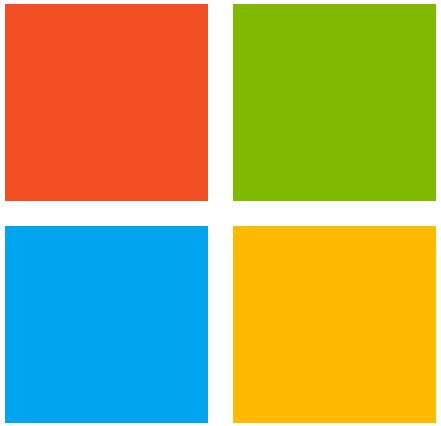


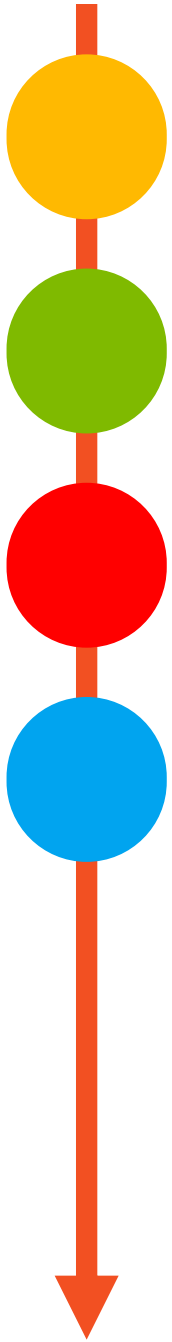
2018 TJMSC Tech. Courses

Machine Learning Algorithm

Yiran Zhuo
Tongji Microsoft Student Club

Nov 4, 2018
Room 516, Ji Shi Building
SSE, Tongji Univ

同济微软俱乐部
Tongji Microsoft Student Club



Information Entropy

Decision Tree

Bayes’ theorem

Naïve Bayes

Information Entropy

Information contained by an event can be described as

$$f(P) = \log \frac{1}{p} = -\log p$$

Thus, when the event has a lower-probability value (i.e., when a low-probability event occurs), the event carries more "information" ("surprisal") than when the event has a higher-probability value

Information Entropy

The information carried by a set of events can be described as:

$$H(X) = -\sum_{x \varepsilon X} P(x) log P(x))$$

Decision trees

Problem: decide whether to wait for a table at a restaurant, based on the following attributes:

1. Alternate: is there an alternative restaurant nearby?
2. Bar: is there a comfortable bar area to wait in?
3. Fri/Sat: is today Friday or Saturday?
4. Hungry: are we hungry?
5. Patrons: number of people in the restaurant (None, Some, Full)
6. Price: price range ($, $$, $$$)
7. Raining: is it raining outside?
8. Reservation: have we made a reservation?
9. Type: kind of restaurant (French, Italian, Thai (泰国菜), Burger)
10. WaitEstimate: estimated waiting time (0-10, 10-30, 30-60, >60)

Attribute-based representations

Examples described by attribute values (Boolean, discrete, continuous)

