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 

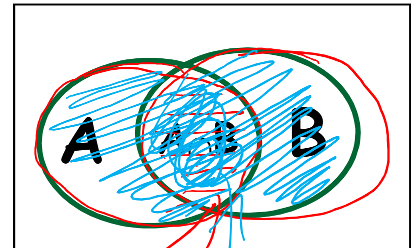
    a. Application to Spam Filtering
3. Decision Trees
4. (Evaluation
5. Unsupervised Learning)
6. Neural Networks
    a. Perceptrons
    b. Multi Layered Neural Networks

Motivation

- How do we represent and reason when the~~e~~r 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 $A_1 \cap \ldots \cap A_n$ is an event that takes place if all the events $A_1, \ldots, A_n$ take place
  - denoted $P(A \cap B)$ or $P(A,B)$
- Sum Rule
  - union $A_1 \cup \ldots \cup A_n$ is an event that takes place if at least one of the events $A_1, \ldots, A_n$ takes place
  - denoted $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

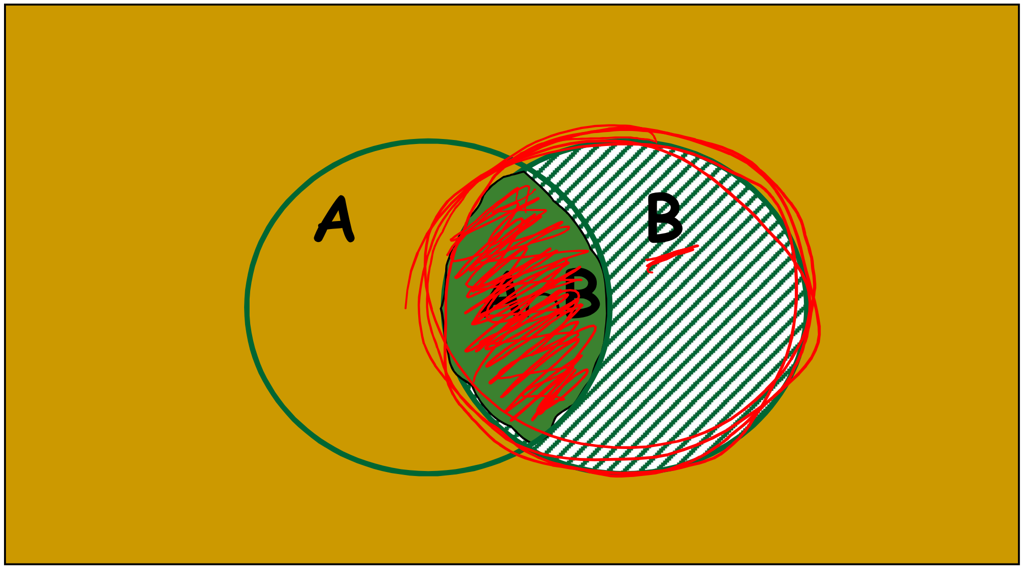
Conditional Probability

- Prior (or unconditional) probability
  - Probability of an event before any evidence is obtained
  - $P(A) = 0.1$ $\qquad$ $P(\text{rain today}) = 0.1$
  - 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 $\qquad$ (B = some evidence)
  - $P(A|B) = 0.8$ $\qquad$ $P(\text{rain today} \mid \text{cloudy}) = 0.8$
  - 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:

$$\boxed{P(A,B) = P(A \mid B) \times P(B)}$$

- 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:

$$P(A_1, A_2, A_3, A_4, \dots, A_n)$$
$$= P(\cap A_i)$$
$$= \boxed{P(A_1)} \times P(A_2 \mid A_1) \times P(A_3 \mid A_1, A_2) \times \dots \times P(A_n \mid \boxed{A_1, A_2, A_3, \dots, A_{n-1}})$$

So what?

