
COMP 472: Artificial Intelligence

Machine Learning

Naive Bayes Classification *video #2*

- Russell & Norvig: Sections 12.2 to 12.6

Today

1. Introduction to ML
2. **Naïve Bayes Classification**
 - a. Application to Spam Filtering
3. Decision Trees
4. (Evaluation
5. Unsupervised Learning)
6. Neural Networks
 - a. Perceptrons
 - b. Multi Layered Neural Networks



Motivation

- How do we represent and reason when there is uncertainly in the necessary knowledge?
 - It might rain tonight
 - If you have red spots on your face, you might have the measles
 - This e-mail is most likely spam
 - I can't read this character, but it looks like a "B"
 - These 2 pictures are very likely of the same person
 - ...
- One way, is to use probability theory

Remember...

■ P is a probability function:

- $0 \leq P(A) \leq 1$
- $P(A) = 0 \Rightarrow$ the event A will never take place
- $P(A) = 1 \Rightarrow$ the event A must take place
- $\sum_i P(A_i) = 1 \Rightarrow$ one of the outcomes A_i will take place
- $P(A) + P(\sim A) = 1$



■ Joint probability

- intersection $A_1 \cap \dots \cap A_n$ is an event that takes place if all the events A₁, ..., A_n take place
- denoted $P(A \cap B)$ or $P(A, B)$

■ Sum Rule

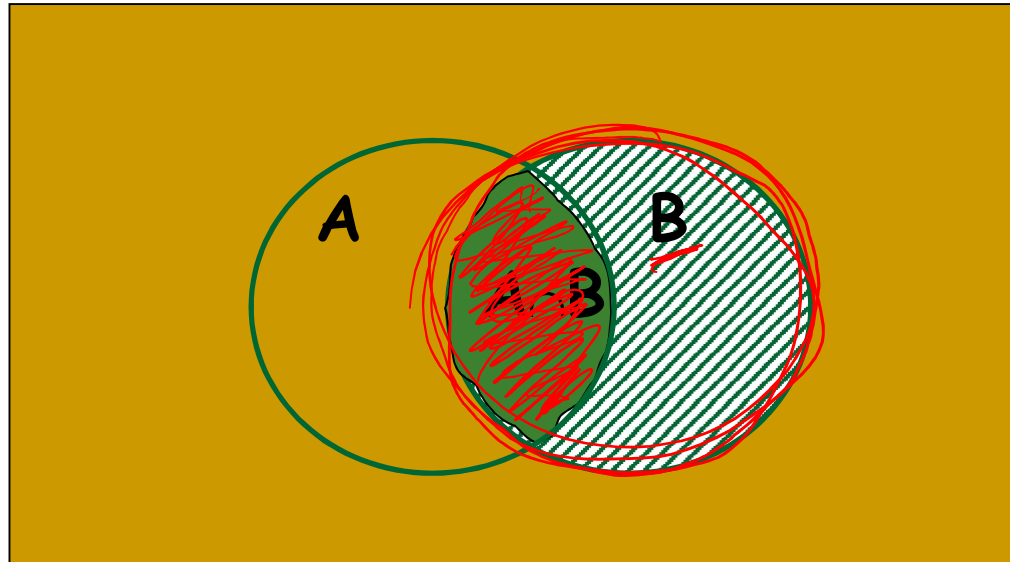
- union $A_1 \cup \dots \cup A_n$ is an event that takes place if at least one of the events A₁, ..., A_n takes place
- denoted $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Conditional Probability

- Prior (or unconditional) probability
 - Probability of an event before any evidence is obtained
 - $P(A) = 0.1$ $P(\text{rain today}) = 0.1$
 - i.e. Your belief about A given that you have no evidence
- Posterior (or conditional) probability
 - Probability of an event given that you know that B is true (B = some evidence)
 - $P(A|B) = 0.8$ $P(\text{rain today} | \text{cloudy}) = 0.8$
 - i.e. Your belief about A given that you know B

Conditional Probability (con't)

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



Chain Rule

- With 2 events, the probability that A and B occur is:

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

- With 3 events, the probability that A, B and C occur is:
 - The probability that A occurs
 - Times, the probability that B occurs, assuming that A occurred
 - Times, the probability that C occurs, assuming that A and B have occurred

- With n events, we can generalize to the Chain rule:

$$P(A_1, A_2, A_3, A_4, \dots, A_n)$$

$$= P(\cap A_i)$$

$$= P(A_1) \times P(A_2 | A_1) \times P(A_3 | A_1, A_2) \times \dots \times P(A_n | A_1, A_2, A_3, \dots, A_{n-1})$$

So what?

