Naive Bayes Text Classification

Sandeep Avula
asandeep@live.unc.edu

Outline

Basic Probability and Notation

Bayes Law and Naive Bayes Classification

Smoothing

Class Prior Probabilities

Naive Bayes Classification

Summary

Crash Course in Basic Probability

Discrete Random Variable

- A is a discrete random variable if:
  - A describes an event with a finite number of possible outcomes (discrete vs continuous)
  - A describes and event whose outcomes have some degree of uncertainty (random vs. pre-determined)

Discrete Random Variables

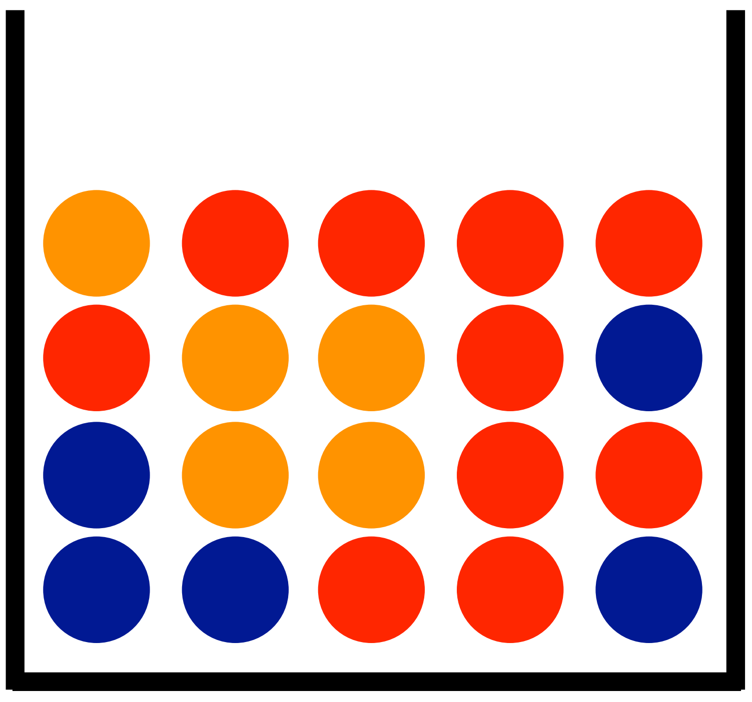
Examples

- A = the outcome of a coin-flip
  - outcomes: heads, tails
- A = it will rain tomorrow
  - outcomes: rain, no rain
- A = you have the flu
  - outcomes: flu, no flu
- A = your final grade in this class
  - outcomes: F, L, P, H

Discrete Random Variables

Examples

- A = the color of a ball pulled out from this bag
  - outcomes: RED, BLUE, ORANGE

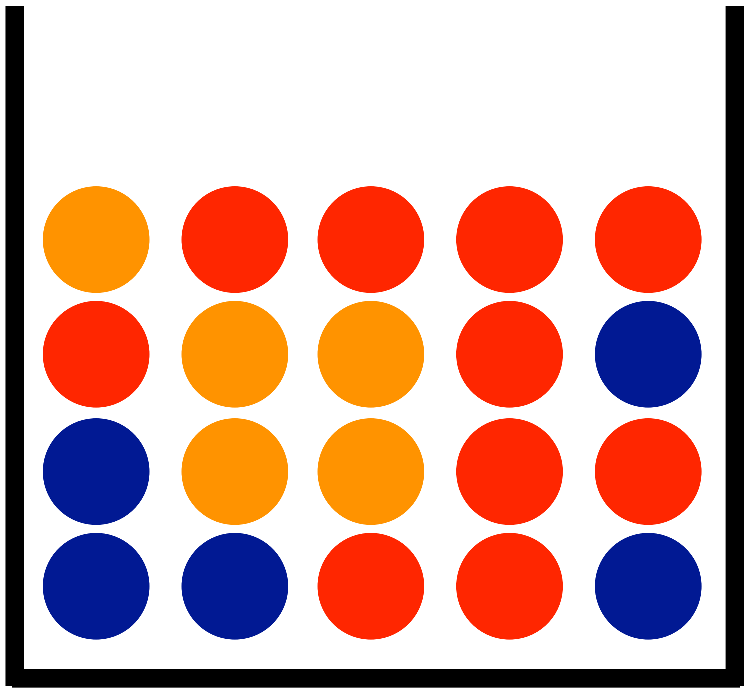


Probabilities

- Let P(A=X) denote the probability that the outcome of event A equals X
- For simplicity, we often express P(A=X) as P(X)
- Ex: P(RAIN), P(NO RAIN), P(FLU), P(NO FLU), ...

Probability Distribution

- A probability distribution gives the probability of each possible outcome of a random variable
- P(RED) = probability of pulling out a red ball
- P(BLUE) = probability of pulling out a blue ball
- P(ORANGE) = probability of pulling out an orange ball



Probability Distribution

- For it to be a probability distribution, two conditions must be satisfied:
  - the probability assigned to each possible outcome must be between 0 and 1 (inclusive)
  - the sum of probabilities assigned to all outcomes must equal 1

$$0 \leq P(\text{RED}) \leq 1$$

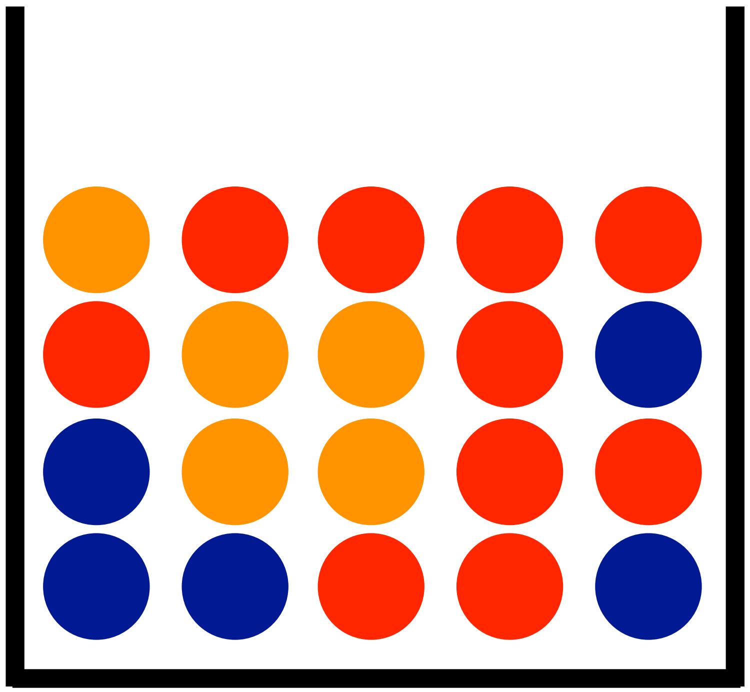
$$0 \leq P(\text{BLUE}) \leq 1$$

$$0 \leq P(\text{ORANGE}) \leq 1$$

$$P(\text{RED}) + P(\text{BLUE}) + P(\text{ORANGE}) = 1$$

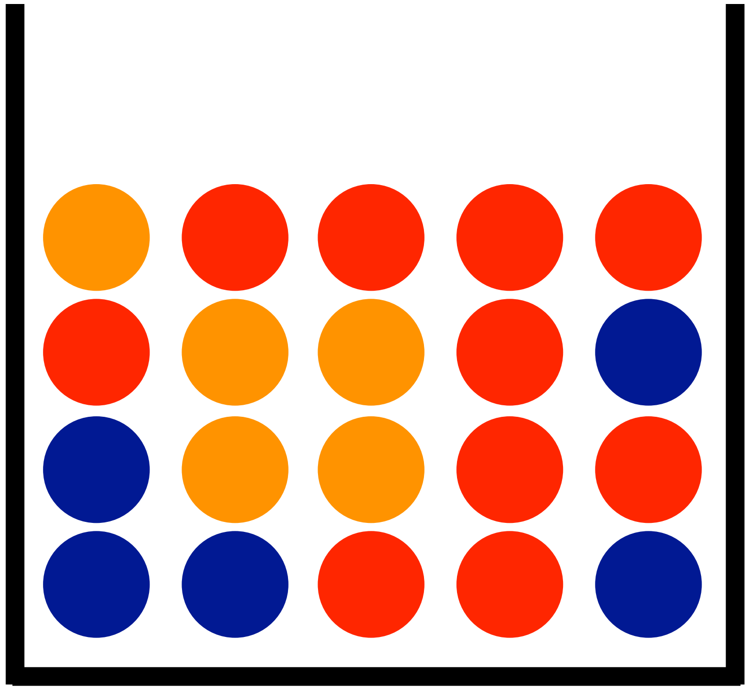
Probability Distribution
Estimation

- Let's estimate these probabilities based on what we know about the contents of the bag
- P(RED) = ?
- P(BLUE) = ?
- P(ORANGE) = ?