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

Example 1

- Joint probability distribution:

P(Toothache $\cap$Cavity)

evidence

hypothesis

	Toothache	~Toothache
Cavity	0.04	0.06
~Cavity	0.01	0.89

$P(\text{cavity}) = 0.1$

$P(\neg \text{cavity}) = 0.9$

$P(\text{toothache}) = 0.05$ $\quad P(\neg \text{toothache}) = 0.9$ $\quad \Sigma = 1$

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

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

Getting the Probabilities

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

$$P(A) = \frac{count\_of\_A}{count\_of\_all\_events}$$

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{count\_of\_A\_and\_B\_together}{count\_of\_all\_B}$$

Combining Evidence

- Assume now 2 pieces of evidence:

- Suppose, we know that
  - P(Cavity | Toothache) = 0.12
  - P(Cavity | Young) = 0.18

- A patient complains about Toothache and is Young…
  - what is P(Cavity | Toothache ∩ Young) ?

Combining Evidence

	Toothache		~Toothache	
	Young	~ Young	Young	~ Young
Cavity	0.108	0.012	0.072	0.008
~Cavity	0.016	0.064	0.144	0.576

evidence #1

evidence #2

P(Toothache ∩Cavity ∩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(Montreal, head) = P(Montreal) * P(head)
    - eg: probability of tossing 2 heads in a row
      - P(head, head) = 1/2 * 1/2 = 1/4
  - some variables are not independent…
    - eg: living in Montreal & wearing boots
      - P(Montreal, boots) ≠ P(Montreal) * P(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 | C) = P(A | C) x P(B | C).

when C is True

Bayes' Theorem

- given: $P(A|B) = \dfrac{P(A,B)}{P(B)}$ so $P(A,B) = P(A|B) \times P(B)$

  $P(B|A) = \dfrac{P(A,B)}{P(A)}$ so $P(A,B) = P(B|A) \times P(A)$

- then: $P(A|B) \times P(B) = P(B|A) \times P(A)$

- and: $P(A|B) = \dfrac{P(B|A) \times P(A)}{P(B)}$

So?

- We typically want to know: P(Hypothesis | Evidence)
  - P(Disease | Symptoms)… P(meningitis | red spots)
  - P(Cause | Side Effect)… P(misaligned brakes| squeaky wheels)
- But P(Hypothesis| Evidence) is hard to gather
  - *ex: out of all people who have red spots… how many have meningitis?*
- However P(Evidence | Hypothesis) is easier to gather
  - *ex: out of all people who have the meningitis … how many have red spots?*

easier to gather

easy to gather

- So

hard to gather

$$P(\text{Hypothesis} \mid \text{Evidence}) = \frac{P(\text{Evidence} \mid \text{Hypothesis}) \times P(\text{Hypothesis})}{P(\text{Evidence})}$$

hard to gather

vaulting

Example 2

Assume we only have 1 hypothesis

Assume:

- P(spots=yes | meningitis=yes) = 0.4 — p(B|A)
- P(meningitis=yes) = 0.00003 — p(A)
- P(spots=yes) = 0.05 — p(B)

$$P(\underbrace{meningitis = yes}_{A} \mid \underbrace{spots = yes}_{B})$$

$$= \frac{P(\underbrace{spots = yes}_{B} \mid \underbrace{meningitis = yes}_{A}) \times P(\underbrace{meningitis = yes}_{A})}{P(\underbrace{spots = yes}_{B})}$$

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

p(A)

→ 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

$\boxed{P(E_2 | H_1)}$

$P(E_x \mid H_i)$	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?
  - $P(H_i \mid E_2)$?
  - predict the weather that maximizes the probability H_i $P(H_i)$
  - select H_i such that $P(H_i \mid E_2)$ is the greatest

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

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

Example 3

$$P(H_1 \mid E_2) = \frac{P(H_1) \times P(E_2 \mid H_1)}{P(E_2)} = \frac{.2 \times .2}{\cancel{.31}} = .129$$

$$P(H_2 \mid E_2) = \frac{P(H_2) \times P(E_2 \mid H_2)}{P(E_2)} = \frac{.5 \times .3}{\cancel{.31}} = .484$$

H2 will still have the highest score

$$P(H_3 \mid E_2) = \frac{P(H_3) \times P(E_2 \mid H_3)}{P(E_2)} = \frac{.3 \times .4}{\cancel{.31}} = .387$$

the argument

$\Rightarrow$ H_2 is the most likely hypothesis, given the evidence

$\boxed{P(H_2 \mid E_2)}$ is the highest

Tomorrow the weather will be bad H_2

select Hi that maximize the function

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

H_2

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:

