COMP 472: Artificial Intelligence
Machine Learning
Naive Bayes Classification

- 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

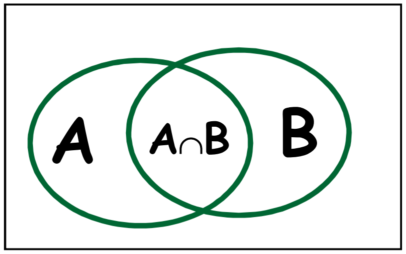
Motivation

推论

- 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 $A_1 \cap \ldots \cap A_n$ 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 先验概率，不知道任何前置条件的概率
  - Probability of an event before any evidence is obtained
  - P(A) = 0.1 *P(rain today) = 0.1*
  - i.e. Your belief about *A* given that you have no evidence

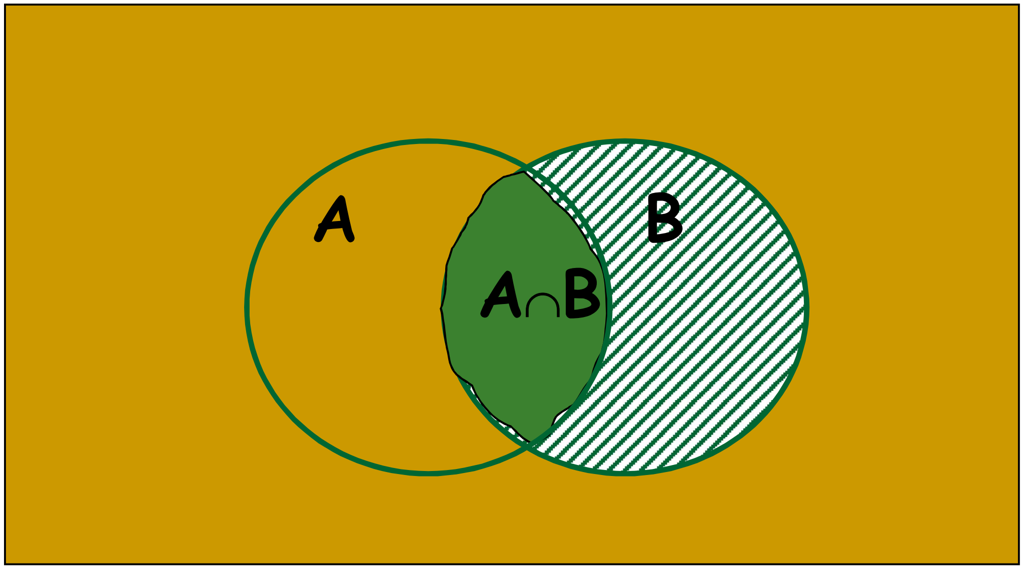
后验概率

- Posterior (or conditional) probability 一定条件下的概率
  - Probability of an event given that you know that B is true (B = some evidence)
  - P(A|B) = 0.8 *P(rain today| cloudy) = 0.8*
  - i.e. Your belief about *A* given that you know B

Conditional Probability (con't)

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

在B已经发生的情况下
发生A的概率



Chain Rule

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

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

链式法则，记忆点就是竖的乘后面那个前置条件的几率就是同时发生的几率

- 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:

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

$= P(\cap A_i)$

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

So what?

这样我们就可以做出几率上的推测了

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

Example 1

- Joint probability distribution:

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

evidence

hypothesis

	Toothache	~Toothache
Cavity 有蛀牙	0.04	0.06
~Cavity	0.01	0.89

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

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

Getting the Probabilities

- in most applications, you just count from a set of observations 绝大多数时候数据来自history data

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

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

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 $\cap$ Young) ?

Combining Evidence

这里实际上是三个variable,Tootheache,Young,Cavity,只不过这是2D图形，所以把Young放到了tootheache下面

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

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

但是variable过多的时候比并不能这样画表

- 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 不能把所有combination都表示出来
  - Ex. with 16 binary variables, we would need 2^{16} entries

Independent Events

- In real life:

有的变量是独立的，同时发生的几率就等于分别发生相乘

  - 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)

不行

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

互不影响，发生B以后发生A的概率等于发生A的概率

- If A and B are independent, then:
  - $P(A,B) = P(A|B) \times P(B)$ (by chain rule)
    $= P(A) \times P(B)$ (by independence)

- To make things work in real applications, we often assume that events are independent 为了更好的让代码运行，我们通常假设event是相互独立的
  - $P(A,B) = P(A) \times P(B)$

