Naïve Bayes

Original Slides by:
Courtney Anne Ngo
Daniel Stanley Tan, PhD
Arren Antioquia

Updated (AY 2023 – 2024 T3) by:
Thomas James Tiam-Lee, PhD

College of
Computer Studies

Naïve Bayes

- **Supervised** machine learning algorithm primarily designed for classification
- Uses a **statistical approach** to machine learning.
- Key idea: model the training data as being **generated from some statistical process**

Probability Review (Discrete)

- Question:

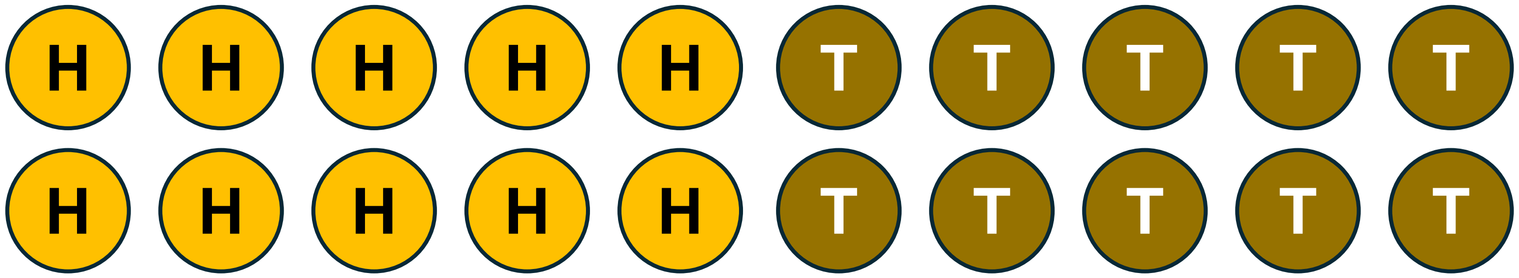
- If we have a fair coin (50% head and 50% tails), and we throw it 20 times, how many heads and how many tails do we expect to observe?

Probability Review (Discrete)

- Question:

- If we have a fair coin (50% head and 50% tails), and we throw it 20 times, how many heads and how many tails do we expect to observe?

- **Answer:** 10 heads and 10 tails



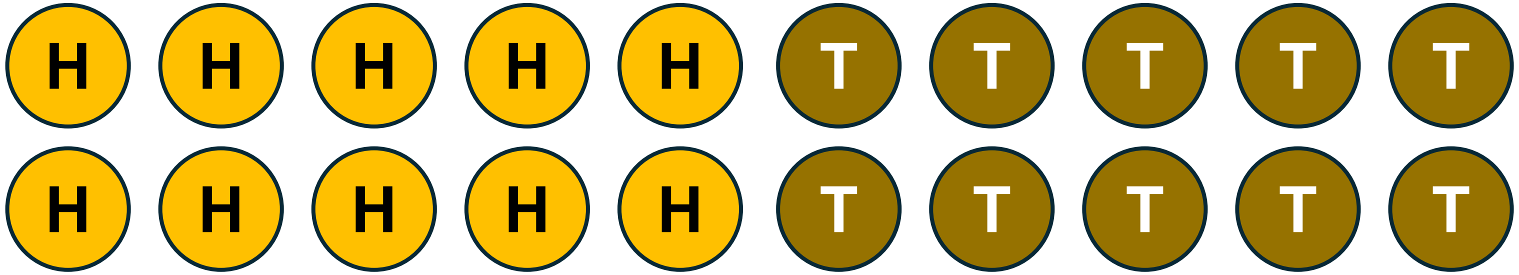
Probability Review (Discrete)

- Question:

Random process that generated the observation

- If we have a fair coin (50% head and 50% tails), and we throw it 20 times, how many heads and how many tails do we expect to observe?

- **Answer:** 10 heads and 10 tails

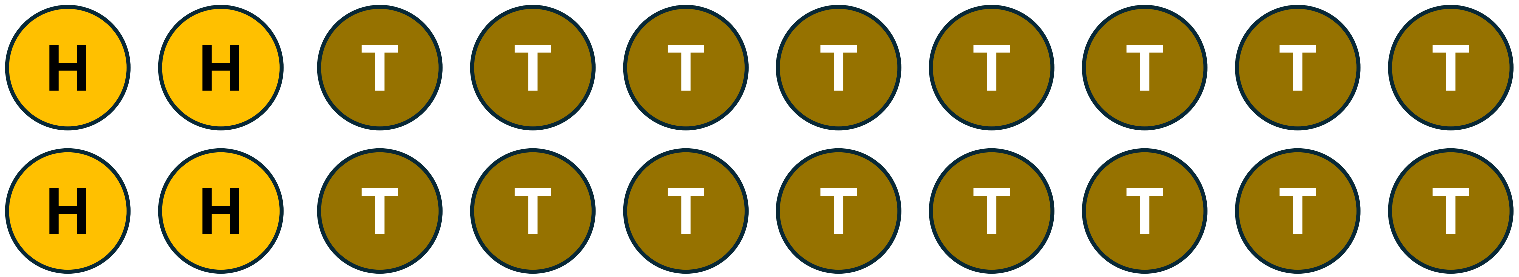


The observation generated by the random process

Probability Review (Discrete)

- Let’s reverse the question:

- You threw a coin 20 times and observed the following:



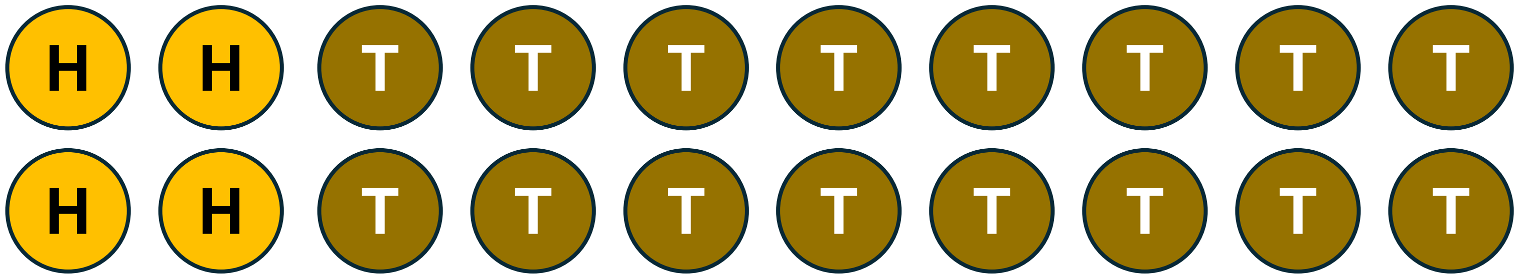
4 heads 16 tails

- Is it a fair coin? If not, what kind of coin is it?