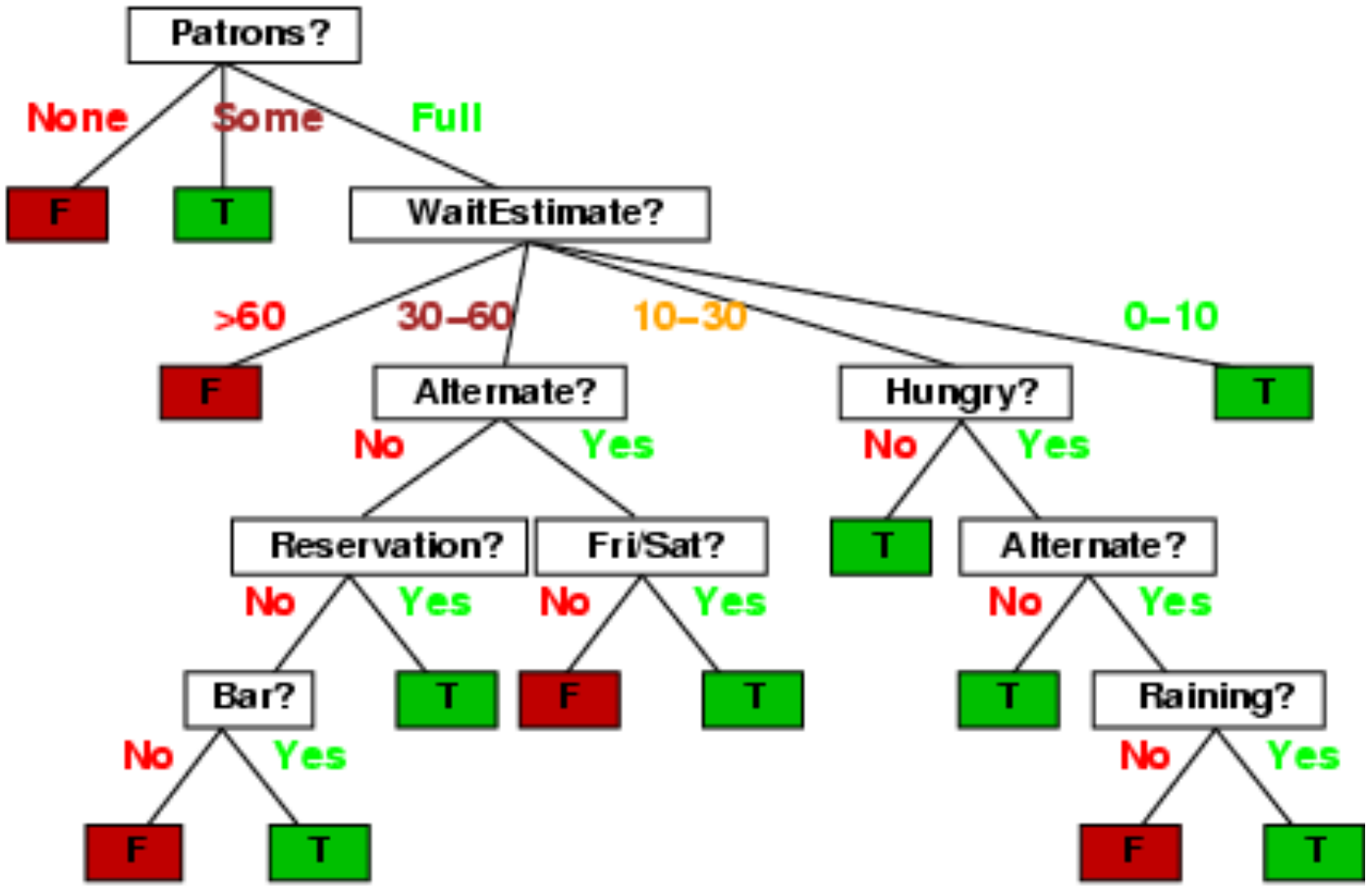
E.g., situations where I will/won't wait for a table:

Example	Attributes										Target
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*Wait*
X_1	T	F	F	T	Some	$$$	F	T	French	0–10	T
X_2	T	F	F	T	Full	$	F	F	Thai	30–60	F
X_3	F	T	F	F	Some	$	F	F	Burger	0–10	T
X_4	T	F	T	T	Full	$	F	F	Thai	10–30	T
X_5	T	F	T	F	Full	$$$	F	T	French	>60	F
X_6	F	T	F	T	Some	$$	T	T	Italian	0–10	T
X_7	F	T	F	F	None	$	T	F	Burger	0–10	F
X_8	F	F	F	T	Some	$$	T	T	Thai	0–10	T
X_9	F	T	T	F	Full	$	T	F	Burger	>60	F
X_{10}	T	T	T	T	Full	$$$	F	T	Italian	10–30	F
X_{11}	F	F	F	F	None	$	F	F	Thai	0–10	F
X_{12}	T	T	T	T	Full	$	F	F	Burger	30–60	T

Classification of examples is positive (T) or negative (F)

