01$$

$$P(B|A) = 0.80$$

$$P(B|\sim A) = 0.096$$

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

What is the probability that she actually has breast cancer?



## BAYES' THEOREM: EXAMPLE

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

A = breast cancer

B = positive mammogram

1% of women at age forty who participate in 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?

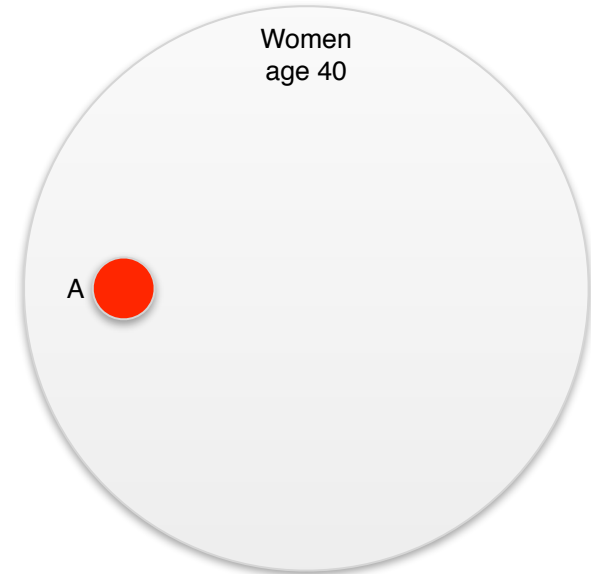
$$P(A) = 0.01$$

$$P(B|A) = 0.80$$

$$P(B|\sim A) = 0.096$$

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

What is the probability that she actually has breast cancer?



## BAYES' THEOREM: EXAMPLE

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

A = breast cancer

B = positive mammogram

1% of women at age forty who participate in 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?

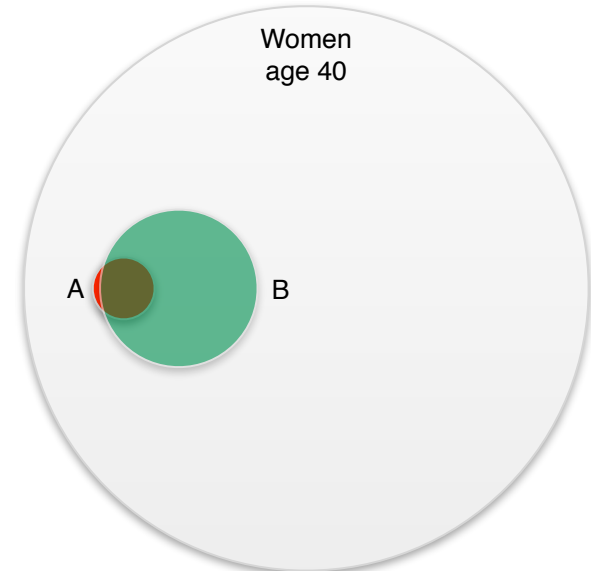
$$P(A) = 0.01$$

$$P(B|A) = 0.80$$

$$P(B|\sim A) = 0.096$$

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

What is the probability that she actually has breast cancer?



## BAYES' THEOREM: EXAMPLE

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

A = breast cancer

B = positive mammogram

1% of women at age forty who participate in 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?