Probability Distribution
estimation

- Let's estimate these probabilities based on what we know about the contents of the bag



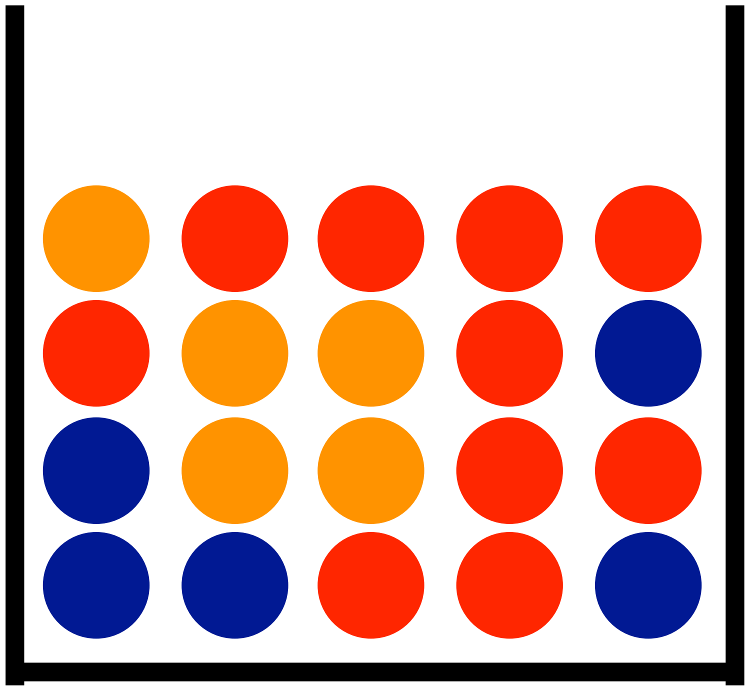
- P(RED) = 10/20 = 0.5
- P(BLUE) = 5/20 = 0.25
- P(ORANGE) = 5/20 = 0.25
- P(RED) + P(BLUE) + P(ORANGE) = 1.0

Probability Distribution

assigning probabilities to outcomes

- Given a probability distribution, we can assign probabilities to different outcomes
- I reach into the bag and pull out an orange ball. What is the probability of that happening?
- I reach into the bag and pull out two balls: one red, one blue. What is the probability of that happening?
- What about three orange balls?

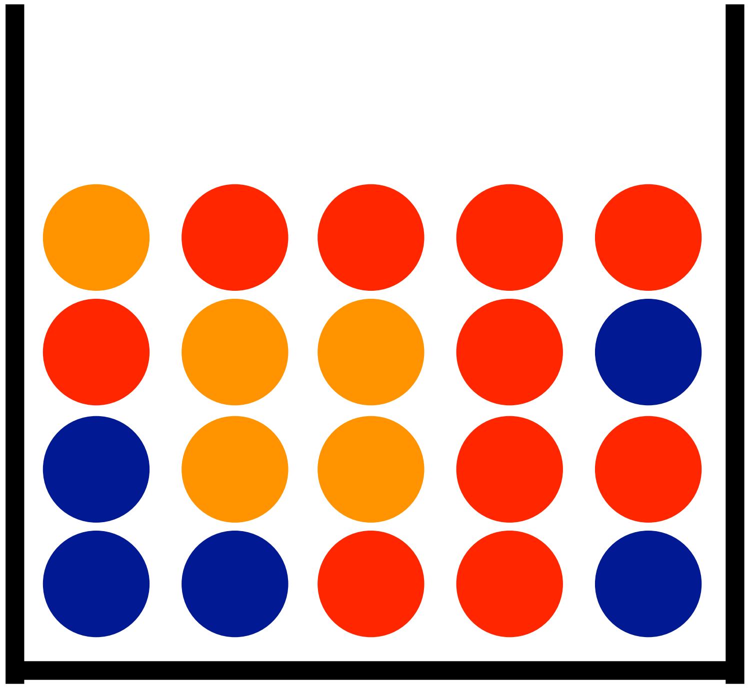
P(RED) = 0.5
P(BLUE) = 0.25
P(ORANGE) = 0.25



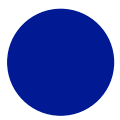
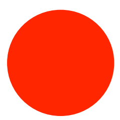
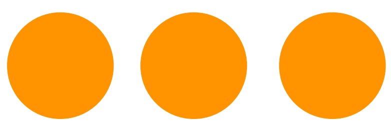
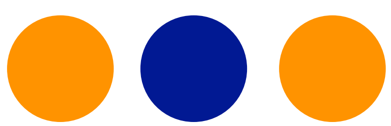
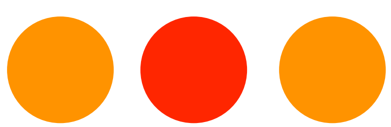
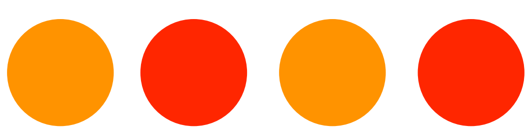
What can we do with a probability distribution?

- If we assume that each outcome is independent of previous outcomes, then the probability of a sequence of outcomes is calculated by multiplying the individual probabilities
- Note: we're assuming that when you take out a ball, you put it back in the bag before taking another

P(RED) = 0.5
P(BLUE) = 0.25
P(ORANGE) = 0.25



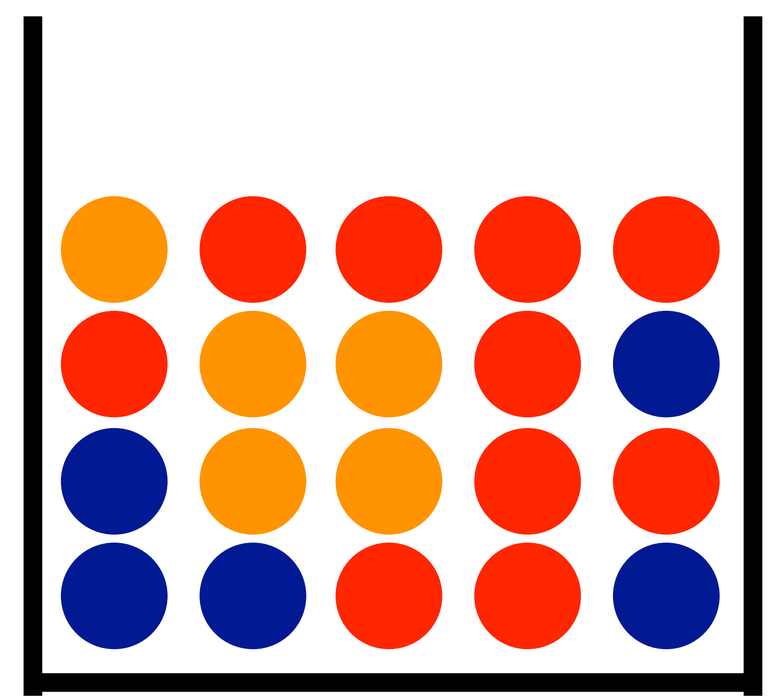
What can we do with a probability distribution?

- P() = ??
- P() = ??
- P() = ??
- P() = ??
- P() = ??
- P() = ??

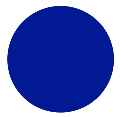
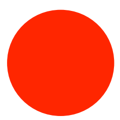
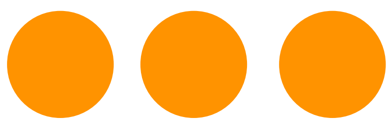
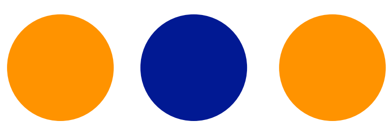
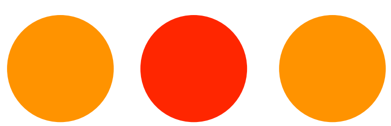
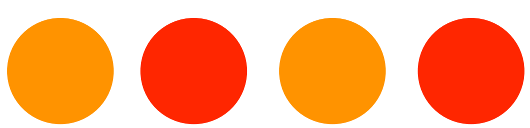
P(RED) = 0.5

P(BLUE) = 0.25

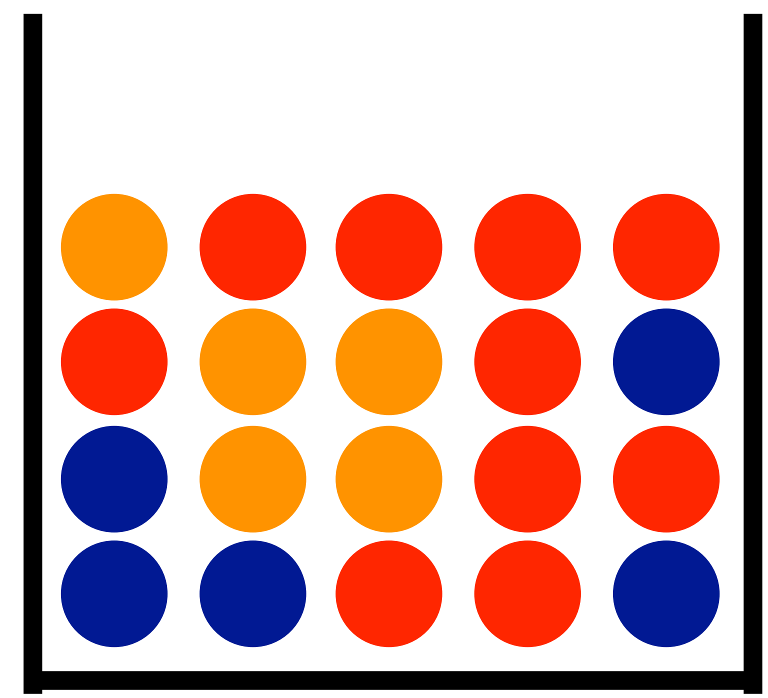
P(ORANGE) = 0.25



What can we do with a probability distribution?