Conditional Independent Events

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

当C成立的时候，B便不再会影响A

Bayes' Theorem

- given: $$P(A|B) = \frac{P(A,B)}{P(B)} \text{ so } P(A,B) = P(A|B) \times P(B)$$
  $$P(B|A) = \frac{P(A,B)}{P(A)} \text{ so } P(A,B) = P(B|A) \times P(A)$$
- 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(Hypothesis | Evidence)
  - P(Disease | Symptoms)… P(meningitis | red spots) 脑膜炎 红斑
  - P(Cause | Side Effect)… P(misaligned brakes| squeaky wheels) 偏离刹车 吱嘎作响的轮胎
- 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?*
- So

Bayes theorem看起来很像废话 但是有的数据是更难得到的

容易得到，例如有脑膜炎的人里面，多高的几率红斑

有红斑的人里面，多高的几率脑膜炎

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

Example 2

Assume we only have 1 hypothesis

Assume:

- P(spots=yes | meningitis=yes) = 0.4
- P(meningitis=yes) = 0.00003
- P(spots=yes) = 0.05

$$P(meningitis = yes \mid spots = yes)$$

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

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

当你有spots的时候，你有meningitis的几率更大

→ 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*

$P(E_2 | H_1)$

$P(E_x \| H_i)$	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

他问你平均明天为average sunset的几率
=H1∩E2+H2∩E2+H3∩E2

- 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
  - select H_i such that $P(H_i \mid E_2)$ is the greatest

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

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

Example 3

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

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

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

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

$P(H_2 \mid E_2)$ is the highest

Tomorrow the weather will be bad

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

Bayes' Reasoning

当你给了证据E时，想要知道最有可能的Hi

- 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 $P(H_i|E)$

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

- 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:

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

Representing the Evidence

- The evidence is typically represented by many attributes/features
  - 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>

Combining Evidence

问你多个evidence下cavity的几率

toothache	young	cavity
yes	yes	?

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

with Bayes Rule: 首先使用bayes rule

$$= \frac{P(Toothache = yes \cap Young = yes \mid Cavity = yes) x P(Cavity = yes)}{P(Toothache = yes \cap Young = yes)}$$

换顺序

evidence part

with independence assumption:

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

下面的拆开

with conditional independence assumption:

$$= \frac{P(Toothache = yes \mid Cavity = yes) \; xP(Young = yes \mid Cavity = yes) x \; P(Cavity = yes)}{P(Toothache = yes) \text{ x } P(Young = yes)}$$

上面的拆开

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

Combining Evidence

toothache	young	cavity
yes	yes	*yes? or no?*

But since we only care about ranking the hypothesis…

$$P(Cavity = yes \mid Toothache = yes \cap Young = yes)$$
$$>$$
$$P(Cavity = no \mid Toothache = yes \cap Young = yes)$$

我们现在只关心Cavity几率的比较，revidence都成立的时候，是否有更大可能cavity

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

上一页的转化，不过Cavity=yes提前了

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

下面的一样

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

注意，H是hypothesis也就是cavity，

E是tootheache young，我们希望知道E固定时（toothache=yes且young=yes），cavity是cavity YES还是cavity no

Example 4

evidence

hypothesis

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

Example 4

然后到了第十五天，我给你a1....an， 希望你预测YES OR NO

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

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

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

这里的HNB hypothesis为YES OR NO，
也就是说当已知A1,A2...AN的时候，Hi是YES还是NO的几率更大

Example 4

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

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

注意了这个是连乘符号

A1到AN我们认为是独立的，所以是a1|hi*a2|hi....

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

$$\left.\begin{aligned}&P(PlayTennis = yes) = 9/14 = 0.64\\&P(PlayTennis = no) = 5/14 = 0.36\end{aligned}\right\}$$

prior probabilities $P(H_i)$

得到hi

$$\left.\begin{aligned}&P(Out = sunny \mid PlayTennis = yes) = 2/9 = 0.22\\&P(Out = sunny \mid PlayTennis = no) = 3/5 = 0.60\\&P(Out = rain \mid PlayTennis = yes) = 3/9 = 0.33\\&P(Out = rain \mid PlayTennis = no) = 2/5 = 0.4\\&\ldots\\&P(Wind = strong \mid PlayTennis = yes) = 3/9 = 0.33\\&P(Wind = strong \mid PlayTennis = no) = 3/5 = 0.60\end{aligned}\right\}$$

conditional probabilities

$P(a_j \mid H_i)$

Example 4

2. TEST:

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

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

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

带入

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

$$\times P(Humidity = high \mid H_i) \times P(Wind = strong \mid H_i)$$

$$\boxed{\begin{aligned} &1)\ P(PlayTennis = yes) \\ &\quad \times P(Outlook = sunny \mid PlayTennis = yes) x P(Temp = cool \mid PlayTennis = yes) x P(Hum = high \mid PlayTennis = yes) x P(Wind = strong \mid PlayTennis = yes) \\ &\quad = 0.0053 \end{aligned}}$$

$$\boxed{\begin{aligned} &2)\ P(PlayTennis = no) \\ &\quad \times P(Outlook = sunny \mid PlayTennis = no) x P(Temp = cool \mid PlayTennis = no) x P(Hum = high \mid PlayTennis = no) x P(Wind = strong \mid PlayTennis = no) \\ &\quad = 0.0206 \end{aligned}}$$