Decision trees

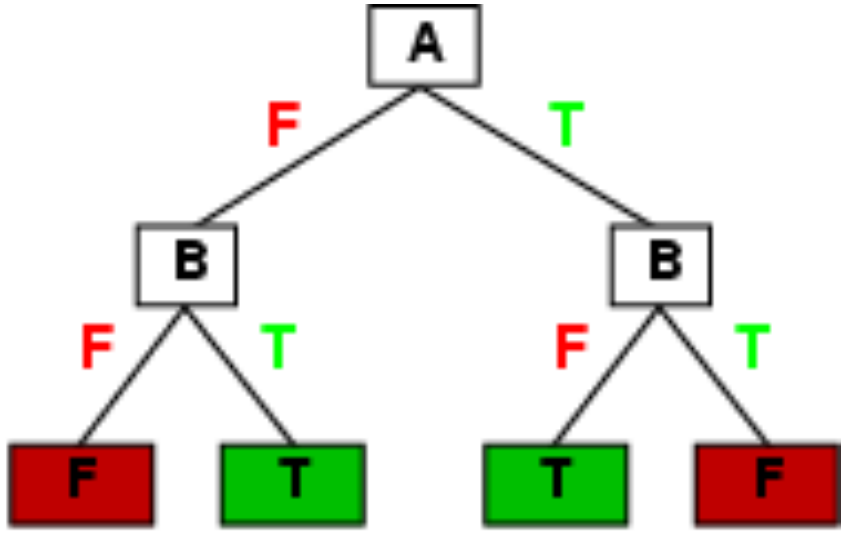
- One possible representation for hypotheses
- E.g., here is the “true” tree for deciding whether to wait:



Expressiveness

- Decision trees can **express any function of** the input attributes.
- E.g., for Boolean functions, truth table row → path to leaf:

A	B	A xor B
F	F	F
F	T	T
T	F	T
T	T	F



- Trivially, there is a consistent decision tree for any training set with one path to leaf for each example (unless *f* nondeterministic in *x*) but it probably won't generalize to new examples
- Prefer to find more compact decision trees

Decision tree learning

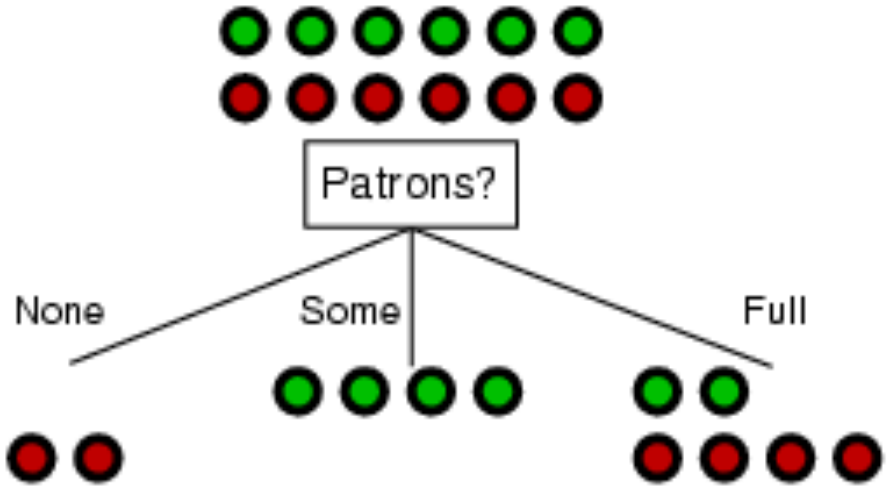
同济微软俱乐部
Tongji Microsoft Student Club

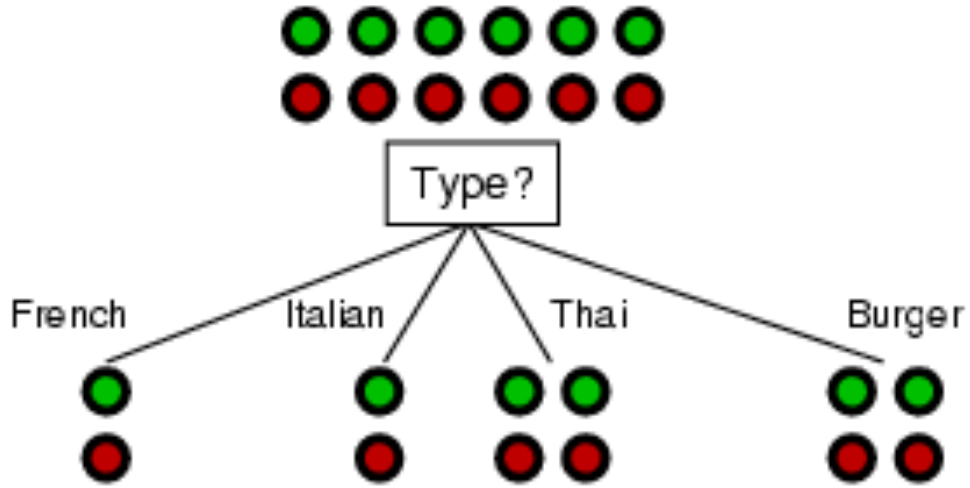
- Aim: find a small tree consistent with the training examples
- Idea: (recursively) choose "most significant" attribute as root of (sub)tree

```
检测数据集中的每个子项是否属于同一分类:
    If so return 类标签
    Else
        寻找划分数据集的最好特征
        划分数据集
        创建分支节点
            for 每个划分的子集调用函数 createBranch 并增加返回结果到分支节点中
        return 分支节点
```