- P() = 0.25
- P() = 0.5
- P() = 0.25 x 0.25 x 0.25
- P() = 0.25 x 0.25 x 0.25
- P() = 0.25 x 0.50 x 0.25
- P() = 0.25 x 0.50 x 0.25 x 0.50

P(RED) = 0.5
P(BLUE) = 0.25
P(ORANGE) = 0.25

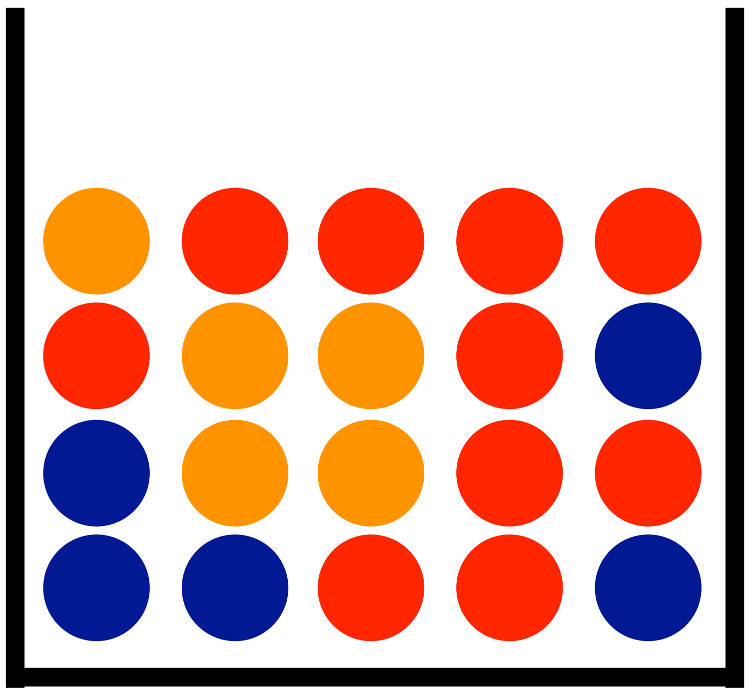


Conditional Probability

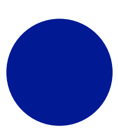
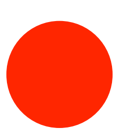
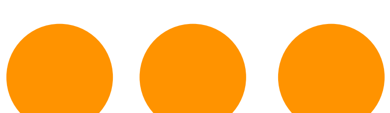
- P(A,B): the probability that event A <u>and</u> event B both occur
- P(A|B): the probability of event A occurring given prior knowledge that event B occurred

Conditional Probability

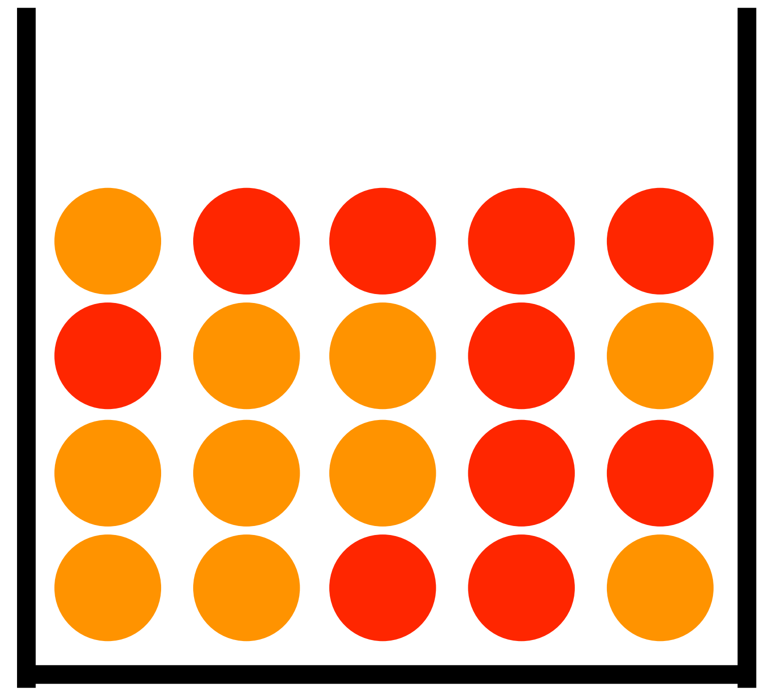
P(RED) = 0.50
P(BLUE) = 0.25
P(ORANGE) = 0.25



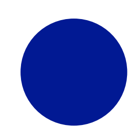
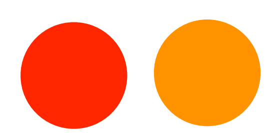
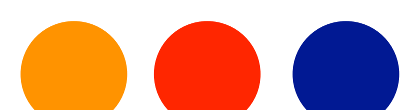
A

- P( | A) = ??
- P( | A) = ??
- P( | A) = ??

P(RED) = 0.50
P(BLUE) = 0.00
P(ORANGE) = 0.50

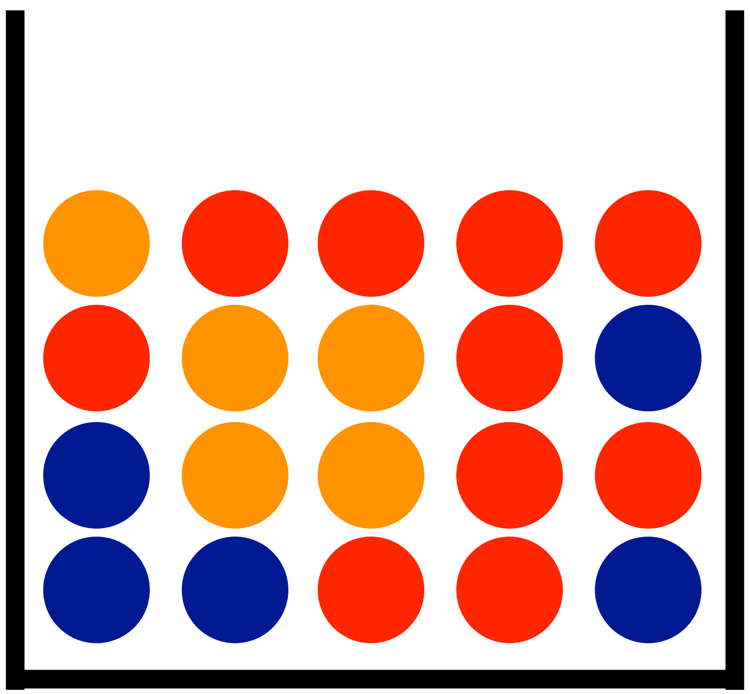


B

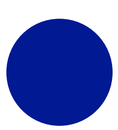
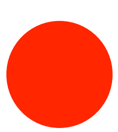
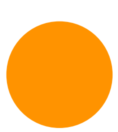
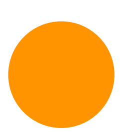
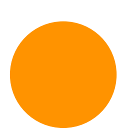
- P( | B) = ??
- P( | B) = ??
- P( | B) = ??

Conditional Probability

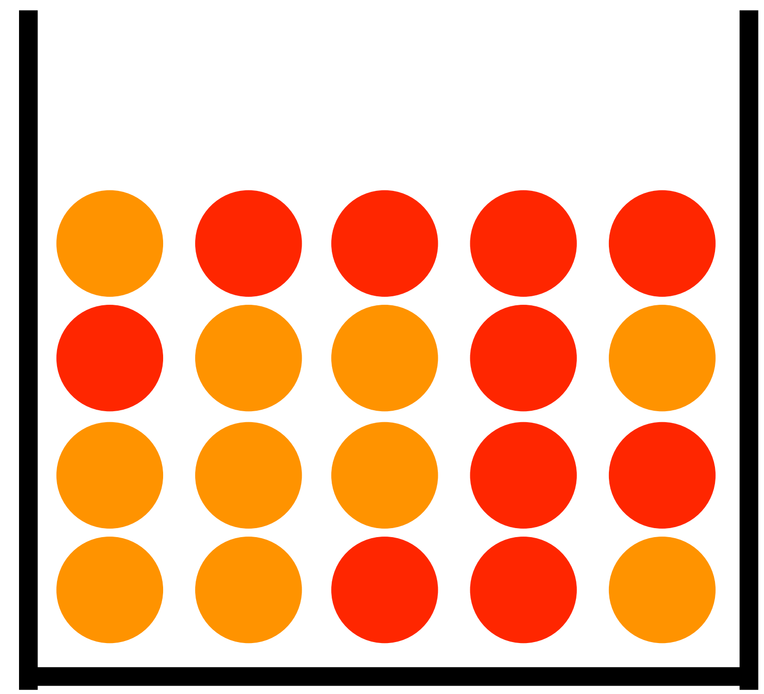
P(RED) = 0.50
P(BLUE) = 0.25
P(ORANGE) = 0.25