$$P(A) = 0.01$$

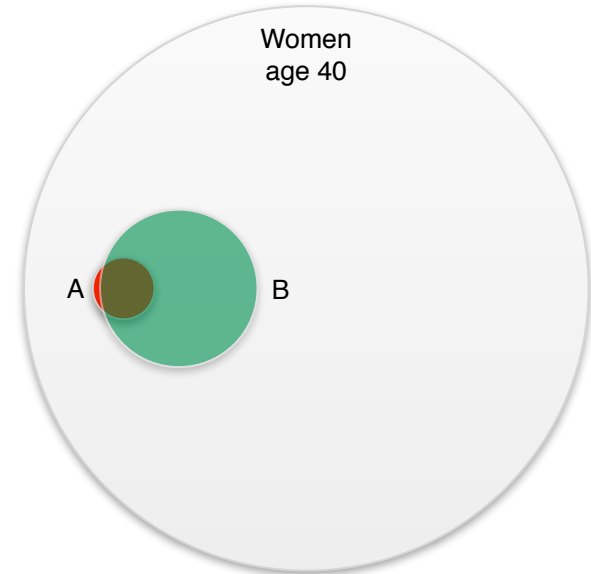
$$P(B|A) = 0.80$$

$$P(B|\sim A) = 0.096$$

$$P(B) = ?$$

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

What is the probability that she actually has breast cancer?



## BAYES' THEOREM: EXAMPLE

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

A = breast cancer

B = positive mammogram

1% of women at age forty who participate in 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?

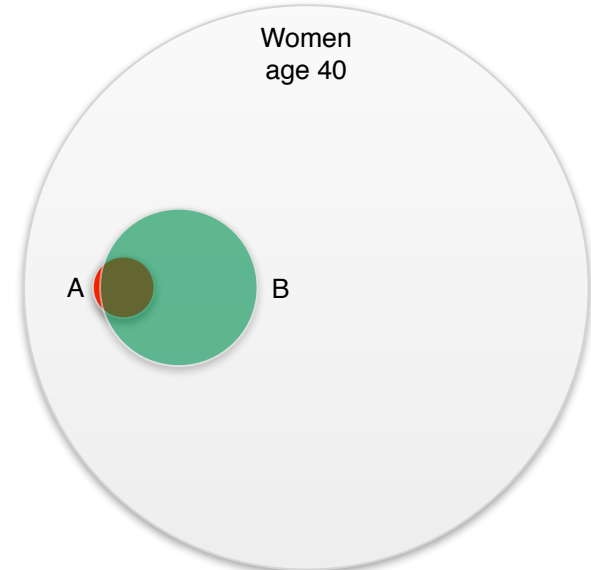
$$P(A) = 0.01$$

$$P(B|A) = 0.80$$

$$P(B|\sim A) = 0.096$$

$$\begin{aligned} P(B) &= P(B|A) * P(A) + P(B|\sim A) * P(\sim A) \\ &= 0.80 * 0.01 + 0.096 * 0.99 = 0.10304 \end{aligned}$$

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



What is the probability that she actually has breast cancer?

## BAYES' THEOREM: EXAMPLE

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

A = breast cancer

B = positive mammogram

1% of women at age forty who participate in 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?

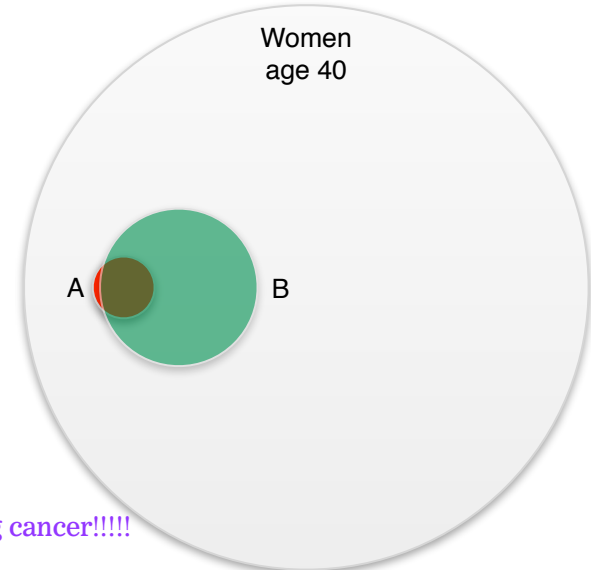
$$P(A) = 0.01$$

$$P(B|A) = 0.80$$

$$P(B|\sim A) = 0.096$$

$$\begin{aligned} P(B) &= P(B|A) * P(A) + P(B|\sim A) * P(\sim A) \\ &= 0.80 * 0.01 + 0.096 * 0.99 = 0.10304 \end{aligned}$$

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} = \frac{0.8 * 0.01}{0.10304} = 0.0776$$



What is the probability that she actually has breast cancer? About 7.8% chance of actually having cancer!!!!