Choosing an attribute

- Idea: a good attribute splits the examples into subsets that are (ideally) "all positive" or "all negative"





- *Patrons?* is a better choice

Using information theory

- To implement `Choose-Attribute` in the DTL algorithm
- Information Content (Entropy):

$$I(P(v_1), \ldots, P(v_n)) = \Sigma_{i=1} -P(v_i) \log_2 P(v_i)$$

- For a training set containing *p* positive examples and *n* negative examples:

$$I(\frac{p}{p+n}, \frac{n}{p+n}) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

Information gain

- A chosen attribute *A* divides the training set *E* into subsets $E_1, \ldots, E_v$ according to their values for *A*, where *A* has v distinct values.

$$remainder(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p + n} I\left(\frac{p_i}{p_i + n_i}, \frac{n_i}{p_i + n_i}\right)$$

- Information Gain (IG) or reduction in entropy from the attribute test:

$$IG(A) = I\left(\frac{p}{p+n}, \frac{n}{p+n}\right) - remainder(A)$$

- Choose the attribute with the largest IG

Information gain

For the training set, $p = n = 6$, $I(6/12, 6/12) = 1$ bit

Consider the attributes *Patrons* and *Type* (and others too)

Example	Attributes										Target
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*Wait*
X_1	T	F	F	T	Some	$$$	F	T	French	0–10	T
X_2	T	F	F	T	Full	$	F	F	Thai	30–60	F
X_3	F	T	F	F	Some	$	F	F	Burger	0–10	T
X_4	T	F	T	T	Full	$	F	F	Thai	10–30	T
X_5	T	F	T	F	Full	$$$	F	T	French	>60	F
X_6	F	T	F	T	Some	$$	T	T	Italian	0–10	T
X_7	F	T	F	F	None	$	T	F	Burger	0–10	F
X_8	F	F	F	T	Some	$$	T	T	Thai	0–10	T
X_9	F	T	T	F	Full	$	T	F	Burger	>60	F
X_{10}	T	T	T	T	Full	$$$	F	T	Italian	10–30	F
X_{11}	F	F	F	F	None	$	F	F	Thai	0–10	F
X_{12}	T	T	T	T	Full	$	F	F	Burger	30–60	T

Information gain

Pat:

4 Some :	4 T	0 F
6 Full :	2 T	4 F
2 None :	0 T	2 F

Type:

2 French :	1 T	1 F
2 Italian :	1 T	1 F
4 Thai :	2 T	2 F
4 Burger:	2 T	2 F

……

Patrons has the highest IG of all attributes and so is chosen by the DTL algorithm as the root