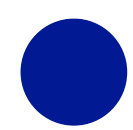
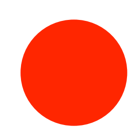
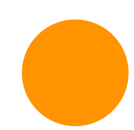
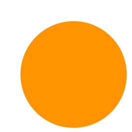
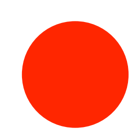
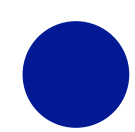
A

- P( | A) = 0.25
- P( | A) = 0.50
- P(   | A) = 0.016

P(RED) = 0.50
P(BLUE) = 0.00
P(ORANGE) = 0.50



B

- P( | B) = 0.00
- P(  | B) = 0.25
- P(   | B) = 0.00

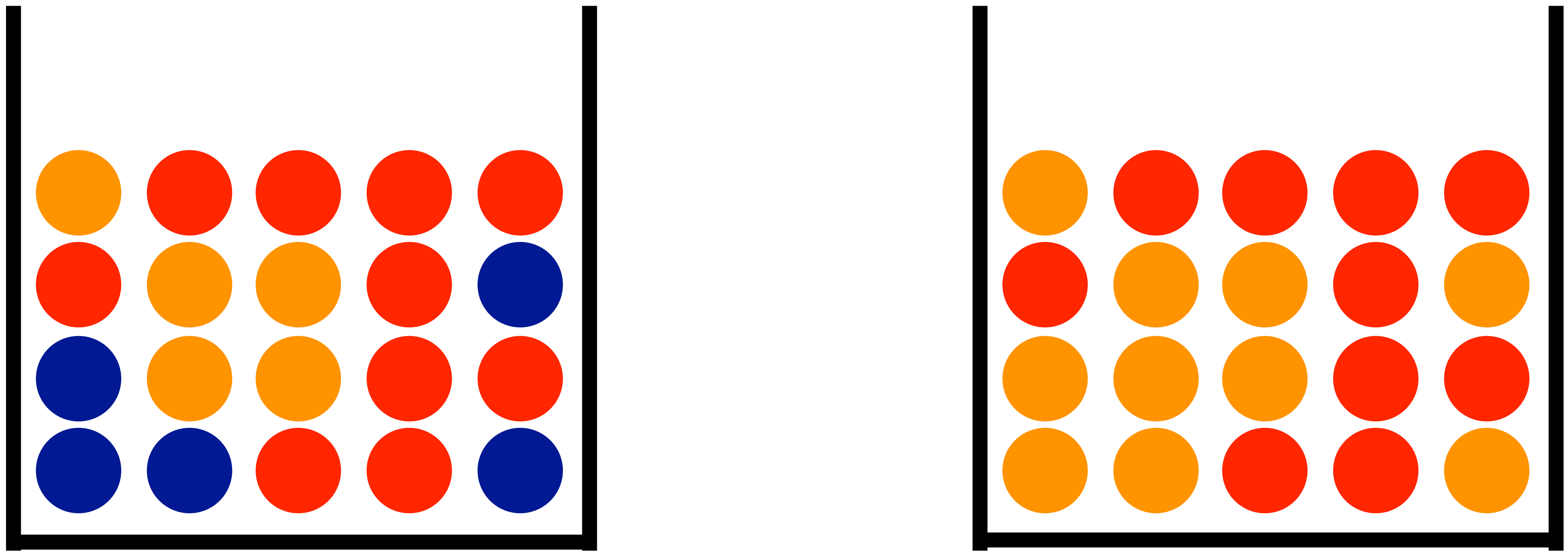
Chain Rule

- $P(A, B) = P(A|B) \times P(B)$
- Example:
  - probability that it will rain today (**B**) <u>and</u> tomorrow (**A**)
  - probability that it will rain today (**B**)
  - probability that it will rain tomorrow (**A**) given that it will rain today (**B**)

Independence

- $P(A, B) = P(A|B) \times P(B) = P(A) \times P(B)$
- Example:
  - probability that it will rain today (**B**) <u>and</u> tomorrow (**A**)
  - probability that it will rain today (**B**)
  - probability that it will rain tomorrow (**A**) given that it will rain today (**B**)
  - probability that it will rain tomorrow (A)

Independence

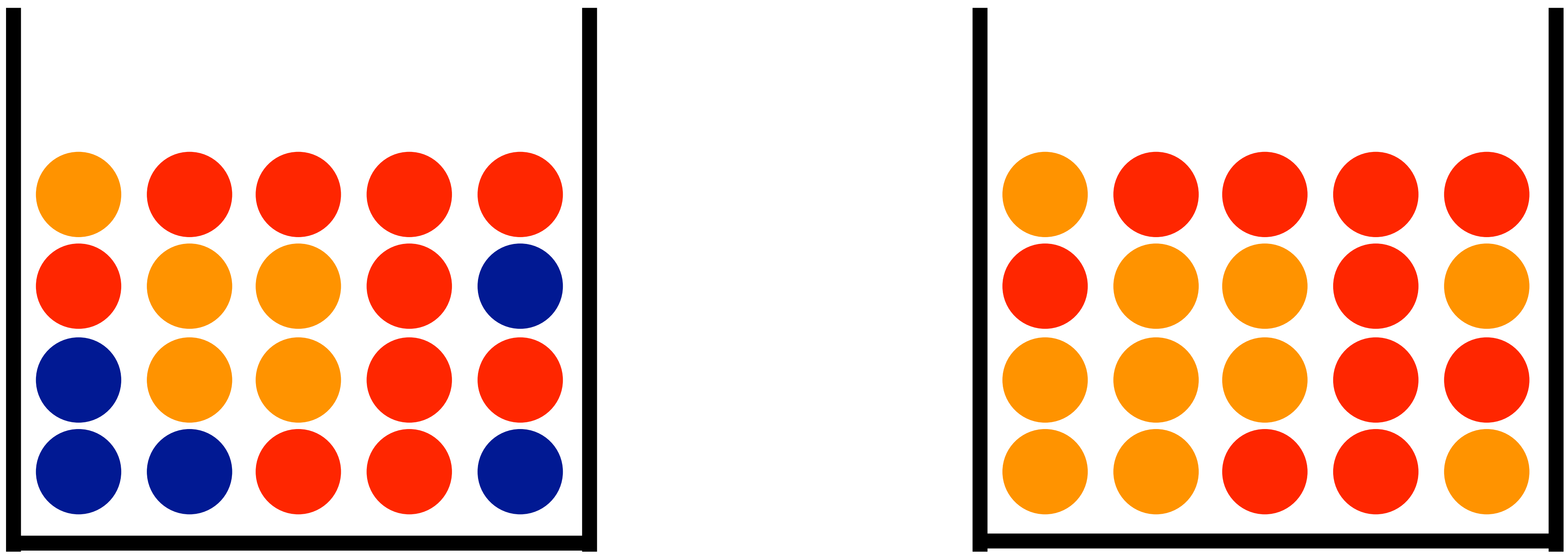


A

B

P(● | A) ?= P(●)