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

NO的时候HNB最大

Application of Bayesian Reasoning

也叫classification 分类 各式各样的特征

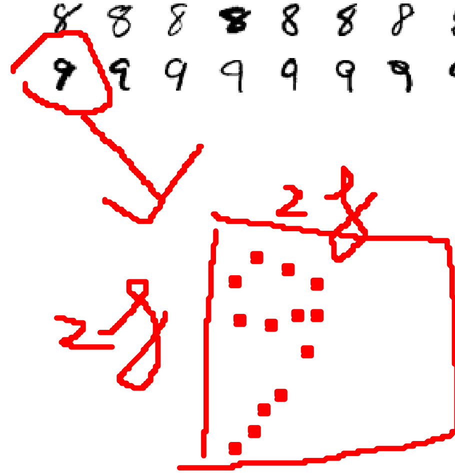
- Categorization: P(Category | Features of Object)
  - Diagnostic systems: P(Disease | Symptoms) 年龄，症状啥的
  - Text classification: P(sports_news | text)
  - Character recognition: P(character | bitmap)
  - Speech recognition: P(words | acoustic signal)
  - Image processing: P(face_person | image features)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

Digit Recognition

- [MNIST dataset](https://en.wikipedia.org/wiki/MNIST_database)
- data set contains handwritten digits from the American Census Bureau employees and American high school students
- 28 x 28 grayscale images

灰度，0到255
0是black,255是白色

- 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



9实际上是一个28*28的图片
784个pixil,每一个pixil都是一个0到255的grayscale

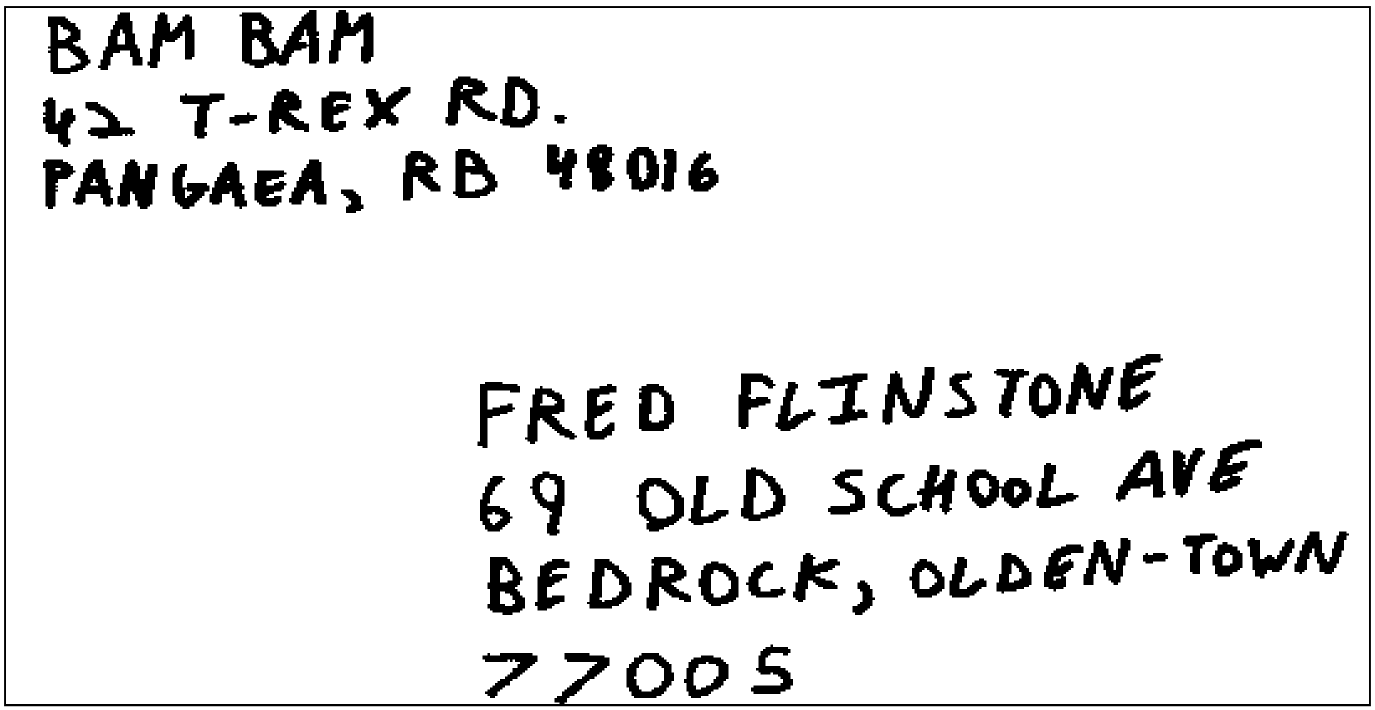
https://en.wikipedia.org/wiki/MNIST_database