Probability Review (Discrete)

- Let’s reverse the question:

- You threw a coin 20 times and observed the following:



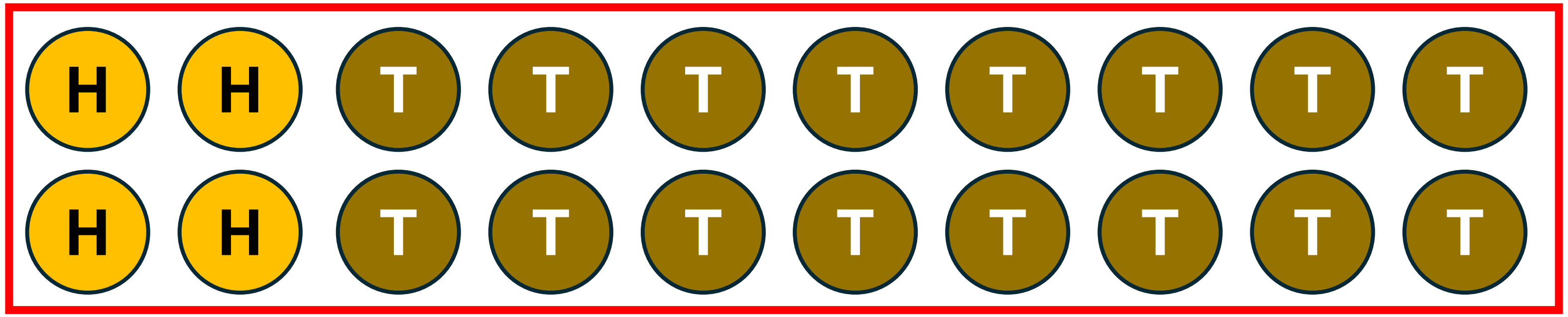
4 heads 16 tails

- **Answer:** It’s a coin with 20% probability of heads and 80% probability of tails

Probability Review (Discrete)

- Let's reverse the question:

- You threw a coin 20 times and observed the following:



The observation generated by the random process

4 heads

16 tails

- **Answer:** It's a coin with 20% probability of heads and 80% probability of tails

Random process that likely generated the observation

Probability Review (Discrete)

- Question: What is the probability that it will rain in this planet?

Raining	Humidity
yes	high
yes	low
no	high
no	low
no	low
no	low
yes	high

Probability Review (Discrete)

- Question: What is the probability that it will rain in this planet?
- **Answer:** $\frac{3}{7} = 0.4286 = 42.86\%$

Raining	Humidity
yes	high
yes	low
no	high
no	low
no	low
no	low
yes	high

Probability Review (Discrete)

- Question: What is the probability that the humidity is low on this planet?

Raining	Humidity
yes	high
yes	low
no	high
no	high
no	high
no	low
yes	high

Probability Review (Discrete)

- Question: What is the probability that the humidity is low on this planet?

- **Answer:** $\frac{2}{7} = 0.2857 = 28.57\%$

Raining	Humidity
yes	high
yes	low
no	high
no	high
no	high
no	low
yes	high

Probability Review (Discrete)

$$P(A) = \frac{number\ of\ times\ event\ A\ happened}{total\ number\ of\ observations}$$

- Note: it should be noted that the probability will more reliable if there are more samples.

Raining	Humidity
yes	high
yes	low
no	high
no	high
no	high
no	low
yes	high

Probability Review (Discrete)

- Question: In this planet, what is the probability that the humidity is low **given that it is not raining?**

Raining	Humidity
yes	high
yes	low
no	high
no	high
no	high
no	low
yes	high

Probability Review (Discrete)

- Question: In this planet, what is the probability that the humidity is low **given that it is not raining?**

- **Answer:** $\frac{1}{4} = 0.25 = 25\%$

Raining	Humidity
yes	high
yes	low
no	high
no	high
no	high
no	low
yes	high

Probability Review (Discrete)

- Question: In this planet, what is the probability that it is not raining **given that the humidity is low?**

Raining	Humidity
yes	high
yes	low
no	high
no	high
no	high
no	low
yes	high

Probability Review (Discrete)

- Question: In this planet, what is the probability that it is not raining **given that the humidity is low?**

- **Answer:** $\frac{1}{2} = 0.5 = 50\%$

Raining	Humidity
yes	high
yes	low
no	high
no	high
no	high
no	low
yes	high

Probability Review (Discrete)

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

- $P(A|B)$ is read as:
  - "probability of A given B"

Raining	Humidity
yes	high
yes	low
no	high
no	high
no	high
no	low
yes	high

Sample Classification Task

- **Task:** predict whether a given student will pass ML class based on their math grade and number of hours studying per week.

- What can we “learn” from the historical data?