# **INTERPRETATIONS OF PROBABILITY**

*There are 2 interpretations of probability:*

*There are 2 interpretations of probability:*

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



*There are 2 interpretations of probability:*

*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.*

# INDEPENDENT EVENTS

*Q: When are 2 events independent?*

*Q: When are 2 events independent?*

*A: Information about one does not affect the probability of the other.*

*Q: When are 2 events independent?*

*A: Information about one does not affect the probability of the other.*

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

*Q: When are 2 events independent?*

*A: Information about one does not affect the probability of the other.*

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

*using the definition of conditional probability:*

$$P(A|B) = P(AB) / P(B) = P(A) \rightarrow P(AB) = P(A) * P(B)$$

---

## ADDITIONAL RESOURCES

---

<http://www.yudkowsky.net/rational/bayes>

[https://en.wikipedia.org/wiki/Bayes%27\\_theorem](https://en.wikipedia.org/wiki/Bayes%27_theorem)

<http://betterexplained.com/articles/an-intuitive-and-short-explanation-of-bayes-theorem/>

<http://alexanderetz.com/2015/08/09/understanding-bayes-visualization-of-bf/>

<http://jakevdp.github.io/blog/2015/08/07/frequentism-and-bayesianism-5-model-selection/>

---

**EXPLAIN IN YOUR OWN WORDS**

---

What's the difference between frequentist and Bayesian interpretations of probability?

What does Bayes Theorem allow us to do?



# **NAÏVE BAYESIAN CLASSIFICATION**

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

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

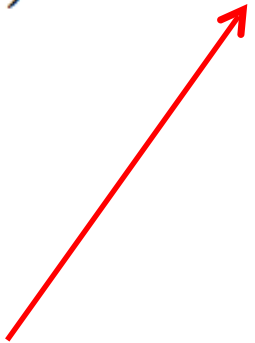
$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

*Bayes' theorem can help us to determine the probability of a record belonging to a class, given the data we observe.*

*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  $C$ .*

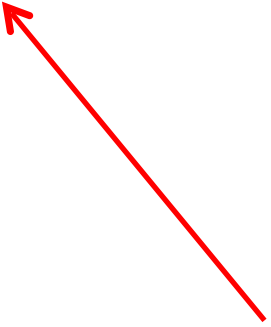
$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$


*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.*

*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 **prior probability** of C. It represents the probability of a record belonging to class C before the data is taken into account.*

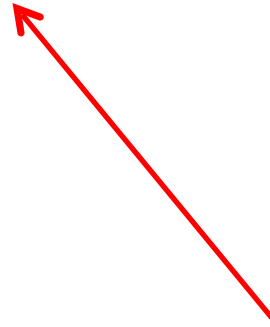
$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

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



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

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



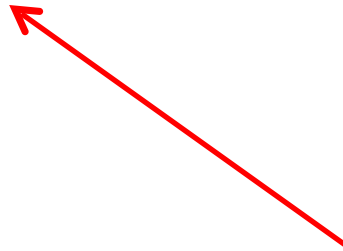
*This term is the **normalization constant**. It doesn't depend on  $C$ , and is generally ignored until the end of the computation.*

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

*The normalization constant doesn't tell us much.*

*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.*

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



*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?*

$$P(\{\mathbf{x}_i\} | C) = P(\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\} | C)$$

*Remember the likelihood function?*

$$P(\{\mathbf{x}_i\} | C) = P(\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\} | C)$$

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



## NAÏVE BAYESIAN 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.*

*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:*

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

*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:*

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

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

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

*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.*

***Advantages:***

- *Fast to train (single scan). Fast to classify*
- *Not sensitive to irrelevant features*
- *Handles real and discrete data*
- *Handles streaming data well*

***Disadvantages:***

- *Assumes independence of features*

# **LAB**

## **IV. NAIVE BAYESIAN CLASSIFICATION**