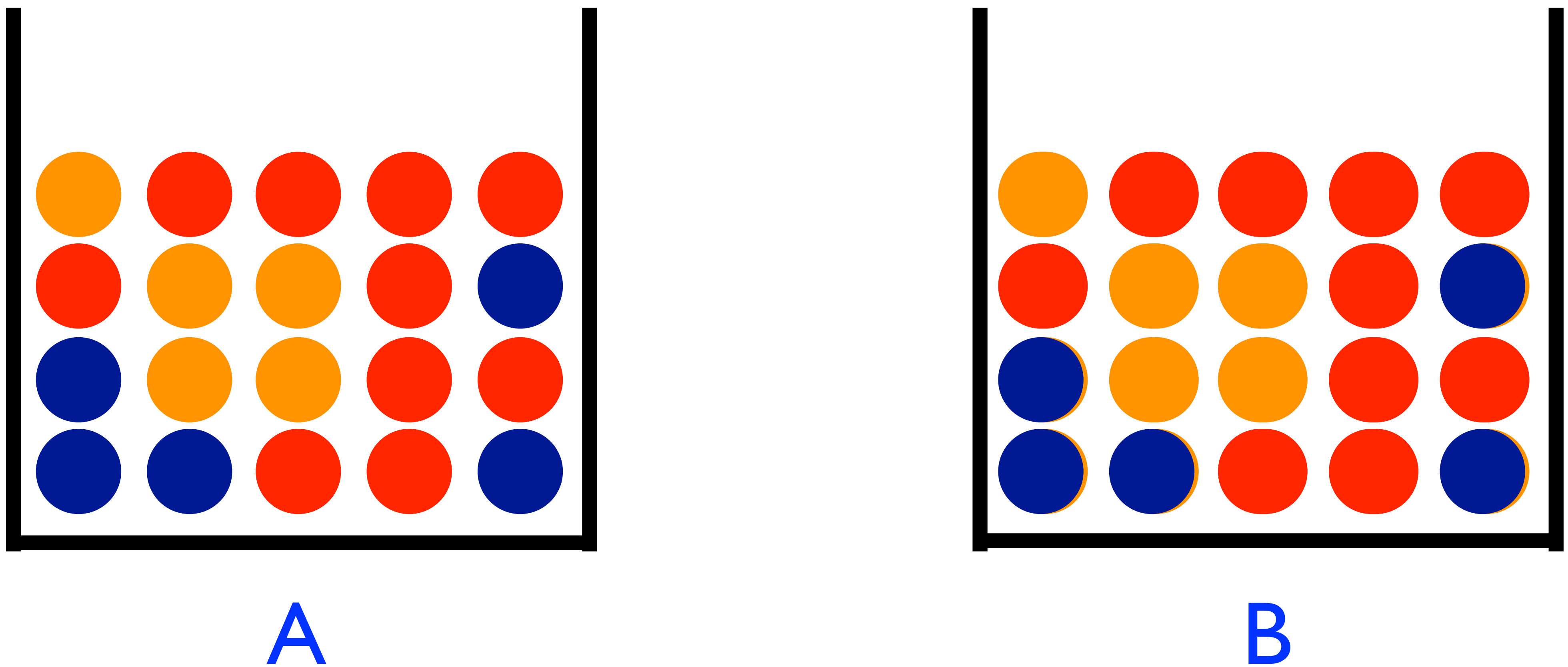
Independence



A B

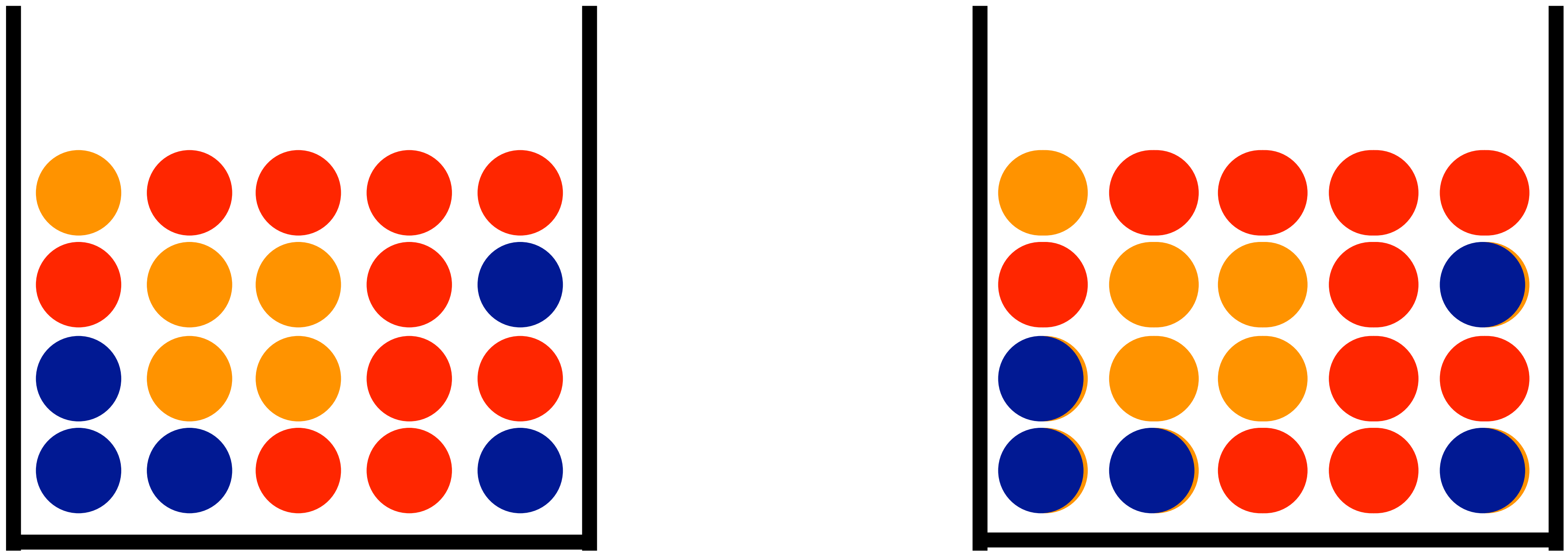
$P(\bullet \mid A) > P(\bullet)$

Independence



P(● | A) ?= P(●)

Independence



A B

$$P(\bullet \mid A) = P(\bullet)$$

Outline

Basic Probability and Notation

Bayes Law and Naive Bayes Classification

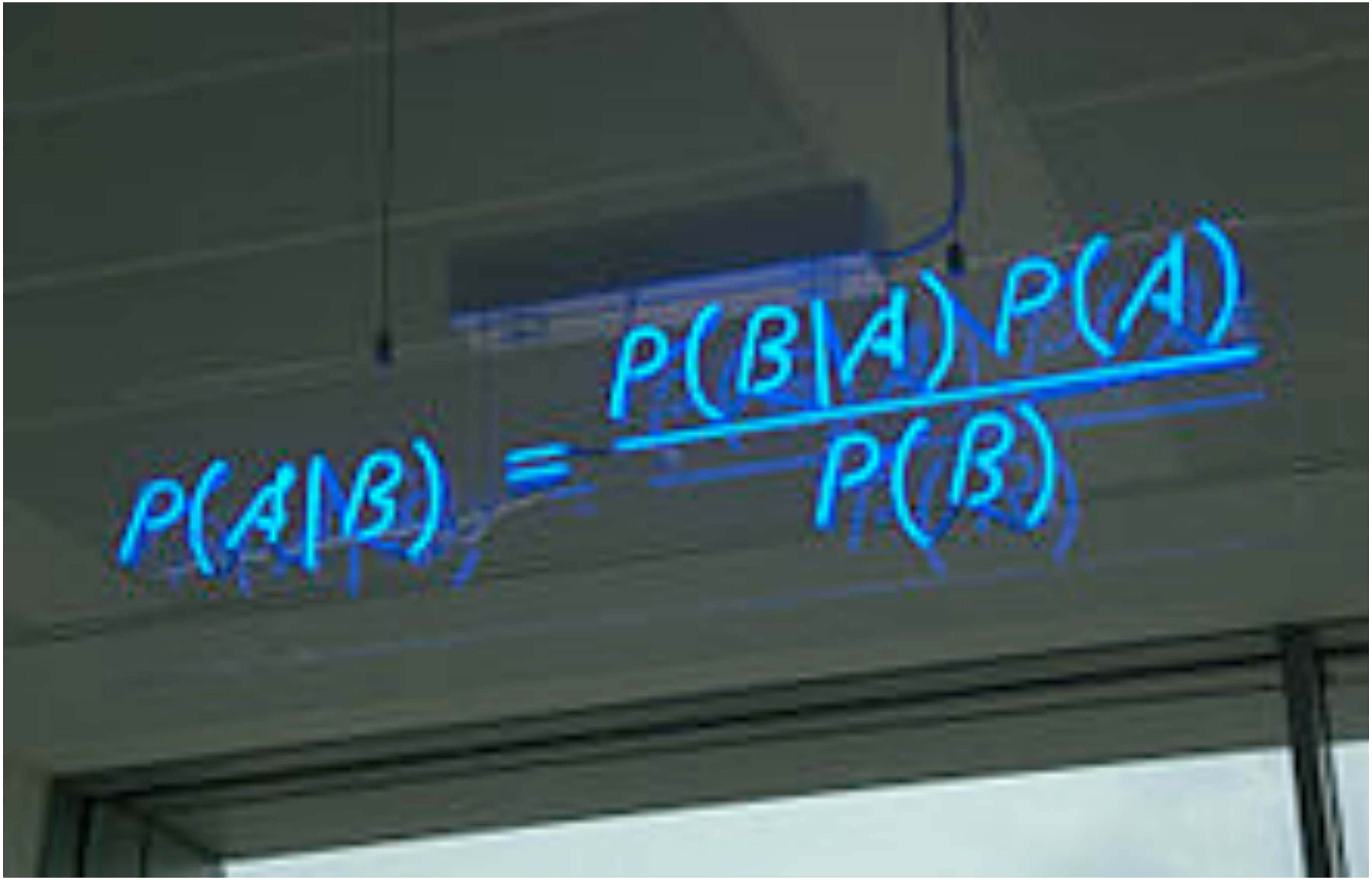
Smoothing

Class Prior Probabilities

Naive Bayes Classification

Summary

Bayes' Law



Bayes' Law

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

Bayes' Law

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

How do we get this?

Derivation of Bayes' Law

$$P(A,B) = P(A,B)$$

Always true!

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

Chain Rule!

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

Divide both sides by P(B)!