Student	Math grade	Hours studying	ML grade
0	Bad	>= 4hrs	passed
1	Good	>= 4hrs	passed
2	Bad	< 4hrs	failed
3	Bad	>= 4hrs	passed
4	Good	>= 4hrs	passed
5	Bad	>= 4hrs	failed
6	Good	< 4hrs	passed
7	Good	< 4hrs	failed
8	Good	>= 4hrs	passed
9	Bad	>= 4hrs	passed

Sample Classification Task

- From a statistical point of view, we can infer:

P(ML=passed | Math=bad ∩ Study>=4hrs)
P(ML=passed | Math=bad ∩ Study<4hrs)
P(ML=passed | Math=good ∩ Study>=4hrs)
P(ML=passed | Math=good ∩ Study<4hrs)
P(ML=failed | Math=bad ∩ Study>=4hrs)
P(ML=failed | Math=bad ∩ Study<4hrs)
P(ML=failed | Math=good ∩ Study>=4hrs)
P(ML=failed | Math=good ∩ Study<4hrs)

Student	Math grade	Hours studying	ML grade
0	Bad	>= 4hrs	passed
1	Good	>= 4hrs	passed
2	Bad	< 4hrs	failed
3	Bad	>= 4hrs	passed
4	Good	>= 4hrs	passed
5	Bad	>= 4hrs	failed
6	Good	< 4hrs	passed
7	Good	< 4hrs	failed
8	Good	>= 4hrs	passed
9	Bad	>= 4hrs	passed

Sample Classification Task

- From a statistical point of view, we can infer:

P(ML=passed | Math=bad ∩ Study>=4hrs) = 0.67
P(ML=passed | Math=bad ∩ Study<4hrs) = 0.00
P(ML=passed | Math=good ∩ Study>=4hrs) = 1.00
P(ML=passed | Math=good ∩ Study<4hrs) = 0.50
P(ML=failed | Math=bad ∩ Study>=4hrs) = 0.33
P(ML=failed | Math=bad ∩ Study<4hrs) = 1.00
P(ML=failed | Math=good ∩ Study>=4hrs) = 0.00
P(ML=failed | Math=good ∩ Study<4hrs) = 0.50

Question: Given a student who has a bad grade in math, and studies for more than 4 hours, predict whether he will pass ML or not.

Student	Math grade	Hours studying	ML grade
0	Bad	>= 4hrs	passed
1	Good	>= 4hrs	passed
2	Bad	< 4hrs	failed
3	Bad	>= 4hrs	passed
4	Good	>= 4hrs	passed
5	Bad	>= 4hrs	failed
6	Good	< 4hrs	passed
7	Good	< 4hrs	failed
8	Good	>= 4hrs	passed
9	Bad	>= 4hrs	passed

Sample Classification Task

- From a statistical point of view, we can infer:

P(ML=passed | Math=bad ∩ Study>=4hrs) = 0.67
P(ML=passed | Math=bad ∩ Study<4hrs) = 0.00
P(ML=passed | Math=good ∩ Study>=4hrs) = 1.00
P(ML=passed | Math=good ∩ Study<4hrs) = 0.50
P(ML=failed | Math=bad ∩ Study>=4hrs) = 0.33
P(ML=failed | Math=bad ∩ Study<4hrs) = 1.00
P(ML=failed | Math=good ∩ Study>=4hrs) = 0.00
P(ML=failed | Math=good ∩ Study<4hrs) = 0.50

Question: Given a student who has a bad grade in math, and studies for more than 4 hours, predict whether he will pass ML or not.

Student	Math grade	Hours studying	ML grade
0	Bad	>= 4hrs	passed
1	Good	>= 4hrs	passed
2	Bad	< 4hrs	failed
3	Bad	>= 4hrs	passed
4	Good	>= 4hrs	passed
5	Bad	>= 4hrs	failed
6	Good	< 4hrs	passed
7	Good	< 4hrs	failed
8	Good	>= 4hrs	passed
9	Bad	>= 4hrs	passed

Prediction: the student will pass!

The Problems with This Approach…

- You must compute all possible combinations of features! (or remember the entire dataset, which will bloat the model size)

P(ML=passed | Math=bad ∩ Study>=4hrs) = 0.67
P(ML=passed | Math=bad ∩ Study<4hrs) = 0.00
P(ML=passed | Math=good ∩ Study>=4hrs) = 1.00
P(ML=passed | Math=good ∩ Study<4hrs) = 0.50
P(ML=failed | Math=bad ∩ Study>=4hrs) = 0.33
P(ML=failed | Math=bad ∩ Study<4hrs) = 1.00
P(ML=failed | Math=good ∩ Study>=4hrs) = 0.00
P(ML=failed | Math=good ∩ Study<4hrs) = 0.50

Student	Math grade	Hours studying	ML grade
0	Bad	>= 4hrs	passed
1	Good	>= 4hrs	passed
2	Bad	< 4hrs	failed
3	Bad	>= 4hrs	passed
4	Good	>= 4hrs	passed
5	Bad	>= 4hrs	failed
6	Good	< 4hrs	passed
7	Good	< 4hrs	failed
8	Good	>= 4hrs	passed
9	Bad	>= 4hrs	passed

If we have 20 binary features, we would need to compute $2^{20} = 1048576$ probabilities!

The Problems with This Approach…

- It also means you need to have enough examples for every possible combination of features (curse of dimensionality)

P(ML=passed | Math=bad ∩ Study>=4hrs) = 0.67
P(ML=passed | Math=bad ∩ Study<4hrs) = 0.00
P(ML=passed | Math=good ∩ Study>=4hrs) = 1.00
P(ML=passed | Math=good ∩ Study<4hrs) = 0.50
P(ML=failed | Math=bad ∩ Study>=4hrs) = 0.33
P(ML=failed | Math=bad ∩ Study<4hrs) = 1.00
P(ML=failed | Math=good ∩ Study>=4hrs) = 0.00
P(ML=failed | Math=good ∩ Study<4hrs) = 0.50

Student	Math grade	Hours studying	ML grade
0	Bad	>= 4hrs	passed
1	Good	>= 4hrs	passed
2	Bad	< 4hrs	failed
3	Bad	>= 4hrs	passed
4	Good	>= 4hrs	passed
5	Bad	>= 4hrs	failed
6	Good	< 4hrs	passed
7	Good	< 4hrs	failed
8	Good	>= 4hrs	passed
9	Bad	>= 4hrs	passed

Probability Review (Independent

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

only if A and B are independent.

Otherwise,

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

- **If events are independent**, the probability of those events all happening can just be the **product of their individual probabilities**

Probability Review (Independent Events)

- Example of independent events:

  - If you flip a fair coin two times, what is the probability that both flips will result in heads?

  - $0.5 \times 0.5 = 0.25$

Probability Review (Independent Events)

- Example of non-independent events:

  - From a 100-day historical data in an unknown planet… it rained for 50 days (0.5 probability) and the ground was wet for 50 days (0.5 probability). What is the probability that that it will be raining and at the same time the ground is wet?