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

Representing the Evidence

- The evidence is typically represented by many attributes/features 100's 1000's
  - beautiful sunset? clouds? temperature? summer?, …
- so often represented as a feature/attribute vector =

	evidence					hypothesis
	sunset a_1	clouds a_2	temp a_3.	summer a_4		weather tomorrow
e1	beautiful	no	high	yes		*Nice*

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

features values

Combining Evidence

toothache	young	cavity
yes	yes	?

$$P(\overbrace{Cavity = yes}^{A} \mid \overbrace{\overbrace{Toothache = yes}^{e_1} \cap \overbrace{Young = yes}^{e_2}}^{B}) = ?$$

with Bayes Rule :

$$= \frac{P(\overbrace{Toothache = yes \cap Young = yes}^{B} \mid \overbrace{Cavity = yes}^{A}) \times P(\overbrace{Cavity = yes}^{A})}{P(\underbrace{\underbrace{Toothache = yes}_{e_1} \cap \underbrace{Young = yes}_{e_2}}_{B})}$$

with independence assumption :

$$= \frac{P(\overbrace{Toothache = yes}^{A} \cap \overbrace{Young = yes}^{B} \mid \overbrace{Cavity = yes}^{C}) \times P(Cavity = yes)}{P(Toothache = yes) \times P(Young = yes)}$$

with conditional independence assumption :

100's 1000's

$$= \frac{P(\overbrace{Toothache = yes}^{A} \mid \overbrace{Cavity = yes}^{C}) \times P(\overbrace{Young = yes}^{B} \mid Cavity = yes) \times P(Cavity = yes)}{P(Toothache = yes) \times P(Young = yes)}$$

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

Combining Evidence

toothache	young	cavity
yes	yes	yes? or no?

e_1 a_1 e_2 a_2

But since we only care about ranking the hypothesis…

$$P(\text{Cavity}=\text{yes} \mid \text{Toothache}=\text{yes} \cap \text{Young}=\text{yes})$$
$$> \quad H_1 > H_2$$
$$P(\text{Cavity}=\text{no} \mid \text{Toothache}=\text{yes} \cap \text{Young}=\text{yes})$$

prior $\prod$

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

prior $\prod$

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

$P(E)$

$\prod(P(e_i))$

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

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

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

$p(H_2) = 9/14$

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

Example 4

- Goal: Given a new instance X=<a_1,…, a_n>, classify as Yes/No

H_2 H_1

$$H_{NB} = \underset{H_i}{\text{argmax}} \frac{P(H_i) \times P(E|H_i)}{P(E)} = \underset{H_i}{\text{argmax}}\ P(H_i) \times P(E|H_i) = \underset{H_i}{\text{argmax}}\ P(H_i) \times P(<a_1, a_2, a_3, \ldots, a_n>|H_i) = \underset{H_i}{\text{argmax}}\ P(H_i) \times \prod_{j=1}^{n} P(a_j|H_i)$$

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

Example 4

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

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

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 $P(a_j \mid H_i)$

Example 4

1. TRAIN:
   - compute the probabilities from the training set

H_2

$P(PlayTennis = yes) = 9/14 = 0.64$
$P(PlayTennis = no) = 5/14 = 0.36$

H_1

prior probabilities P(H_i)

feature value H_2

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

H_1

$P(Out = sunny \mid PlayTennis = no) = 3/5 = 0.60$
$P(Out = rain \mid PlayTennis = yes) = 3/9 = 0.33$
$P(Out = rain \mid PlayTennis = no) = 2/5 = 0.4$
$P(out = overcast)$
...
$P(Wind = strong \mid PlayTennis = yes) = 3/9 = 0.33$
$P(Wind = strong \mid PlayTennis = no) = 3/5 = 0.60$

conditional probabilities

$P(a_j \mid H_i)$

Example 4

2. TEST:

prior

conditional.

classify the new case: X=(Outlook: Sunny, Temp: Cool, Hum: High, Wind: Strong)

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

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

should have been computed in the previous slide

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

2/9

$$9/14 \quad \times P(Humidity = high \mid H_i) \times P(Wind = strong \mid H_i)$$