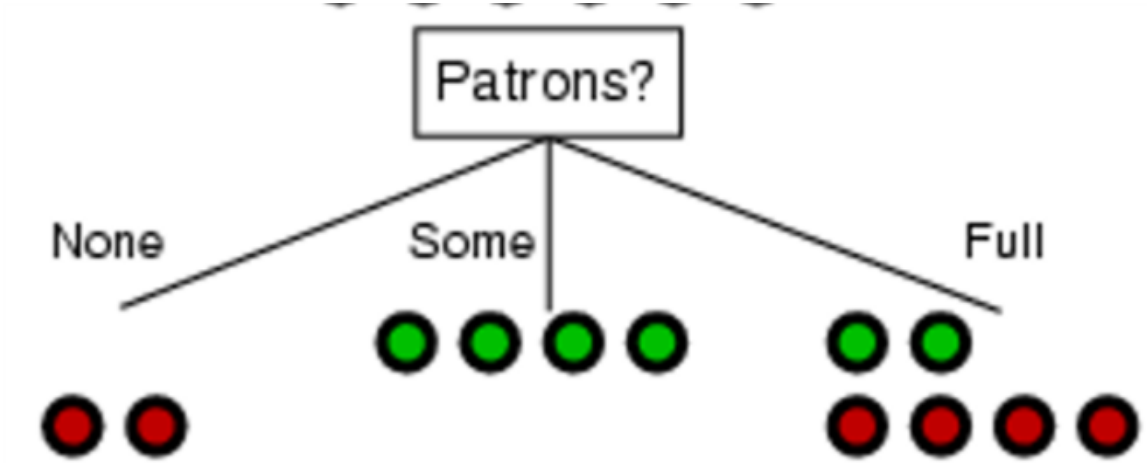
Example	Attributes										Target
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*Wait*
X_1	T	F	F	T	Some	$$$	F	T	French	0–10	T
X_2	T	F	F	T	Full	$	F	F	Thai	30–60	F
X_3	F	T	F	F	Some	$	F	F	Burger	0–10	T
X_4	T	F	T	T	Full	$	F	F	Thai	10–30	T
X_5	T	F	T	F	Full	$$$	F	T	French	>60	F
X_6	F	T	F	T	Some	$$	T	T	Italian	0–10	T
X_7	F	T	F	F	None	$	T	F	Burger	0–10	F
X_8	F	F	F	T	Some	$$	T	T	Thai	0–10	T
X_9	F	T	T	F	Full	$	T	F	Burger	>60	F
X_{10}	T	T	T	T	Full	$$$	F	T	Italian	10–30	F
X_{11}	F	F	F	F	None	$	F	F	Thai	0–10	F
X_{12}	T	T	T	T	Full	$	F	F	Burger	30–60	T

$$IG(Patrons) = 1 - [\frac{2}{12} I(0,1) + \frac{4}{12} I(1,0) + \frac{6}{12} I(\frac{2}{6}, \frac{4}{6})] = .0541 \text{ bits}$$

$$IG(Type) = 1 - [\frac{2}{12} I(\frac{1}{2}, \frac{1}{2}) + \frac{2}{12} I(\frac{1}{2}, \frac{1}{2}) + \frac{4}{12} I(\frac{2}{4}, \frac{2}{4}) + \frac{4}{12} I(\frac{2}{4}, \frac{2}{4})] = 0 \text{ bits}$$

Information gain

After choosing *Patrons* as the root, we still need to choose another node *at the situation 'Full'*



Information gain

Hun:

4 T : 2 T 2 F

2 F : 0 T 2 F

Type:

1 French : 0 T 1 F

1 Italian : 0 T 1 F

2 Thai : 1 T 1 F

2 Burger: 1 T 1 F

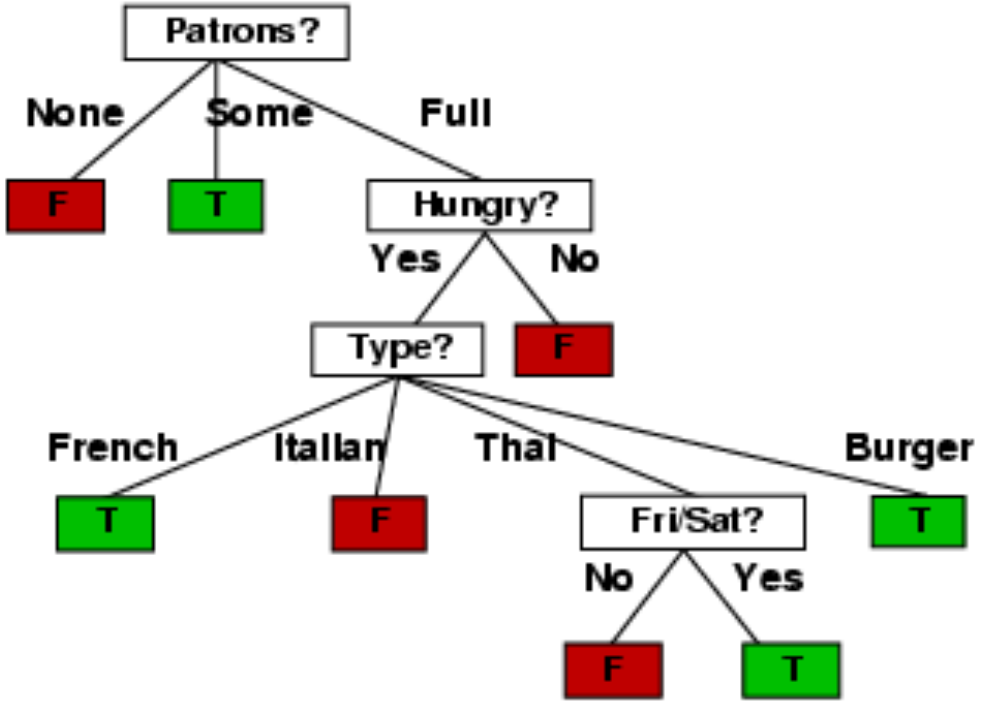
……

Hungrey has the highest IG of all attributes and so is chosen by the DTL algorithm as the root

Example	Attributes										Target
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*Wait*
X_2	T	F	F	T	Full	$	F	F	Thai	30–60	F
X_4	T	F	T	T	Full	$	F	F	Thai	10–30	T
X_5	T	F	T	F	Full	$$$	F	T	French	>60	F
X_9	F	T	T	F	Full	$	T	F	Burger	>60	F
X_{10}	T	T	T	T	Full	$$$	F	T	Italian	10–30	F
X_{12}	T	T	T	T	Full	$	F	F	Burger	30–60	T

Example contd.

- Decision tree learned from the 12 examples:



- Substantially simpler than “true” tree---a more complex hypothesis isn’t justified by small amount of data

Bayes’ theorem

Bayes' theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event

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

Naïve Bayes

Preliminaries and Notations

$\omega i \in \{ \omega 1, \omega 2, \dots , \omega c \}$: a state of nature
x : feature vector
$p(\omega i)$: prior probability
$p(x|\omega i)$: class-conditional density
$p(\omega i|x)$: posterior probability

Naïve Bayes

Fish Sorting Example

-Define ω as the type of fish we observe (state of nature, class) where

$\omega = \omega 1$ for sea bass;

$\omega = \omega 2$ for salmon.

-$p(\omega 1)$ is the a priori probability that the next fish is a sea bass.

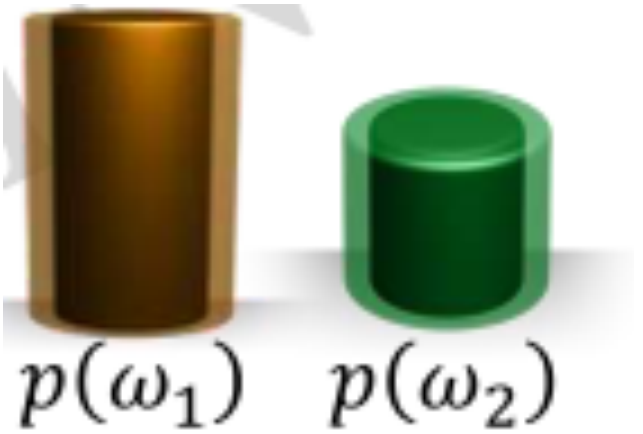
-$p(\omega 2)$ is the a priori probability that the next fish is a salmon.

Naïve Bayes

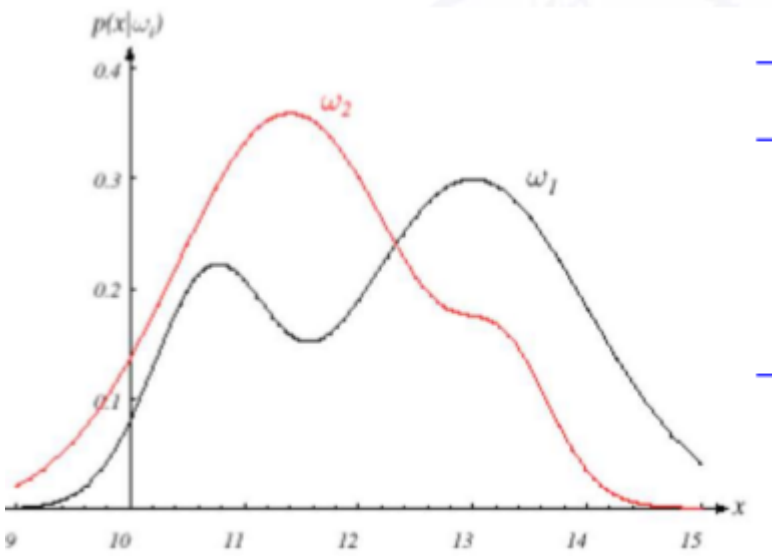
If we make decision with only the prior information