  - $0.5 \times 0.5 = 0.25$ (???)
    - This is obviously wrong

Addressing the Problem…

$$P(\underbrace{y = \square}_{\text{target}} | \underbrace{X_1 = \square \cap X_2 = \square \cap \cdots \cap X_d = \square}_{\text{features}})$$

$$\Downarrow$$

$$P(T \mid F)$$

Addressing the Problem...

$$P(T \mid F) = \frac{P(T \cap F)}{P(F)}$$

Addressing the Problem…

These events are not independent!
(the fact that we are using F to predict T means that we believe T is dependent on F)

$$P(T \mid F) = \frac{\boxed{P(T \cap F)}}{P(F)}$$

Addressing the Problem…

$$P(T \mid F) = \frac{P(F \mid T)P(T)}{P(F)}$$

Addressing the Problem...

$$P(T \mid F) = \frac{P(F \mid T)P(T)}{P(F)}$$

This is known as the Bayes Rule

Addressing the Problem...

$$P(T \mid F) = \frac{P(\boxed{F} \mid T)P(T)}{P(F)}$$

Addressing the Problem…

$$P(T \mid F) = \frac{P(X_1 = \square \cap X_2 = \square \cap \cdots \cap X_d = \square \mid T)P(T)}{P(F)}$$

Addressing the Problem...

Are these events independent?

$$P(T \mid F) = \frac{P(\boxed{X_1 = \square \cap X_2 = \square \cap \cdots \cap X_d = \square} \mid T)P(T)}{P(F)}$$

Addressing the Problem...

Are these events independent?

$$P(T \mid F) = \frac{P(\boxed{X_1 = \square \cap X_2 = \square \cap \cdots \cap X_d = \square} \mid T)P(T)}{P(F)}$$

Technically, we **cannot** say that the features are independent. For example, students with good grade in math may be more likely to study more in general.

Addressing the Problem...

Are these events independent?

$$P(T \mid F) = \frac{P(\boxed{X_1 = \square \cap X_2 = \square \cap \cdots \cap X_d = \square} \mid T)P(T)}{P(F)}$$

However, to make things more manageable, we will just **assume that the features are independent!**

This is the principle of <u>Naïve</u> Bayes

Naïve Bayes

$$P(T \mid F) = \frac{P(X_1 = \square | T) \times P(X_2 = \square | T) \times \cdots \times P(X_d = \square | T) \times P(T)}{P(F)}$$

- Notice that now, we never have to compute the joint probability of multiple features anymore!
- We just compute the probability of the features **independently**.
- Statistically speaking, this is wrong, but it turns out that this “naïve” assumption can still yield decent predictions!

Naïve Bayes

Student	Math grade	Hours studying	ML grade
0	Bad	>= 4hrs	passed
1	Good	>= 4hrs	passed
2	Bad	< 4hrs	failed
3	Bad	>= 4hrs	passed
4	Good	>= 4hrs	passed
5	Bad	>= 4hrs	failed
6	Good	< 4hrs	passed
7	Good	< 4hrs	failed
8	Good	>= 4hrs	passed
9	Bad	>= 4hrs	failed

Parameters of NB Model:

Given	Grade = Good	Grade = Bad	Hours >= 4hrs	Hours < 4 hrs
Passed				
Failed				

P(ML = Pass)	
P(ML = Fail)	

Naïve Bayes

Student	Math grade	Hours studying	ML grade
0	Bad	>= 4hrs	passed
1	Good	>= 4hrs	passed
2	Bad	< 4hrs	failed
3	Bad	>= 4hrs	passed
4	Good	>= 4hrs	passed
5	Bad	>= 4hrs	failed
6	Good	< 4hrs	passed
7	Good	< 4hrs	failed
8	Good	>= 4hrs	passed
9	Bad	>= 4hrs	failed

Parameters of NB Model:

Given	Grade = Good	Grade = Bad	Hours >= 4hrs	Hours < 4 hrs
Passed				
Failed				

P(ML = Pass)	
P(ML = Fail)	

Naïve Bayes

Test Instance:

Math grade is **bad**, studies < **4hrs**

Student	Math grade	Hours studying	ML grade
0	Bad	>= 4hrs	passed
1	Good	>= 4hrs	passed
2	Bad	< 4hrs	failed
3	Bad	>= 4hrs	passed
4	Good	>= 4hrs	passed
5	Bad	>= 4hrs	failed
6	Good	< 4hrs	passed
7	Good	< 4hrs	failed
8	Good	>= 4hrs	passed
9	Bad	>= 4hrs	failed

Parameters of NB Model:

Given	Grade = Good	Grade = Bad	Hours >= 4hrs	Hours < 4 hrs
Passed				
Failed				

P(ML = Pass)	
P(ML = Fail)	

Naïve Bayes

Test Instance:

Math grade is **bad**, studies < **4hrs**

Student	Math grade	Hours studying	ML grade
0	Bad	>= 4hrs	passed
1	Good	>= 4hrs	passed
2	Bad	< 4hrs	failed
3	Bad	>= 4hrs	passed
4	Good	>= 4hrs	passed
5	Bad	>= 4hrs	failed
6	Good	< 4hrs	passed
7	Good	< 4hrs	failed
8	Good	>= 4hrs	passed
9	Bad	>= 4hrs	failed

Parameters of NB Model:

Given	Grade = Good	Grade = Bad	Hours >= 4hrs	Hours < 4 hrs
Passed				
Failed				

P(ML = Pass)	
P(ML = Fail)	

$$P(Pass|Bad, < 4hrs) = \frac{P(Bad|Passed)P(< 4hrs|Passed)P(Passed)}{P(Bad, < 4hrs)}$$

$$P(Fail|Bad, < 4hrs) = \frac{P(Bad|Fail)P(< 4hrs|Fail)P(Fail)}{P(Bad, < 4hrs)}$$

Naïve Bayes

Test Instance:

Math grade is **bad**, studies < **4hrs**

Student	Math grade	Hours studying	ML grade
0	Bad	>= 4hrs	passed
1	Good	>= 4hrs	passed
2	Bad	< 4hrs	failed
3	Bad	>= 4hrs	passed
4	Good	>= 4hrs	passed
5	Bad	>= 4hrs	failed
6	Good	< 4hrs	passed
7	Good	< 4hrs	failed
8	Good	>= 4hrs	passed
9	Bad	>= 4hrs	failed