1,1:255,1,2:255,1,3:255,1,4:0,..........28,28:255

784个attribute

H0=0,h1=1,h2=2,h3=.. h9=9,十个hypothesis
给你这么一组数据，让你求每个hypothesis的可能性

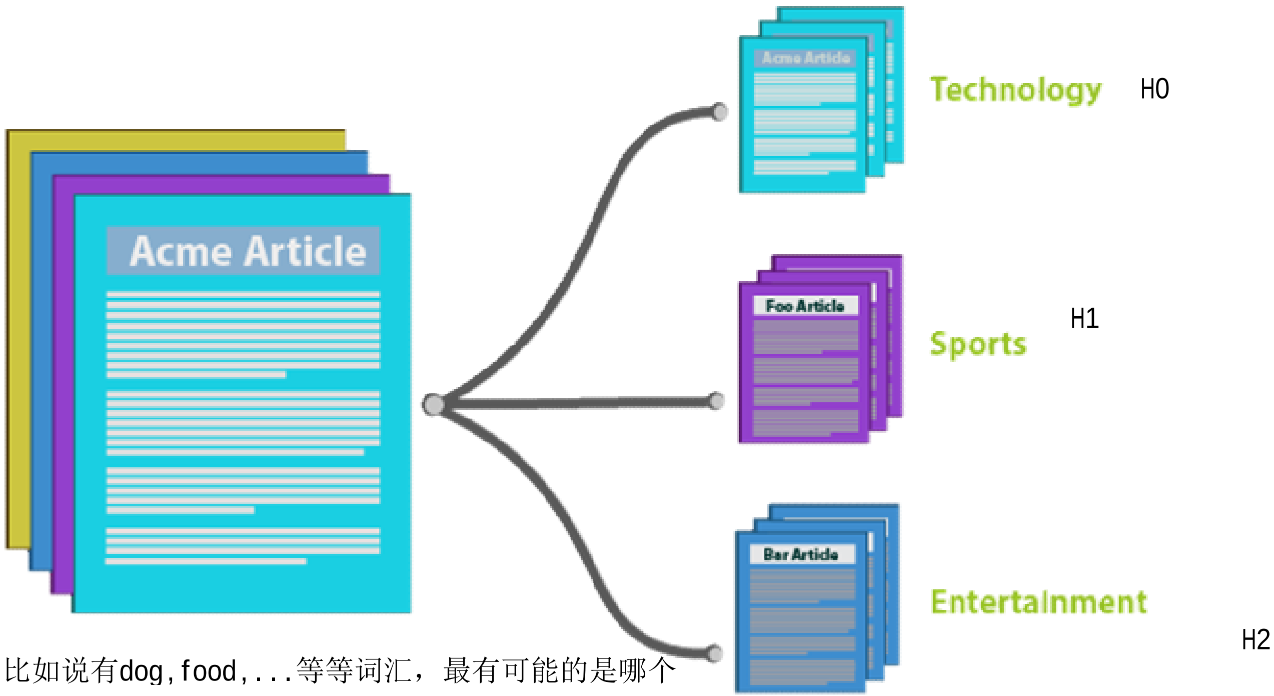
Postal Code Recognition



处理postal code时很有用

Text Classification

10000features,features for each word in the dictionary



比如说有dog,food,...等等词汇，最有可能的是哪个

TEXT<DOG:2,AIRPLANE:0,FOOD:1.....>

2是frequency of the word dog
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

上面的这类叫做naive bayes classification，是具有强独立性的假设

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

- BUT:
  - fast, simple 快速简单
  - gives confidence in its class predictions (i.e., the scores) 分类的时候很有效
  - surprisingly very effective on real-world tasks 现实世界中牛逼
  - basis of many spam filters
    垃圾邮件过滤器

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

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