Naive Bayes Classification

example: positive/negative movie reviews

Bayes Rule

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

Confidence of POS prediction given instance D

$$P(POS|D) = \frac{P(D|POS) \times P(POS)}{P(D)}$$

Confidence of NEG prediction given instance D

$$P(NEG|D) = \frac{P(D|NEG) \times P(NEG)}{P(D)}$$

Naive Bayes Classification

example: positive/negative movie reviews

- Given instance D, predict positive (POS) if:

$$P(POS|D) \geq P(NEG|D)$$

- Otherwise, predict negative (NEG)

Naive Bayes Classification

example: positive/negative movie reviews

- Given instance D, predict positive (POS) if:

$$\frac{P(D|POS) \times P(POS)}{P(D)} \geq \frac{P(D|NEG) \times P(NEG)}{P(D)}$$

- Otherwise, predict negative (NEG)

Naive Bayes Classification

example: positive/negative movie reviews

- Given instance D, predict positive (POS) if:

$$\frac{P(D|POS) \times P(POS)}{P(D)} \geq \frac{P(D|NEG) \times P(NEG)}{P(D)}$$

- Otherwise, predict negative (NEG)

Are these necessary?

Naive Bayes Classification

example: positive/negative movie reviews

- Given instance D, predict positive (POS) if:

$$P(D|POS) \times P(POS) \geq P(D|NEG) \times P(NEG)$$

- Otherwise, predict negative (NEG)

Naive Bayes Classification

example: positive/negative movie reviews

- Our next goal is to estimate these parameters from the training data!

- P(NEG) = ??
- P(POS) = ??
- P(D|NEG) = ??
- P(D|POS) = ??

Easy!

Not so easy!

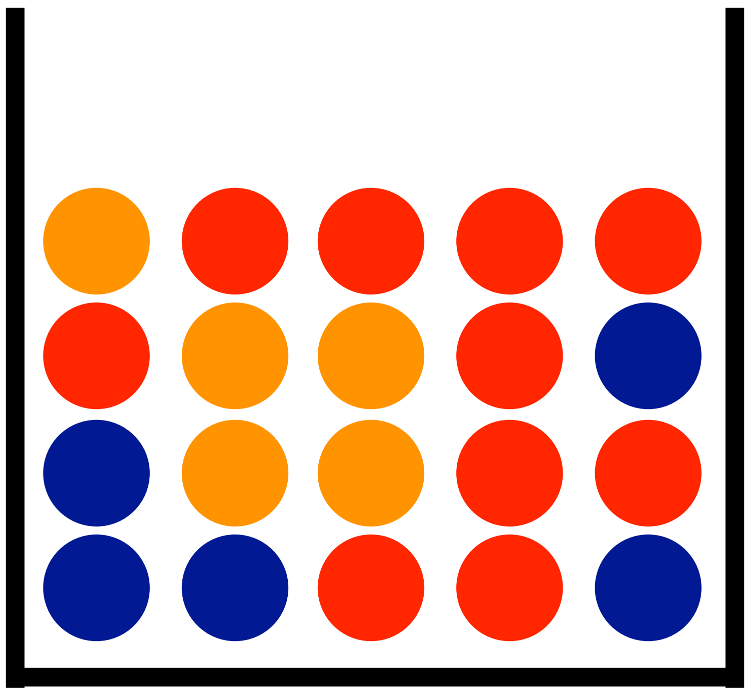
Naive Bayes Classification

example: positive/negative movie reviews

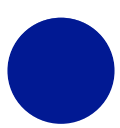
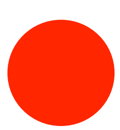
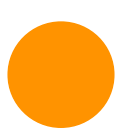
- Our next goal is to estimate these parameters from the training data!
- P(NEG) = % of training set documents that are NEG
- P(POS) = % of training set documents that are POS
- P(D|NEG) = ??
- P(D|POS) = ??

Remember Conditional Probability?

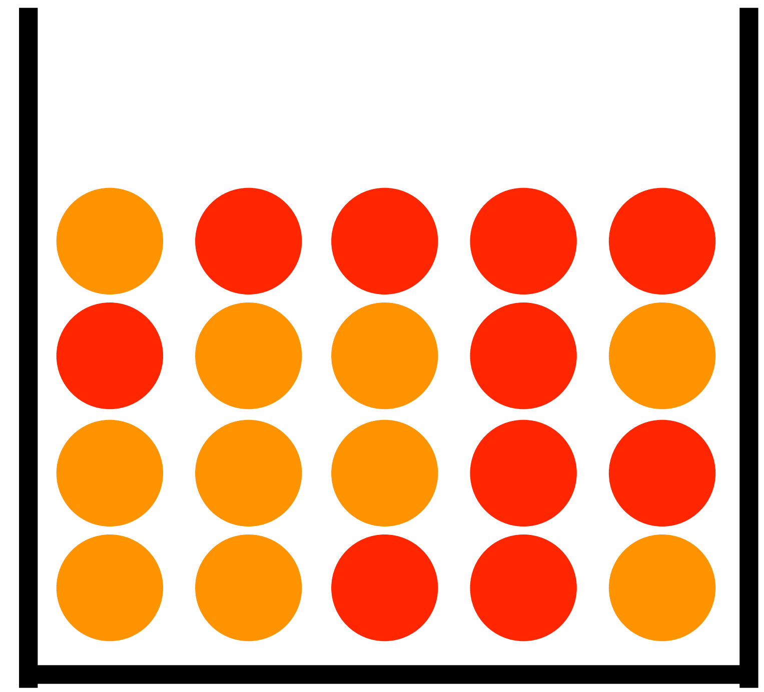
P(RED) = 0.50
P(BLUE) = 0.25
P(ORANGE) = 0.25



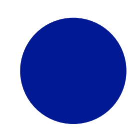
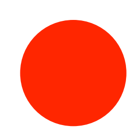
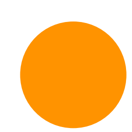
A

- P( | A) = 0.25
- P( | A) = 0.50
- P( | A) = 0.25

P(RED) = 0.50
P(BLUE) = 0.00
P(ORANGE) = 0.50

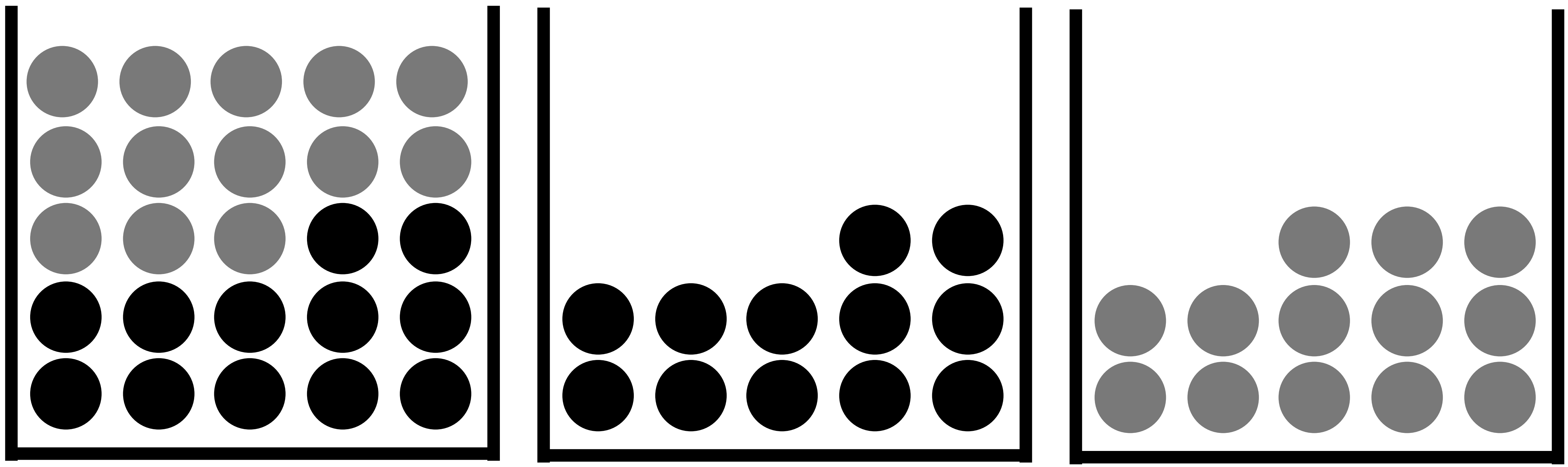


B

- P( | B) = 0.00
- P( | B) = 0.50
- P( | B) = 0.50

Naive Bayes Classification

example: positive/negative movie reviews



Training Instances

Positive Training Instances

Negative Training Instances

P(D|POS) = ??   P(D|NEG) = ??

Naive Bayes Classification

example: positive/negative movie reviews

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	...	w_n	sentiment
1	0	1	0	1	0	0	1	...	0	positive
0	1	0	1	1	0	1	1	...	0	positive
0	1	0	1	1	0	1	0	...	0	positive
0	0	1	0	1	1	0	1	...	1	positive
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	...	⋮	⋮
1	1	0	1	1	0	0	1	...	1	positive

Naive Bayes Classification

example: positive/negative movie reviews

- Our next goal is to estimate these parameters from the training data!

- P(NEG) = ??
- P(POS) = ??
- P(D|NEG) = ??
- P(D|POS) = ??