- we can do probabilistic inference
 - i.e. infer new knowledge from observed evidence

Example 1

- Joint probability distribution:

$P(\text{Toothache} \cap \text{Cavity})$

evidence

hypothesis

	Toothache	\sim Toothache
Cavity	0.04	0.06
\sim Cavity	0.01	0.89

$P(\text{toothache}) = 0.05$ $P(\sim \text{toothache}) = 0.9$ $\Sigma = 1$
 $P(\text{cavity}) = 0.1$ $P(\sim \text{cavity}) = 0.9$

$$P(H|E) = \frac{P(H \cap E)}{P(E)}$$

$$P(\text{cavity} | \text{toothache}) = \frac{P(\text{cavity} \cap \text{toothache})}{P(\text{toothache})} = \frac{0.04}{0.04 + 0.01} = 0.8$$

Getting the Probabilities

- in most applications, you just count from a set of observations

$$\underline{P(A)} = \frac{\underline{\text{count_of_A}}}{\underline{\text{count_of_all_events}}}$$

$$\underline{P(A|B)} = \frac{\underline{P(A \cap B)}}{\underline{P(B)}} = \frac{\underline{\text{count_of_A_and_B_together}}}{\underline{\text{count_of_all_B}}}$$

Combining Evidence

- Assume now 2 pieces of evidence:
- Suppose, we know that
 - $P(\text{Cavity} \mid \text{Toothache}) = 0.12$
 - $P(\text{Cavity} \mid \text{Young}) = 0.18$
- A patient complains about Toothache and is Young...
 - what is $P(\text{Cavity} \mid \text{Toothache} \cap \text{Young})$?

Combining Evidence

	Toothache		~Toothache		evidence #1
	Young	~ Young	Young	~ Young	evidence #2
Cavity	0.108	0.012	0.072	0.008	
~Cavity	0.016	0.064	0.144	0.576	

$P(\text{Toothache} \cap \text{Cavity} \cap \text{Young})$

- But how do we get the data ?
- In reality, we may have dozens, hundreds of variables
- We cannot have a table with the probability of all possible combinations of variables
 - Ex. with 16 binary variables, we would need 2^{16} entries

Independent Events

- In real life:

- some variables are independent...

- eg: living in Montreal & tossing a coin

- $P(\text{Montreal, head}) = P(\text{Montreal}) * P(\text{head})$

- eg: probability of tossing 2 heads in a row

- $P(\text{head, head}) = 1/2 * 1/2 = 1/4$

- some variables are not independent...

- eg: living in Montreal & wearing boots

- $P(\text{Montreal, boots}) \neq P(\text{Montreal}) * P(\text{boots})$

Independent Events

- Two events A and B are independent:
 - if the occurrence of one of them does not influence the occurrence of the other
 - i.e. A is independent of B if $P(A) = P(A|B)$
 - If A and B are independent, then:
 - $P(A, B) = P(A|B) \times P(B)$ (by chain rule) // see previous slide 7
= $P(A)$ $\times P(B)$ (by independence)
- To make things work in real applications, we often assume that events are independent
 - $P(A, B) = P(A) \times P(B)$

Conditional Independent Events

- Two events A and B are conditionally independent given C :
 - Given that C is true, then any evidence about B cannot change our belief about A
 - $P(A, B \mid C) = P(A \mid C) \times P(B \mid C)$.

when C is True

Bayes' Theorem

■ given:

$$P(A|B) = \frac{P(A,B)}{P(B)} \quad \text{so } P(A,B) = P(A|B) \times P(B)$$
$$P(B|A) = \frac{P(A,B)}{P(A)} \quad \text{so } P(A,B) = P(B|A) \times P(A)$$

■ then:

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

■ and:

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

So?