Parameters of NB Model:

Given	Grade = Good	Grade = Bad	Hours >= 4hrs	Hours < 4 hrs
Passed				
Failed				

P(ML = Pass)	
P(ML = Fail)	

$$P(Pass|Bad, < 4hrs) = \frac{(0.33)(0.17)(0.6)}{P(Bad, < 4hrs)} = \frac{0.03366}{P(Bad, < 4hrs)}$$

$$P(Fail|Bad, < 4hrs) = \frac{(0.75)(0.5)(0.4)}{P(Bad, < 4hrs)} = \frac{0.15}{P(Bad, < 4hrs)}$$

Naïve Bayes

Test Instance:

Math grade is **bad**, studies < **4hrs**

Student	Math grade	Hours studying	ML grade
0	Bad	>= 4hrs	passed
1	Good	>= 4hrs	passed
2	Bad	< 4hrs	failed
3	Bad	>= 4hrs	passed
4	Good	>= 4hrs	passed
5	Bad	>= 4hrs	failed
6	Good	< 4hrs	passed
7	Good	< 4hrs	failed
8	Good	>= 4hrs	passed
9	Bad	>= 4hrs	failed

Parameters of NB Model:

Given	Grade = Good	Grade = Bad	Hours >= 4hrs	Hours < 4 hrs
Passed				
Failed				

P(ML = Pass)	
P(ML = Fail)	

$$P(Pass|Bad, < 4hrs) = \frac{(0.33)(0.17)(0.6)}{P(Bad, < 4hrs)} = \frac{0.03366}{P(Bad, < 4hrs)}$$

$$P(Fail|Bad, < 4hrs) = \frac{(0.75)(0.5)(0.4)}{P(Bad, < 4hrs)} = \frac{0.15}{P(Bad, < 4hrs)}$$

We don’t even need to compute the denominator anymore (it’s always going to be the same)

Naïve Bayes

Student	Math grade	Hours studying	ML grade
0	Bad	>= 4hrs	passed
1	Good	>= 4hrs	passed
2	Bad	< 4hrs	failed
3	Bad	>= 4hrs	passed
4	Good	>= 4hrs	passed
5	Bad	>= 4hrs	failed
6	Good	< 4hrs	passed
7	Good	< 4hrs	failed
8	Good	>= 4hrs	passed
9	Bad	>= 4hrs	failed

Test Instance:

Math grade is **bad**, studies < **4hrs**

Parameters of NB Model:

Given	Grade = Good	Grade = Bad	Hours >= 4hrs	Hours < 4 hrs
Passed				
Failed				

P(ML = Pass)	
P(ML = Fail)	

$$P(Pass|Bad, < 4hrs) = \frac{(0.33)(0.17)(0.6)}{P(Bad, < 4hrs)} = \frac{0.03366}{P(Bad, < 4hrs)}$$

$$P(Fail|Bad, < 4hrs) = \frac{(0.75)(0.5)(0.4)}{P(Bad, < 4hrs)} = \frac{0.15}{P(Bad, < 4hrs)}$$

Prediction: The student will fail!

Probability Review (Continuous)

- Question:

- If the average age of students in a university is 20, and we pick 100 students randomly from this university, what do expect their ages are going to be?

Probability Review (Continuous)

- Question:

- If the average age of students in a university is 20 with a standard deviation of 1.5, and we pick 100 students randomly from this university, what do expect their ages are going to be?

- **Answer:** mostly 20 and close to 20, with the frequency decreasing as move farther away from 20.

(note: we assume that the variable is normally distributed)

Probability Review (Continuous)

- Question:

Random process that generated the observation

- If the average age of students in a university is 20 with a standard deviation of 1.5, and we pick 100 students randomly from this university, what do expect their ages are going to be?

Observation that was generated from the random process

- **Answer:** mostly 20 and close to 20, with the frequency decreasing as move farther away from 20.

(note: we assume that the variable is normally distributed)

Probability Review (Continuous)

- Let's reverse the question:

- If we pick 10 random students from a school, and their ages are:

  - 19, 18, 19, 20, 19, 20, 19, 19, 18, 19

- What is the average and standard deviation of the age of students in that school?

Probability Review (Continuous)

- Let's reverse the question:

- If we pick 10 random students from a school, and their ages are:

  - 19, 18, 19, 20, 19, 20, 19, 19, 18, 19

- What is the average and standard deviation of the age of students in that school?
- **Answer:** We can estimate it to be: $\mu = 19$ and $\sigma = 0.67$

Probability Review (Continuous)

- Let's reverse the question:

- If we pick 10 random students from a school, and their ages are:

  - 19, 18, 19, 20, 19, 20, 19, 19, 18, 19

Observation that was generated from the random process

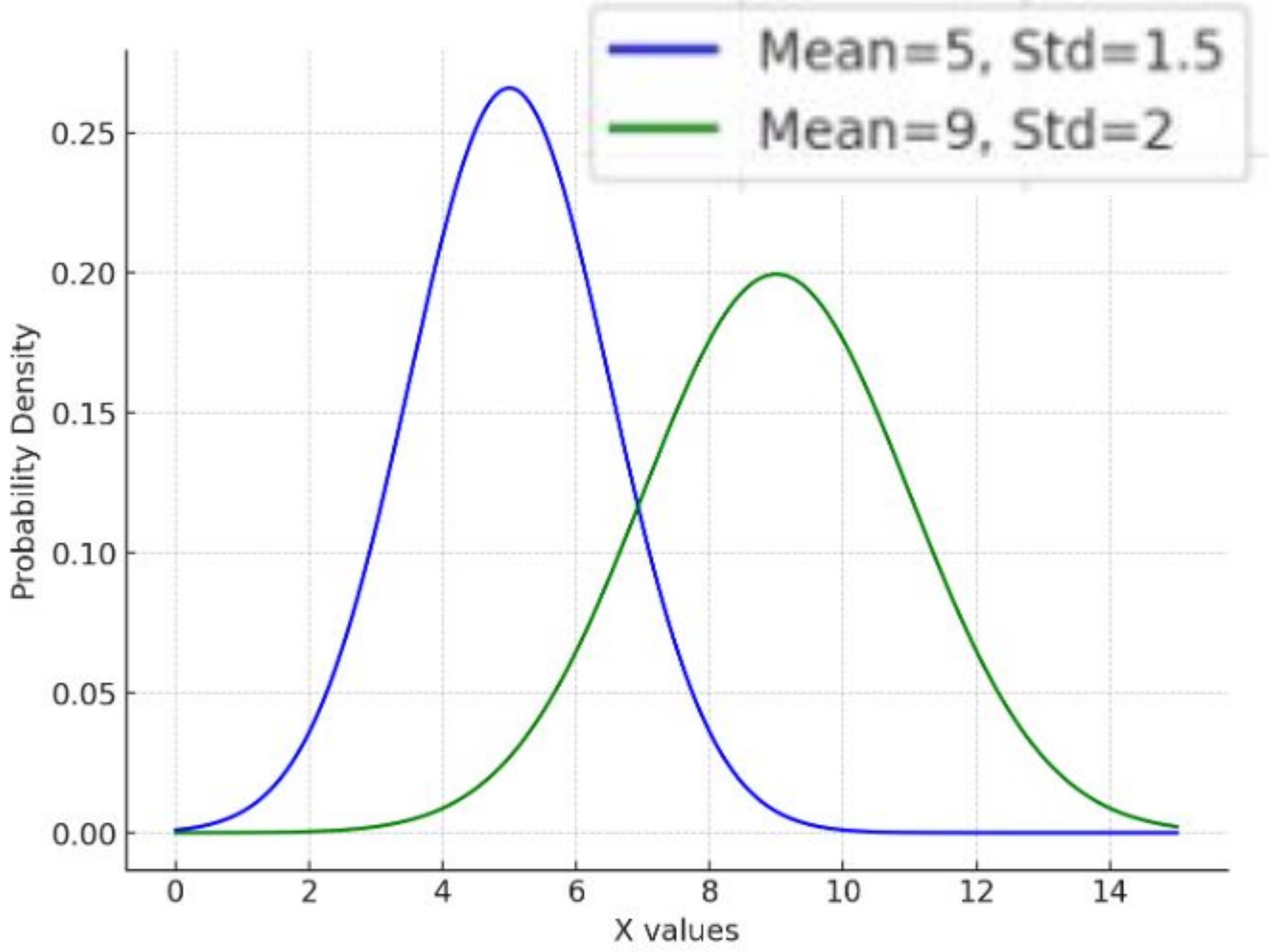
- What is the average and standard deviation of the age of students in that school?
- **Answer:** We can estimate it to be: $\mu = 19$ and $\sigma = 0.67$

Random process that likely generated the observation

Probability Review (Continuous)

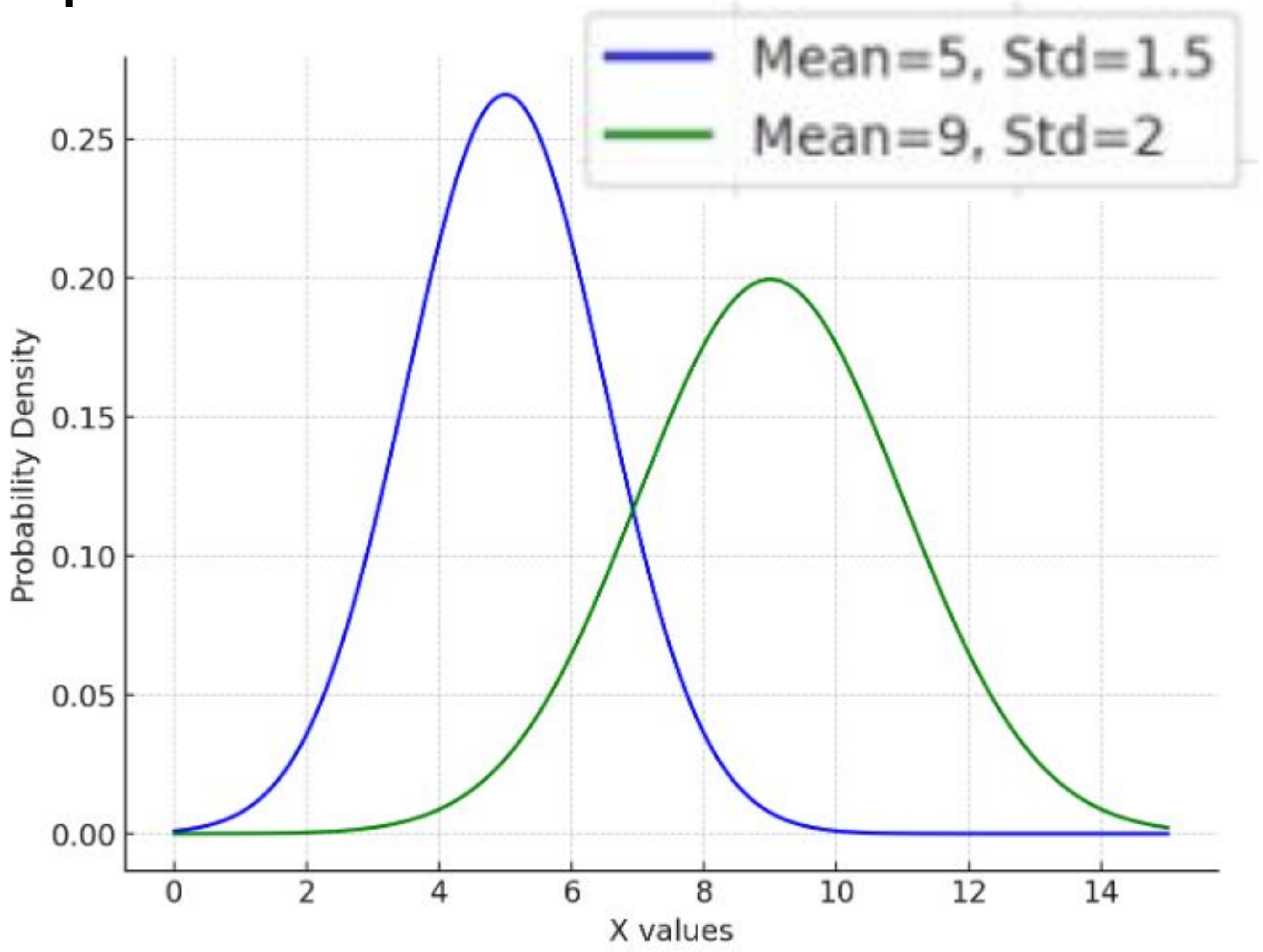
- Given two random normal processes:

- If you observe a value of 6, is it more likely to have been generated by the blue or green process?