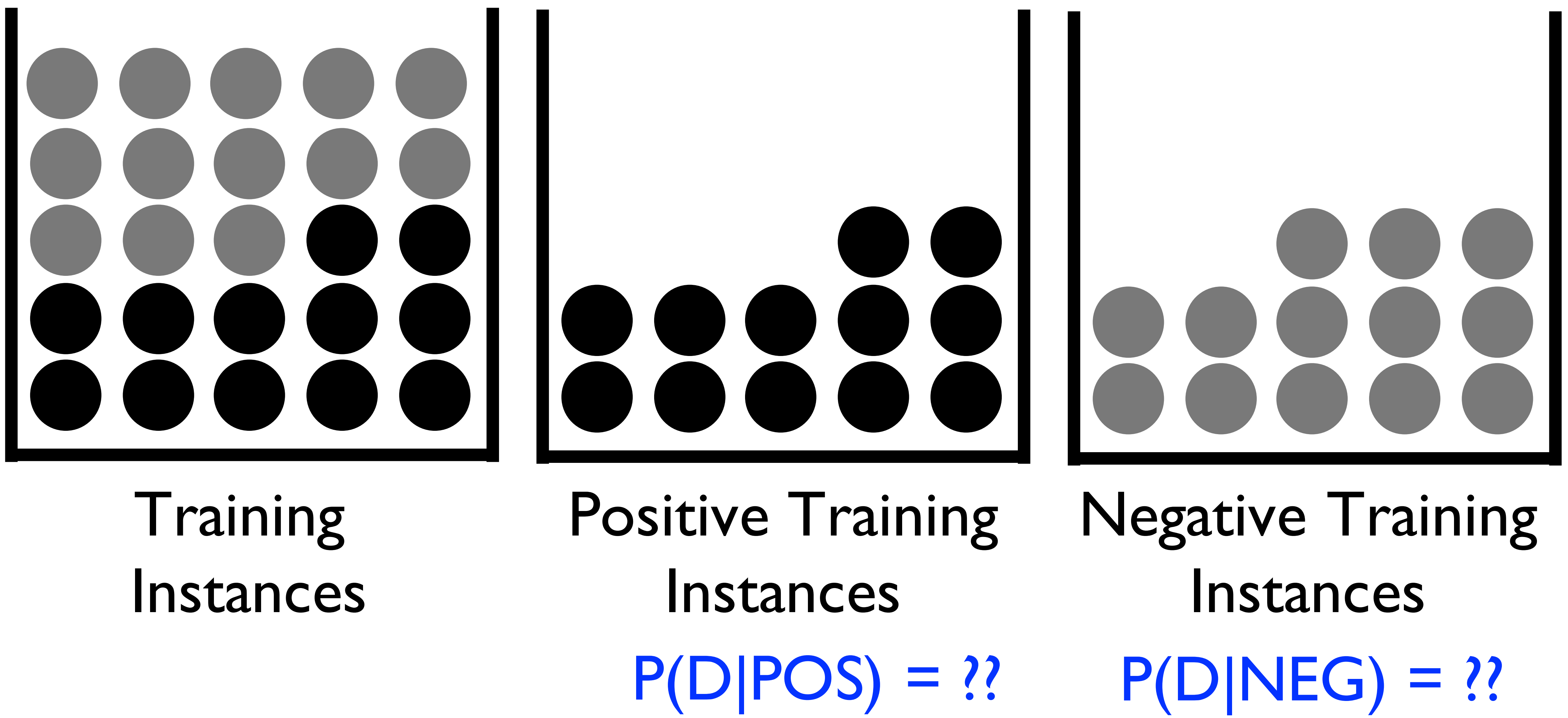
Easy!

Not so easy!

Naive Bayes Classification

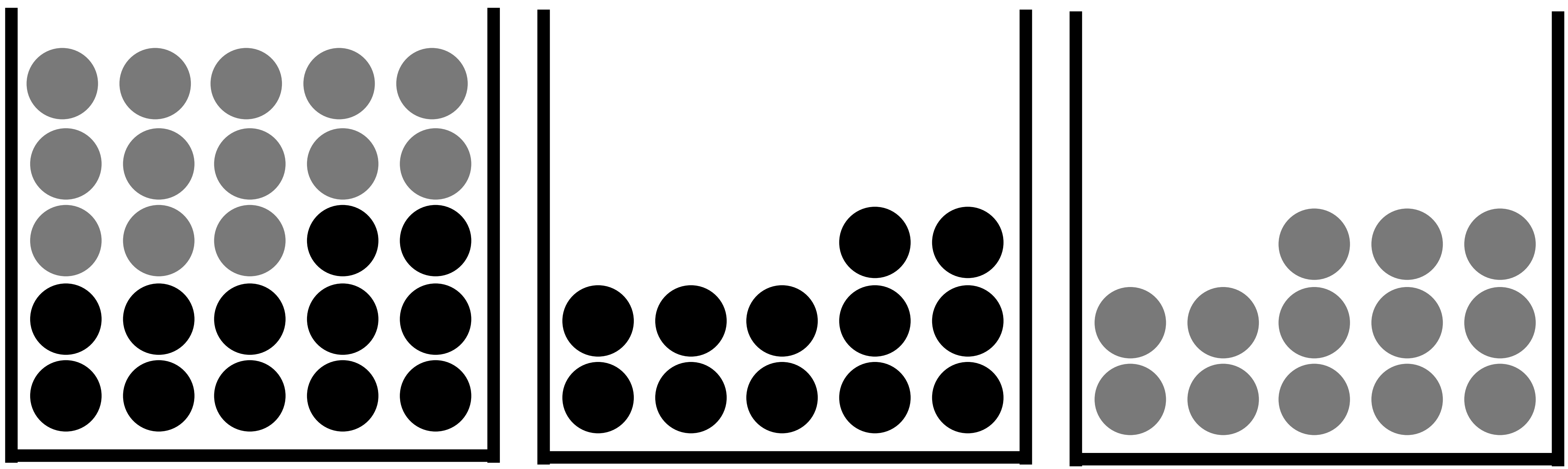
example: positive/negative movie reviews

- We have a problem! What is it?



Naive Bayes Classification
example: positive/negative movie reviews

- We have a problem! What is it?



Training Instances

Positive Training Instances

Negative Training Instances

P(D|POS) = ??

P(D|NEG) = ??

What is D?

What is the Document representation?

- Lets say that our feature space is limited by the following words: w_1, w_2, and w_3.
- Total possible features are:
- w_1, w_2, w_3
- w_1 w_2, w_1 w_3, w_2 w_3
- w_1 w_2 w_3, w_1 w_2 w_4
- Total = 8
- If say our feature space was: w_1, w_2, w_3, and w_4. Then total feature possible combinations are: ?

Feature combinations

- Lets say that our feature space is limited by the following words: w_1, w_2, and w_3.
- Total possible features are:
- w_1, w_2, w_3
- w_1 w_2, w_1 w_3, w_2 w_3
- w_1 w_2 w_3
- Total = 8 => 7 (above combinations) + 1 (no combination)
- If say our feature space was: w_1, w_2, w_3, and w_4. Then total feature possible combinations are: 2^4

Naive Bayes Classification

example: positive/negative movie reviews

- We have a problem! What is it?
- Assuming **n** binary features, the number of possible combinations is 2^n
- 2^{1000} = 1.071509e+301
- And in order to estimate the probability of each combination, we would require multiple occurrences of each combination in the training data!
- We could never have enough training data to reliably estimate P(D|NEG) or P(D|POS)!

Naive Bayes Classification

example: positive/negative movie reviews

- Assumption: given a particular class value (i.e, POS or NEG), the value of a particular feature is independent of the value of other features
- In other words, the value of a particular feature is only dependent on the class value

Naive Bayes Classification

example: positive/negative movie reviews

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	...	w_n	sentiment
1	0	1	0	1	0	0	1	...	0	positive
0	1	0	1	1	0	1	1	...	0	positive
0	1	0	1	1	0	1	0	...	0	positive
0	0	1	0	1	1	0	1	...	1	positive
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	...	⋮	⋮
1	1	0	1	1	0	0	1	...	1	positive

Example of document representation

- Assume the feature space to be the following words: hello, world, bad, movie.
- Review 1: “this is a bad movie!”
- Review 2: “hello world is a movie?”
- What is the bag-of-words document representation for review 1 and review 2?
- Review 1?
- Review 2?

Example of document representation

- Assume the feature space to be the following words: hello, world, bad, movie.
- Review 1: “this is a bad movie!”
- Review 2: “hello world is a movie?”
- What is the bag-of-words document representation for review 1 and review 2?
- Review 1: 0011
- Review 2: 1101

Naive Bayes Classification

example: positive/negative movie reviews

- Assumption: given a particular class value (i.e, POS or NEG), the value of a particular feature is independent of the value of other features

- Example: we have <u>seven</u> features and D = **1011011**

- P(**1011011**|POS) =

  P(w_1=**1**|POS) x P(w_2=**0**|POS) x P(w_3=**1**|POS) x P(w_4=**1**|POS) x P(w_5=**0**|POS) x P(w_6=**1**|POS) x P(w_7=**1**|POS)

- P(**1011011**|NEG) =

  P(w_1=**1**|NEG) x P(w_2=**0**|NEG) x P(w_3=**1**|NEG) x P(w_4=**1**|NEG) x P(w_5=**0**|NEG) x P(w_6=**1**|NEG) x P(w_7=**1**|NEG)

Naive Bayes Classification

example: positive/negative movie reviews

- Assumption: given a particular class value (i.e, POS or NEG), the value of a particular feature is independent of the value of other features

- Example: we have seven features and D = **1011011**

- $\boxed{P(1011011|POS) =}$

You should compute both!