- We typically want to know: $P(\text{Hypothesis} | \text{Evidence})$
 - $P(\text{Disease} | \text{Symptoms}) \dots P(\text{meningitis} | \text{red spots})$
 - $P(\text{Cause} | \text{Side Effect}) \dots P(\text{misaligned brakes} | \text{squeaky wheels})$
- But $P(\text{Hypothesis} | \text{Evidence})$ is hard to gather
 - ex: out of all people who have red spots... how many have meningitis?
- However $P(\text{Evidence} | \text{Hypothesis})$ is easier to gather
 - ex: out of all people who have the meningitis ... how many have red spots?
- So
$$P(\text{Hypothesis} | \text{Evidence}) = \frac{P(\text{Evidence} | \text{Hypothesis}) \times P(\text{Hypothesis})}{P(\text{Evidence})}$$

Example 2

Assume we only have 1 hypothesis

Assume:

- $P(\text{spots=yes} \mid \text{meningitis=yes}) = 0.4$

- $P(\text{meningitis=yes}) = 0.00003$

- $P(\text{spots=yes}) = 0.05$

$$P(\text{meningitis=yes} \mid \text{spots=yes})$$

$$= \frac{P(\text{spots=yes} \mid \text{meningitis=yes}) \times P(\text{meningitis=yes})}{P(\text{spots=yes})}$$

$$= \frac{0.4 \times 0.00003}{0.05} = 0.00024$$

→ If you have spots... you are more likely to have meningitis than if we don't know about you having spots

Example 3

- Predict the weather tomorrow based on tonight's sunset...
- Assume we have 3 hypothesis...
 - H_1 : weather will be nice $P(H_1) = 0.2$
 - H_2 : weather will be bad $P(H_2) = 0.5$
 - H_3 : weather will be mixed $P(H_3) = 0.3$
- And 1 piece of evidence with 3 possible values
 - E_1 : today, there's a beautiful sunset
 - E_2 : today, there's a average sunset
 - E_3 : today, there's no sunset

evidence
feature

$$P(E_2 | H_1)$$

$P(E_x H_i)$	E_1	E_2	E_3
H_1	0.7	0.2	0.1
H_2	0.3	0.3	0.4
H_3	0.4	0.4	0.2

Example 3

- Observation: average sunset (E_2)
- Question: how will be the weather tomorrow?
 - $P(H_i | E_2)$?
 - predict the weather that maximizes the probability
 - select H_i such that $P(H_i | E_2)$ is the greatest

$$P(H_i | E_2) = \frac{P(H_i) \times P(E_2 | H_i)}{P(E_2)}$$

$$P(E_2) = P(H_1) \times P(E_2 | H_1) + P(H_2) \times P(E_2 | H_2) + P(H_3) \times P(E_2 | H_3)$$
$$= .2 \times .2 + .5 \times .3 + .3 \times .4 = .04 + .15 + .12 = 0.31$$

Example 3

$P(H_1 E_2) = \frac{P(H_1) \times P(E_2 H_1)}{P(E_2)} = \frac{.2 \times .2}{.31} = .129$
$P(H_2 E_2) = \frac{P(H_2) \times P(E_2 H_2)}{P(E_2)} = \frac{.5 \times .3}{.31} = .484$
$P(H_3 E_2) = \frac{P(H_3) \times P(E_2 H_3)}{P(E_2)} = \frac{.3 \times .4}{.31} = .387$

*H₂ will still have the highest score
the argument*

⇒ H₂ is the most likely hypothesis, given the evidence

P(H₂ | E₂) is the highest

Tomorrow the weather will be bad *H₂*

$$H_{NB} = \operatorname{argmax}_{H_i} \frac{P(H_i) \times P(E | H_i)}{P(E)}$$

select H_i that maximize the function

H₂

Bayes' Reasoning

- Out of n hypothesis...
 - we want to find the most probable H_i given the evidence E
- So we choose the H_i with the largest $P(H_i|E)$

$$H_{NB} = \underset{H_i}{\operatorname{argmax}} P(H_i | E) = \underset{H_i}{\operatorname{argmax}} \frac{P(H_i) \times P(E | H_i)}{P(E)}$$