H2 =

3/9

score H_2

1) P(PlayTennis = yes)
x P(Outlook = sunny | PlayTennis = yes)xP(Temp = cool | PlayTennis = yes)xP(Hum = high | PlayTennis = yes)xP(Wind = strong | PlayTennis = yes)
= 0.0053

H_1

5/14

score H_1

2) P(PlayTennis = no)
x P(Outlook = sunny | PlayTennis = no)xP(Temp = cool | PlayTennis = no)xP(Hum = high | PlayTennis = no)xP(Wind = strong | PlayTennis = no)
= 0.0206

3/5

3/5

$\Rightarrow$ answer : PlayTennis(X) = no

Application of Bayesian Reasoning

- Categorization: P(Category | Features of Object)
  - Diagnostic systems: P(Disease | Symptoms)
    Hn
    < t, temp, smoke >
  - Text classification: P(sports_news | text)
  - Character recognition: P(character | bitmap)
  - Speech recognition: P(words | acoustic signal)
  - Image processing: P(face_person | image features)
  - Spam filter: P(spam_message | words in e-mail)
  - …

sports

obituary

politics

Digit Recognition

H_0 = "0"
H_1 = "1"
...
H_9

10 hyp.

< 1,1: 255, 1,2: 100, ... 28,28: 255

784 features

- 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 28x28 = 784 features
  - each feature = 256 greyscale value
- Task: classify new digits into one of the 10 classes