$P(w_1=1|POS) \times P(w_2=0|POS) \times P(w_3=1|POS) \times P(w_4=1|POS) \times P(w_5=0|POS) \times P(w_6=1|POS) \times P(w_7=1|POS)$

- $\boxed{P(1011011|NEG) =}$

$P(w_1=1|NEG) \times P(w_2=0|NEG) \times P(w_3=1|NEG) \times P(w_4=1|NEG) \times P(w_5=0|NEG) \times P(w_6=1|NEG) \times P(w_7=1|NEG)$

Naive Bayes Classification
example: positive/negative movie reviews

- Question: How do we estimate $P(w_I = \mathbf{I} | POS)$?

Naive Bayes Classification

example: positive/negative movie reviews

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	...	w_n	sentiment
1	0	1	0	1	0	0	1	...	0	positive
0	1	0	1	1	0	1	1	...	0	negative
0	1	0	1	1	0	1	0	...	0	negative
0	0	1	0	1	1	0	1	...	1	positive
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	...	⋮	⋮
1	1	0	1	1	0	0	1	...	1	negative

Naive Bayes Classification

example: positive/negative movie reviews

- Question: How do we estimate $P(w_1 = 1 | POS)$?

	POS	NEG
$w_1 = 1$	a	b
$w_1 = 0$	c	d

$P(w_1 = 1 | POS) = ??$

Naive Bayes Classification

example: positive/negative movie reviews

- Question: How do we estimate $P(w_1 = 1 | POS)$?

	POS	NEG
$w_1 = 1$	a	b
$w_1 = 0$	c	d

$P(w_1 = 1 | POS) = a / (a + c)$

Naive Bayes Classification

example: positive/negative movie reviews

- Question: How do we estimate $P(w_1 = 1/0 | POS/NEG)$?

	POS	NEG
$w_1 = 1$	a	b
$w_1 = 0$	c	d

$P(w_1 = 1 | POS) = a / (a + c)$

$P(w_1 = 0 | POS) = ??$

$P(w_1 = 1 | NEG) = ??$

$P(w_1 = 0 | NEG) = ??$

Naive Bayes Classification

example: positive/negative movie reviews

- Question: How do we estimate $P(w_1 = \mathbf{1/0} | POS/NEG)$?

	POS	NEG
$w_1 = \mathbf{1}$	a	b
$w_1 = \mathbf{0}$	c	d

$P(w_1 = \mathbf{1} | POS) = a / (a + c)$

$P(w_1 = \mathbf{0} | POS) = c / (a + c)$

$P(w_1 = \mathbf{1} | NEG) = b / (b + d)$

$P(w_1 = \mathbf{0} | NEG) = d / (b + d)$

Naive Bayes Classification

example: positive/negative movie reviews

- Question: How do we estimate $P(w_2 = 1/0 | POS/NEG)$?

	POS	NEG
$w_2 = 1$	a	b
$w_2 = 0$	c	d

$P(w_2 = 1 | POS) = a / (a + c)$

$P(w_2 = 0 | POS) = c / (a + c)$

$P(w_2 = 1 | NEG) = b / (b + d)$

$P(w_2 = 0 | NEG) = d / (b + d)$

- The value of a, b, c, and d would be different for different features w_1, w_2, w_3, w_4, w_5, ... , w_n

Naive Bayes Classification

example: positive/negative movie reviews

- Given instance D, predict positive (POS) if:

$$P(D|POS) \times P(POS) \geq P(D|NEG) \times P(NEG)$$

- Otherwise, predict negative (NEG)

Naive Bayes Classification

example: positive/negative movie reviews

- Given instance D, predict positive (**POS**) if:

$$P(POS) \times \prod_{i=1}^{n} P(w_i = D_i|POS) \geq P(NEG) \times \prod_{i=1}^{n} P(w_i = D_i|NEG)$$

- Otherwise, predict negative (**NEG**)

Naive Bayes Classification

example: positive/negative movie reviews

- Given instance D = **1011011**, predict positive (POS) if:

$P(w_1=\mathbf{1}|POS) \times P(w_2=\mathbf{0}|POS) \times P(w_3=\mathbf{1}|POS) \times P(w_4=\mathbf{1}|POS) \times P(w_5=\mathbf{0}|POS) \times P(w_6=\mathbf{1}|POS) \times P(w_7=\mathbf{1}|POS) \times P(POS)$

$$\geq$$

$P(w_1=\mathbf{1}|NEG) \times P(w_2=\mathbf{0}|NEG) \times P(w_3=\mathbf{1}|NEG) \times P(w_4=\mathbf{1}|NEG) \times P(w_5=\mathbf{0}|NEG) \times P(w_6=\mathbf{1}|NEG) \times P(w_7=\mathbf{1}|NEG) \times P(NEG)$

- Otherwise, predict negative (NEG)

Naive Bayes Classification
example: positive/negative movie reviews

- We still have a problem! What is it?

Naive Bayes Classification

example: positive/negative movie reviews

- Given instance D = **1011011**, predict positive (POS) if:

$P(w_1=\mathbf{1}|POS) \times P(w_2=\mathbf{0}|POS) \times P(w_3=\mathbf{1}|POS) \times P(w_4=\mathbf{1}|POS) \times P(w_5=\mathbf{0}|POS) \times \mathbf{P(w_6=1|POS)} \times P(w_7=\mathbf{1}|POS) \times P(POS)$

$$\geq$$

$P(w_1=\mathbf{1}|NEG) \times P(w_2=\mathbf{0}|NEG) \times P(w_3=\mathbf{1}|NEG) \times P(w_4=\mathbf{1}|NEG) \times P(w_5=\mathbf{0}|NEG) \times P(w_6=\mathbf{1}|NEG) \times P(w_7=\mathbf{1}|NEG) \times P(NEG)$

- Otherwise, predict negative (NEG)

What if this never happens in the training data?

Smoothing Probability Estimates

- When estimating probabilities, we tend to ...
  - Over-estimate the probability of observed outcomes
  - Under-estimate the probability of unobserved outcomes
- The goal of smoothing is to ...
  - Decrease the probability of observed outcomes
  - Increase the probability of unobserved outcomes
- It's usually a good idea
- You probably already know this concept!

Smoothing Probability Estimates

- YOU: Are there mountain lions around here?
- YOUR FRIEND: Nope.
- YOU: How can you be so sure?
- YOUR FRIEND: Because I've been hiking here five times before and have never seen one.
- YOU: ????

Smoothing Probability Estimates

- YOU: Are there mountain lions around here?
- YOUR FRIEND: Nope.
- YOU: How can you be so sure?
- YOUR FRIEND: Because I've been hiking here five times before and have never seen one.
- MOUNTAIN LION: You should have learned about smoothing by taking INLS 613. Yum!

Add-One Smoothing

- Question: How do we estimate $P(w_2=1/0|POS/NEG)$?

	POS	NEG
$w_2 = 1$	a	b
$w_2 = 0$	c	d

$P(w_2=1|POS) = a / (a + c)$

$P(w_2=0|POS) = c / (a + c)$

$P(w_2=1|NEG) = b / (b + d)$

$P(w_2=0|NEG) = d / (b + d)$

Add-One Smoothing

- Question: How do we estimate $P(w_2 = \mathbf{1/0} | POS/NEG)$?

	POS	NEG
$w_2 = \mathbf{1}$	a + 1	b + 1
$w_2 = \mathbf{0}$	c + 1	d + 1

$P(w_2 = \mathbf{1} | POS) = ??$

$P(w_2 = \mathbf{0} | POS) = ??$

$P(w_2 = \mathbf{1} | NEG) = ??$

$P(w_2 = \mathbf{0} | NEG) = ??$

Add-One Smoothing

- Question: How do we estimate $P(w_2 = 1/0 | POS/NEG)$?

	POS	NEG
$w_2 = 1$	$a + 1$	$b + 1$
$w_2 = 0$	$c + 1$	$d + 1$

$P(w_2=1|POS) = (a + 1) / (a + c + 2)$

$P(w_2=0|POS) = (c + 1) / (a + c + 2)$

$P(w_2=1|NEG) = (b + 1) / (b + d + 2)$

$P(w_2=0|NEG) = (d + 1) / (b + d + 2)$

Naive Bayes Classification

example: positive/negative movie reviews

- Given instance D, predict positive (POS) if:

$$P(POS) \times \prod_{i=1}^{n} P(w_i = D_i | POS) \geq P(NEG) \times \prod_{i=1}^{n} P(w_i = D_i | NEG)$$

- Otherwise, predict negative (NEG)

Naive Bayes Classification

- Naive Bayes Classifiers are simple, effective, robust, and very popular
- Assumes that feature values are conditionally independent given the target class value
- This assumption does not hold in natural language
- Even so, NB classifiers are very powerful
- Smoothing is necessary in order to avoid zero probabilities