- But... $P(E)$
 - is the same for all possible H_i (and is hard to gather anyways)
 - so we can drop it
- So Bayesian reasoning:

$$H_{NB} = \underset{H_i}{\operatorname{argmax}} \frac{P(H_i) \times P(E | H_i)}{P(E)} = \underset{H_i}{\operatorname{argmax}} P(H_i) \times P(E | H_i)$$

prior

Representing the Evidence

- The evidence is typically represented by many attributes/features
100's 1000's
- so often represented as a feature/attribute vector
=

	evidence					hypothesis
	sunset	clouds	temp	summer		weather
	a ₁	a ₂	a ₃	a ₄		tomorrow
<i>e1</i>	<u>beautiful</u>	<u>no</u>	<u>high</u>	<u>yes</u>		<i>Nice</i>

- e1 = < sunset:beautiful, clouds:no, temp:high, summer:yes >
features values

Combining Evidence

toothache	young	cavity
yes	yes	?

$$P(\text{Cavity} = \text{yes} | \text{Toothache} = \text{yes} \cap \text{Young} = \text{yes}) = ?$$

with Bayes Rule :

$$= \frac{P(\text{Toothache} = \text{yes} \cap \text{Young} = \text{yes} | \text{Cavity} = \text{yes}) \times P(\text{Cavity} = \text{yes})}{P(\text{Toothache} = \text{yes} \cap \text{Young} = \text{yes})}$$

with independence assumption :

$$= \frac{P(\text{Toothache} = \text{yes} | \text{Cavity} = \text{yes}) \times P(\text{Young} = \text{yes} | \text{Cavity} = \text{yes}) \times P(\text{Cavity} = \text{yes})}{P(\text{Toothache} = \text{yes}) \times P(\text{Young} = \text{yes})}$$

with conditional independence assumption :

$$= \frac{P(\text{Toothache} = \text{yes} | \text{Cavity} = \text{yes}) \times P(\text{Young} = \text{yes} | \text{Cavity} = \text{yes}) \times P(\text{Cavity} = \text{yes})}{P(\text{Toothache} = \text{yes}) \times P(\text{Young} = \text{yes})}$$

Now we have decomposed the joint probability distribution into much smaller pieces...

Combining Evidence

^{a₁} _{e₁} toothache	^{a₂} _{e₂} young	cavity
yes	yes	yes? or no?

But since we only care about ranking the hypothesis...

? $P(\text{Cavity} = \text{yes} \mid \text{Toothache} = \text{yes} \cap \text{Young} = \text{yes})$

$P(\text{Cavity} = \text{no} \mid \text{Toothache} = \text{yes} \cap \text{Young} = \text{yes})$

?
$$\frac{P(\text{Cavity} = \text{yes}) \times P(\text{Toothache} = \text{yes} \mid \text{Cavity} = \text{yes}) \times P(\text{Young} = \text{yes} \mid \text{Cavity} = \text{yes})}{P(\text{Toothache} = \text{yes}) \times P(\text{Young} = \text{yes})}$$

$$\frac{P(\text{Cavity} = \text{no}) \times P(\text{Toothache} = \text{yes} \mid \text{Cavity} = \text{no}) \times P(\text{Young} = \text{yes} \mid \text{Cavity} = \text{no})}{P(\text{Toothache} = \text{yes}) \times P(\text{Young} = \text{yes})}$$

? $P(\text{Cavity} = \text{yes}) \times P(\text{Toothache} = \text{yes} \mid \text{Cavity} = \text{yes}) \times P(\text{Young} = \text{yes} \mid \text{Cavity} = \text{yes})$

$P(\text{Cavity} = \text{no}) \times P(\text{Toothache} = \text{yes} \mid \text{Cavity} = \text{no}) \times P(\text{Young} = \text{yes} \mid \text{Cavity} = \text{no})$

$$H_{\text{NB}} = \underset{H_i}{\operatorname{argmax}} \frac{P(H_i) \times P(E \mid H_i)}{P(E)} = \underset{H_i}{\operatorname{argmax}} P(H_i) \times P(E \mid H_i) = \underset{H_i}{\operatorname{argmax}} P(H_i) \times P(\langle a_1, a_2, a_3, \dots, a_n \rangle \mid H_i) = \underset{H_i}{\operatorname{argmax}} P(H_i) \times \prod_{j=1}^n P(a_j \mid H_i)$$