Probability Review (Continuous)

- Given two random normal processes:



- **Answer:**
- Compute the probability density function at point 6

- pdf(6) for blue = 0.21
- pdf(6) for green = 0.06

- Blue is more likely to generate 6!

Naïve Bayes

Student	Math grade	Hours studying	ML grade
0	1.0	6	passed
1	4.0	5	passed
2	1.5	2	failed
3	2.0	6	passed
4	3.5	7	passed
5	1.5	4	failed
6	3.0	3	passed
7	4.0	2	failed
8	3.5	6	passed
9	1.0	4	failed

Parameters of NB Model:

Given	Math grade		Hours studying	
Passed				
Failed				

P(ML = Pass)	
P(ML = Fail)	

Naïve Bayes

Student	Math grade	Hours studying	ML grade
0	1.0	6	passed
1	4.0	5	passed
2	1.5	2	failed
3	2.0	6	passed
4	3.5	7	passed
5	1.5	4	failed
6	3.0	3	passed
7	4.0	2	failed
8	3.5	6	passed
9	1.0	4	failed

Parameters of NB Model:

Given	Math grade		Hours studying	
Passed				
Failed				

P(ML = Pass)	
P(ML = Fail)	

Naïve Bayes

Test Instance:

Math grade is **3.5**, studies **5** hours

Student	Math grade	Hours studying	ML grade
0	1.0	6	passed
1	4.0	5	passed
2	1.5	2	failed
3	2.0	6	passed
4	3.5	7	passed
5	1.5	4	failed
6	3.0	3	passed
7	4.0	2	failed
8	3.5	6	passed
9	1.0	4	failed

Parameters of NB Model:

Given	Math grade		Hours studying	
Passed				
Failed				

P(ML = Pass)	
P(ML = Fail)	

Naïve Bayes

Test Instance:

Math grade is 3.5, studies 5 hours

Student	Math grade	Hours studying	ML grade
0	1.0	6	passed
1	4.0	5	passed
2	1.5	2	failed
3	2.0	6	passed
4	3.5	7	passed
5	1.5	4	failed
6	3.0	3	passed
7	4.0	2	failed
8	3.5	6	passed
9	1.0	4	failed

Parameters of NB Model:

Given	Math grade		Hours studying	
Passed				
Failed				

P(ML = Pass)	
P(ML = Fail)	

$$P(Pass|3.5,5) = \frac{(0.292)(0.271)(0.6)}{P(3.5,5)} = \frac{0.0474792}{P(3.5,5)}$$

$$P(Fail|3.5,5) = \frac{(0.16)(0.077)(0.4)}{P(3.5,5)} = \frac{0.004928}{P(3.5,5)}$$

Test Instance:

Math grade is 3.5, studies 5 hours

Naïve Bayes

Student	Math grade	Hours studying	ML grade
0	1.0	6	passed
1	4.0	5	passed
2	1.5	2	failed
3	2.0	6	passed
4	3.5	7	passed
5	1.5	4	failed
6	3.0	3	passed
7	4.0	2	failed
8	3.5	6	passed
9	1.0	4	failed

Parameters of NB Model:

Given	Math grade		Hours studying	
Passed				
Failed				

P(ML = Pass)	
P(ML = Fail)	

$$P(Pass|3.5,5) = \frac{(0.292)(0.271)(0.6)}{P(3.5,5)} = \frac{0.0474792}{P(3.5,5)}$$

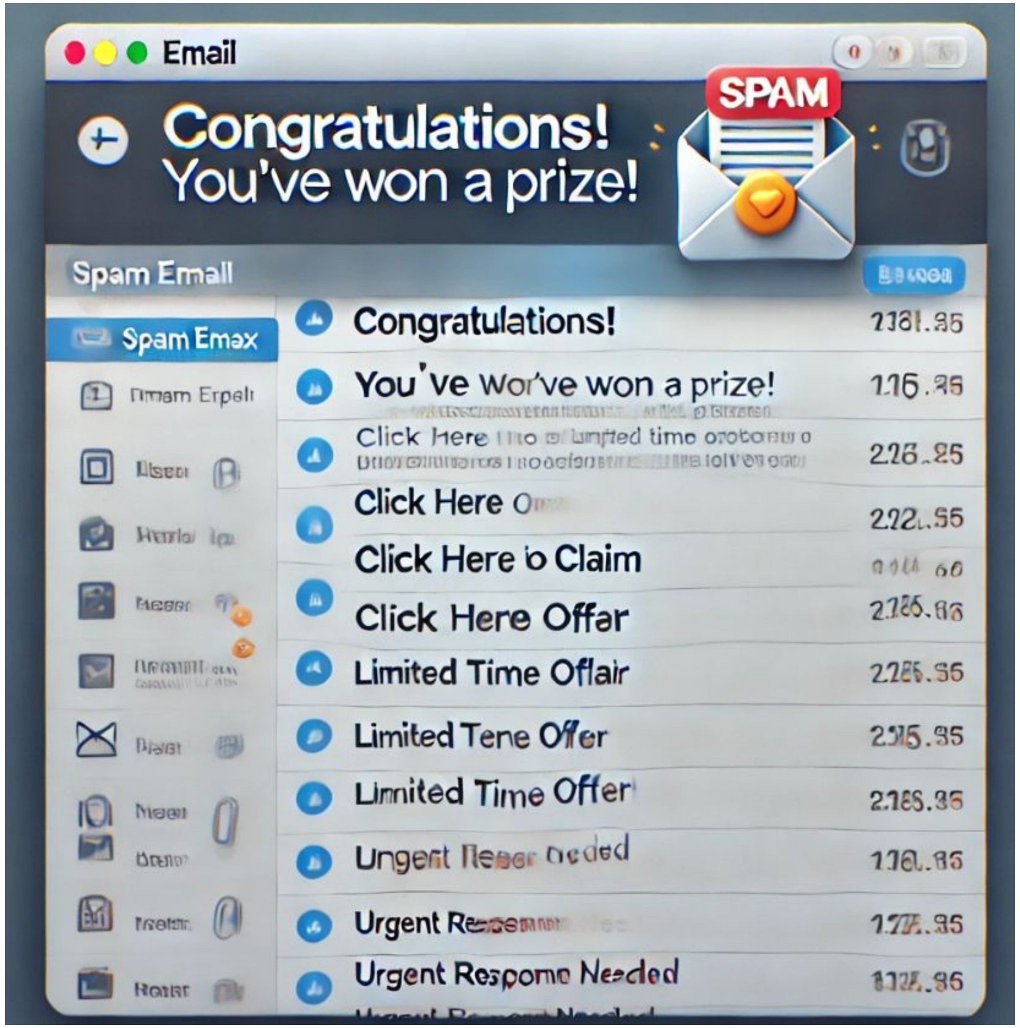
$$P(Fail|3.5,5) = \frac{(0.16)(0.077)(0.4)}{P(3.5,5)} = \frac{0.004928}{P(3.5,5)}$$