28

28

255

0 .. 100 255

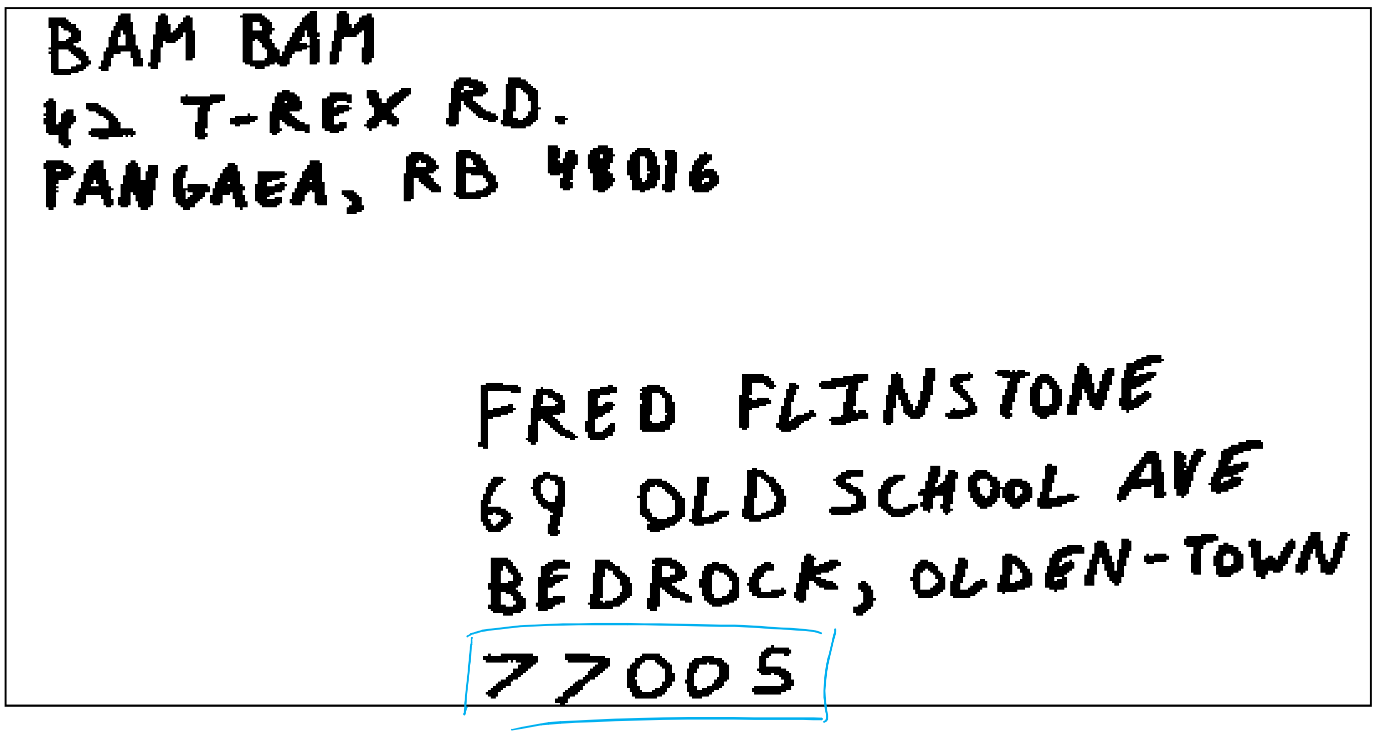
black

dark grey

white

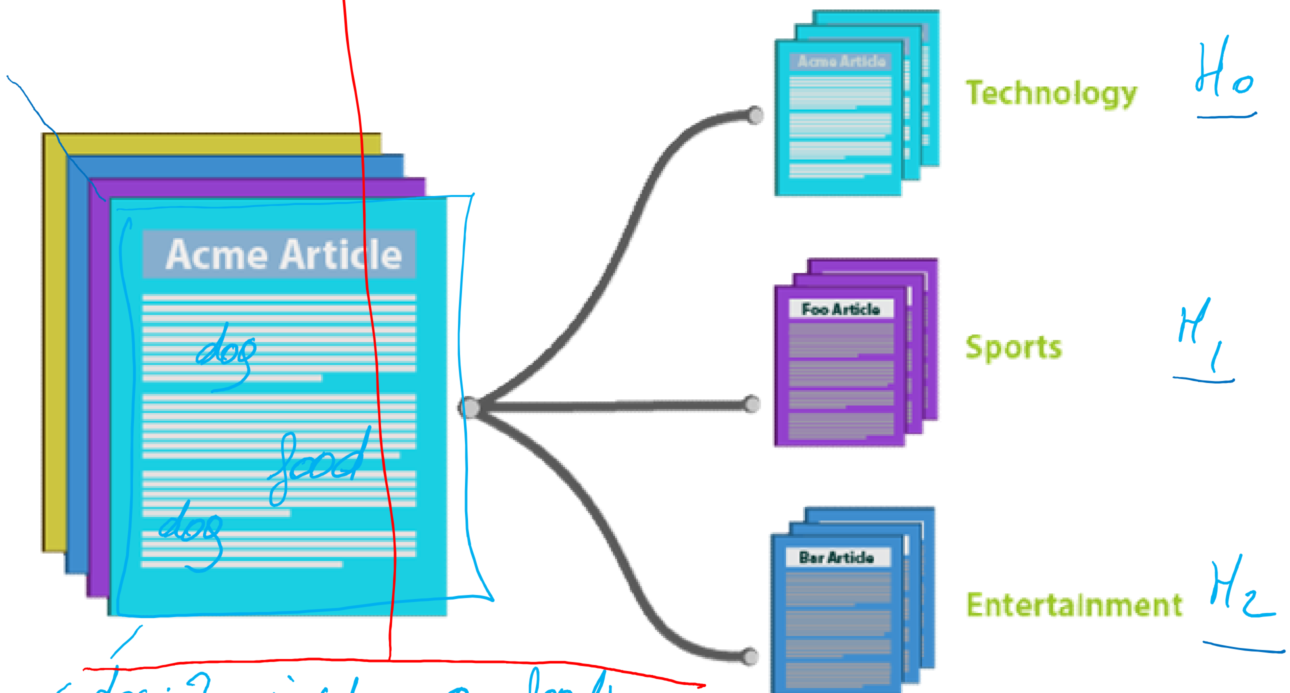
https://en.wikipedia.org/wiki/MNIST_database

Postal Code Recognition



Text Classification

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



text <dog: 2, airplane: 0, food: 1, ... >

frequency of the word in the document

features: actual words in English

https://towardsdatascience.com/machine-learning-nlp-text-classification-using-scikit-learn-python-and-nltk-c52b92a7c73a

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

$P(\langle E \rangle / H_i)$

$\approx \prod P(e_i / H_i)$

- 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

score$(H_1) = 0.5$ 0.47

score$(H_2) = 0.00$ 0.489

score$(H_3) = 0.00001$

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. 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