Example 4

many pieces of evidence

features / attributes

2 hypothesis
2 classes

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

14 days

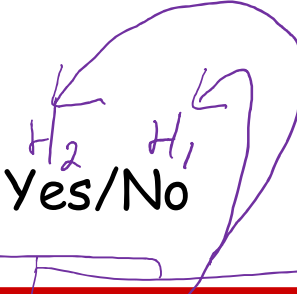
H_1 = no tennis
 H_2 = play tennis

$$P(H_2) = \frac{9}{14}$$

$$P(H_1) = \frac{5}{14}$$

Example 4

- Goal: Given a new instance $X = \langle a_1, \dots, a_n \rangle$, classify as Yes/No


$$H_{NB} = \operatorname{argmax}_{H_i} \frac{P(H_i) \times P(E | H_i)}{P(E)} = \operatorname{argmax}_{H_i} P(H_i) \times P(E | H_i) = \operatorname{argmax}_{H_i} P(H_i) \times P(\langle a_1, a_2, a_3, \dots, a_n \rangle | H_i) = \operatorname{argmax}_{H_i} P(H_i) \times \prod_{j=1}^n P(a_j | H_i)$$

- Naïve Bayes: Assumes that the attributes/features are conditionally independent given the hypothesis

Example 4

- Goal: Given a new instance $X = \langle a_1, \dots, a_n \rangle$, classify as Yes/No

$$H_{NB} = \operatorname{argmax}_{H_i} P(H_i) \times \prod_{j=1}^n P(a_j | H_i)$$

1. 1st estimate the probabilities from the training examples:
 - a) For each hypothesis H_i estimate $P(H_i)$
 - b) For each attribute value a_j of each instance (evidence) estimate $P(a_j | H_i)$

Example 4

1. TRAIN:

- compute the probabilities from the training set

$$P(\text{PlayTennis} = \text{yes}) = 9/14 = 0.64$$

$$P(\text{PlayTennis} = \text{no}) = 5/14 = 0.36$$

prior probabilities $P(H_i)$

feature H₁ value

$$P(\text{Out} = \text{sunny} \mid \text{PlayTennis} = \text{yes}) = 2/9 = 0.22$$

$$P(\text{Out} = \text{sunny} \mid \text{PlayTennis} = \text{no}) = 3/5 = 0.60$$

$$P(\text{Out} = \text{rain} \mid \text{PlayTennis} = \text{yes}) = 3/9 = 0.33$$

$$P(\text{Out} = \text{rain} \mid \text{PlayTennis} = \text{no}) = 2/5 = 0.4$$

P(out = overcast)
...

$$P(\text{Wind} = \text{strong} \mid \text{PlayTennis} = \text{yes}) = 3/9 = 0.33$$

$$P(\text{Wind} = \text{strong} \mid \text{PlayTennis} = \text{no}) = 3/5 = 0.60$$

conditional probabilities

$P(a_j \mid H_i)$

Example 4

2. TEST:

classify the new case: $X = (\text{Outlook: Sunny, Temp: Cool, Hum: High, Wind: Strong})$

$$H_{NB} = \underset{H_i \in [\text{yes}, \text{no}]}{\operatorname{argmax}} P(H_i) \times P(X | H_i)$$

$$= \underset{H_i \in [\text{yes}, \text{no}]}{\operatorname{argmax}} P(H_i) \times \prod_j P(a_j | H_i)$$

$$= \underset{H_i \in [\text{yes}, \text{no}]}{\operatorname{argmax}} P(H_i) \times P(\text{Outlook} = \text{sunny} | H_i) \times P(\text{Temp} = \text{cool} | H_i)$$

$$\times P(\text{Humidity} = \text{high} | H_i) \times P(\text{Wind} = \text{strong} | H_i)$$

$$1) P(\text{PlayTennis} = \text{yes})$$

$$\times P(\text{Outlook} = \text{sunny} | \text{PlayTennis} = \text{yes}) \times P(\text{Temp} = \text{cool} | \text{PlayTennis} = \text{yes}) \times P(\text{Hum} = \text{high} | \text{PlayTennis} = \text{yes}) \times P(\text{Wind} = \text{strong} | \text{PlayTennis} = \text{yes})$$