Prediction: The student will pass!

Multinomial Naïve Bayes

- **Problem:** Determine whether an email is spam or not spam.

- **Training Data:** a collection of spam and non-spam (ham) emails.



Multinomial Naïve Bayes

P(spam)	0.7
P(ham)	0.3

Ratio of spam and ham documents

	Spam	Ham
buy	0.20	0.10
products	0.10	0.19
please	0.50	0.01
promise	0.10	0.10
DLSU	0.05	0.30
learn	0.05	0.30

These are estimated from the email documents themselves.

- Out of all the spam emails, how many of them contain "buy"?
- Out of all the ham emails, how many of them contain the word "buy"?

In practice, the word list will be much longer – a dictionary of sorts.

Multinomial Naïve Bayes

P(spam)	0.7
P(ham)	0.3

	Spam	Ham
buy	0.20	0.10
products	0.10	0.19
please	0.50	0.01
promise	0.10	0.10
DLSU	0.05	0.30
learn	0.05	0.30

Test Instance:

Buy products from us please

Multinomial Naïve Bayes

P(spam)	0.7
P(ham)	0.3

	Spam	Ham
buy	0.20	0.10
products	0.10	0.19
please	0.50	0.01
promise	0.10	0.10
DLSU	0.05	0.30
learn	0.05	0.30

Test Instance:

Buy products from us please

	Buy	products	from	us	please
spam	0.20	0.10			0.50
ham	0.10	0.19			0.01

$$P(spam|X) = (0.20)(0.10)(0.50)(0.7) = 0.007$$

$$P(ham|X) = (0.10)(0.19)(0.01)(0.3) = 0.000057$$

Multinomial Naïve Bayes

P(spam)	0.7
P(ham)	0.3

	Spam	Ham
buy	0.20	0.10
products	0.10	0.19
please	0.50	0.01
promise	0.10	0.10
DLSU	0.05	0.30
learn	0.05	0.30

Test Instance:

Buy products from us please

	Buy	products	from	us	please
spam	0.20	0.10			0.50
ham	0.10	0.19			0.01

$$P(spam|X) = (0.20)(0.10)(0.50)(0.7) = 0.007$$

$$P(ham|X) = (0.10)(0.19)(0.01)(0.3) = 0.000057$$

Prediction: The email is spam!

Maximum A Posteriori Estimation (MAP)

Problem: If a certain feature value never appears for a given class, it will have a probability of 0.

Student	Math grade	Hours studying	ML grade
0	Bad	>= 4hrs	passed
1	Good	>= 4hrs	passed
2	Bad	< 4hrs	failed
3	Bad	>= 4hrs	passed
4	Good	>= 4hrs	passed
5	Bad	>= 4hrs	failed
6	Good	< 4hrs	passed
7	Bad	< 4hrs	failed
8	Good	>= 4hrs	passed
9	Bad	>= 4hrs	failed

Parameters of NB Model:

Given	Grade = Good	Grade = Bad	Hours >= 4hrs	Hours < 4 hrs
Passed				
Failed				

P(ML = Pass)	
P(ML = Fail)	

Now, every time the test instance has Math Grade = Good, the probability of failing will always be 0, ignoring all other features!

Maximum A Posteriori Estimation (MAP)

- MAP allows to inject a prior "belief" to the probabilities, so that the probability is not automatically 0 even if there is no observation.

- β_H and β_T are hyperparameters. Usually they are set to
- $\beta_H = 2, \beta_T = 2$ or $\beta_H = 1, \beta_T = 1$.

$$\hat{\theta}_{MAP} = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H - 1 + \alpha_T + \beta_T - 1}$$

Maximum A Posteriori Estimation (MAP)

Student	Math grade	Hours studying	ML grade
0	Bad	>= 4hrs	passed
1	Good	>= 4hrs	passed
2	Bad	< 4hrs	failed
3	Bad	>= 4hrs	passed
4	Good	>= 4hrs	passed
5	Bad	>= 4hrs	failed
6	Good	< 4hrs	passed
7	Bad	< 4hrs	failed
8	Good	>= 4hrs	passed
9	Bad	>= 4hrs	failed

Parameters of NB Model:

Given	Grade = Good	Grade = Bad	Hours >= 4hrs	Hours < 4 hrs
Passed				
Failed				

P(ML = Pass)	
P(ML = Fail)	

$\beta = 2$

$$\hat{\theta}_{MAP} = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H - 1 + \alpha_T + \beta_T - 1}$$

Given	Grade = Good	Grade = Bad	Hours = >= 4 hrs	Hours = < 4 hrs
Passed				
Failed				

Maximum A Posteriori Estimation (MAP)

- There are also ways to incorporate MAP to continuous and multinomial features.

  - For continuous features, Gaussian distribution is used.
  - For multinomial features, Dirichlet distribution is used.

- More information: https://medium.com/@gokcenazakyol/mle-map-naive-bayes-machine-learning-7-2e13b27ba14f

Naïve Bayes Pipeline