$$= 0.0053$$

$$2) P(\text{PlayTennis} = \text{no})$$

$$\times P(\text{Outlook} = \text{sunny} | \text{PlayTennis} = \text{no}) \times P(\text{Temp} = \text{cool} | \text{PlayTennis} = \text{no}) \times P(\text{Hum} = \text{high} | \text{PlayTennis} = \text{no}) \times P(\text{Wind} = \text{strong} | \text{PlayTennis} = \text{no})$$

$$= 0.0206$$

\Rightarrow answer : $\text{PlayTennis}(X) = \text{no}$

prior

conditional

should have been computed in the previous slide

score H_2

score H_1

Application of Bayesian Reasoning

■ Categorization: $P(\text{Category} \mid \text{Features of Object})$

- Diagnostic systems: $P(\text{Disease} \mid \text{Symptoms})$
- Text classification: $P(\text{sports_news} \mid \text{text})$ *< .t., tennis, smoke*
- Character recognition: $P(\text{character} \mid \text{bitmap})$
- Speech recognition: $P(\text{words} \mid \text{acoustic signal})$
- Image processing: $P(\text{face_person} \mid \text{image features})$
- Spam filter: $P(\text{spam_message} \mid \text{words in e-mail})$
- ...

*sports -
obituary
politics*

Digit Recognition

$< 1, 1:255, 1, 2:100, \dots \dots \dots$

$28, 28:255$

784 features

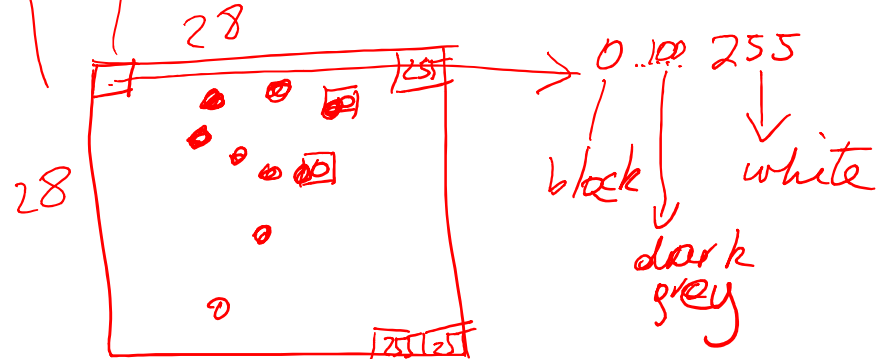
$H_0 = "0"$
 $H_1 = "1"$
 \dots
 H_9

10 hyp.

MNIST dataset

- data set contains handwritten digits from the American Census Bureau employees and American high school students
- 28 x 28 grayscale images
- training set: 60,000 examples
- test set: 10,000 examples.
- Features: each pixel is used as a feature so:
 - there are $28 \times 28 = 784$ features
 - each feature = 256 grayscale value
- Task: classify new digits into one of the 10 classes

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
9	9	9	9	9	9	9	9	9	9	9	9	9	9	9



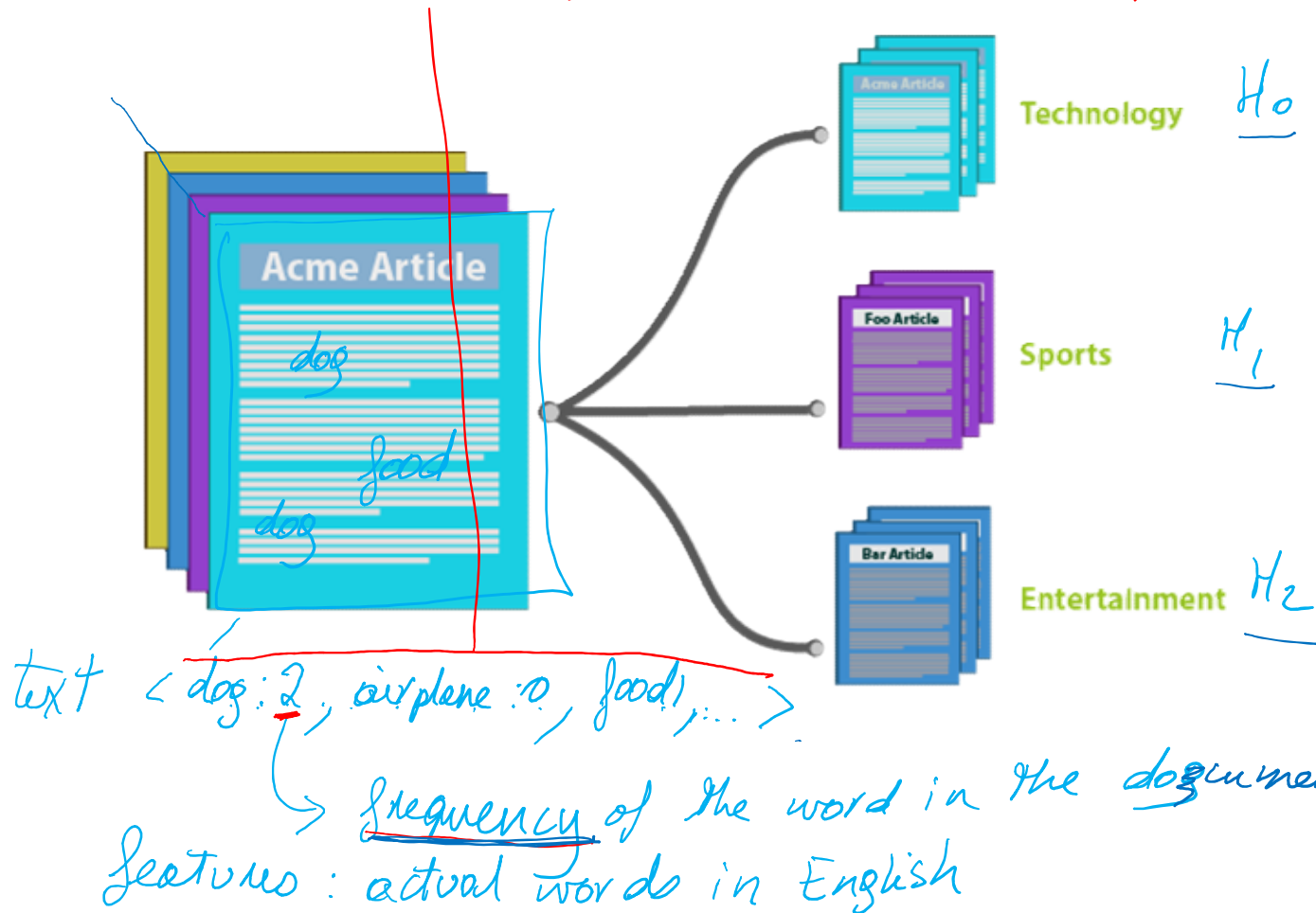
Postal Code Recognition

BAM BAM
42 T-REX RD.
PANGAEA, RB 48016

FRED FLINSTONE
69 OLD SCHOOL AVE
BEDROCK, OLDEN-TOWN
77005

Text Classification

10 000 features (i.e. 1 feature for each word in the dictionary)



Comments on Naïve Bayes Classification

- A simple probabilistic classifier based on Bayes' theorem
 - with strong (naive) independence assumption
 - i.e. the features/attributes are conditionally independent given the classes
 - eg: assumes that the word ambulance is conditionally independent of the word accident given the class SPORTS



$$p(\langle \underline{E} \rangle / H_i) \\ \hookrightarrow \prod p(\underline{e}_i / H_i)$$

- BUT:

- fast, simple
- gives confidence in its class predictions (i.e., the scores)
- surprisingly very effective on real-world tasks
- basis of many spam filters

$$\begin{aligned} \text{score}(H_1) &= 0.5 \\ \text{score}(H_2) &= 0.00 \\ \text{score}(H_3) &= 0.00001 \end{aligned}$$

Today

1. Introduction to ML 
2. Naïve Bayes Classification *video #2*
 - a. Application to Spam Filtering *video #3*
3. Decision Trees
4. (Evaluation
5. Unsupervised Learning)
6. Neural Networks
 - a. Perceptrons
 - b. Multi Layered Neural Networks

Up Next

1. Introduction to ML
2. Naive Bayes Classification
 - a. Application to Spam Filtering
3. Decision Trees
4. (Evaluation
5. Unsupervised Learning)
6. Neural Networks
 - a. Perceptrons
 - b. Multi Layered Neural Networks