$$\text{Decision}\begin{cases}\omega 1 & p(\omega 1) > p(\omega 2) \\ \omega 2 & p(\omega 1) \le p(\omega 2)\end{cases}$$

p(error) = min{ p(ω1) , p(ω2)}



Naïve Bayes



x : continuous random variable. In this example, x is the lightness of fish.
- p(x|ωi) : the class-conditional probability density (probability of x given that the state of nature is ωi for i = 1,2)
- p(x|ω1) and p(x|ω2) : the difference in lightness between populations of sea bass and
salmon.

Naïve Bayes

How can we make a decision after observing the value of x?

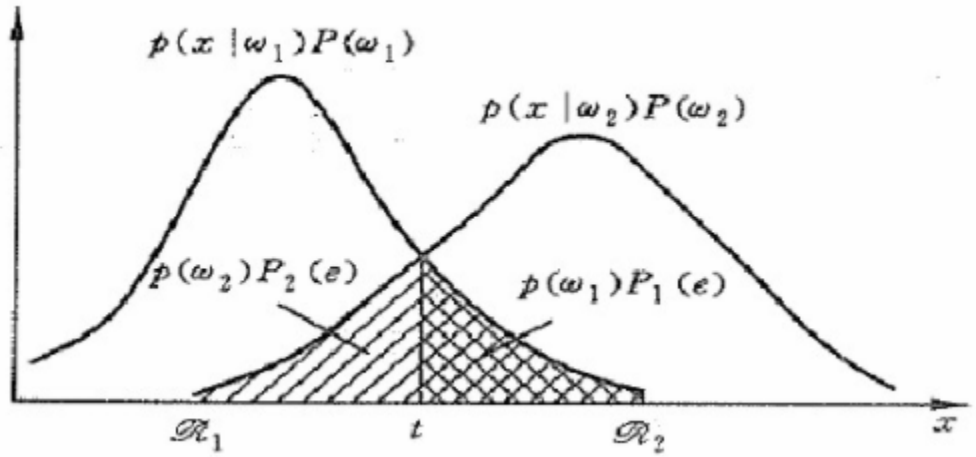
$$\text{Decision}\begin{cases}\omega 1 & p(\omega 1|x) > p(\omega 2|x)\\ \omega 2 & p(\omega 1|x) \le p(\omega 2|x)\end{cases}$$

$$\mathrm{p}(\omega i|x) = \frac{p(x|\omega i)\,p(\omega i)}{p(x)}$$

$$\text{Decision}\begin{cases}\omega 1 & p(\mathrm{x}|\omega 1)\,p(\omega 1) > p(x|\omega 2)\,p(\omega 2)\\ \omega 2 & p(x|\omega 1)\,p(\omega 1) \le p(x|\omega 2)\,p(\omega 2)\end{cases}$$

Naïve Bayes

$$\mathrm{p(error)} = \begin{cases} p(\omega 1) & decision = \omega 2 \\ p(\omega 2) & decision = \omega 2 \end{cases} = \int p(error)dx$$



Naïve Bayes

Exercise1:

Two types of cells：normal(ω1) and abnormal(ω2), the prior probabilities are known：

Normal(ω1)： $p(\omega1) = 0.9$

Abnormal(ω2)： $p(\omega2) = 0.1$

For the given cell x，we know its class-conditional probabilities:

$p(x|\omega1) = 0.2$, $p(x|\omega2) = 0.4$

This cell x is normal or abnormal?

Naïve Bayes

Exercise1:

p(ω1) = 0.9

p(ω2) = 0.1

p(x|ω1) = 0.2 , p(x|ω2) = 0.4

$$\text{Decision}\begin{cases}\omega 1 & p(\text{x}|\omega 1)p(\omega 1) > p(x|\omega 2)p(\omega 2)\\ \omega 2 & p(x|\omega 1)p(\omega 1) \le p(x|\omega 2)p(\omega 2)\end{cases}$$

$$\text{p}(x|\omega 1)p(\omega 1) = 0.9\times 0.2 = 0.18$$

$$\text{p}(x|\omega 2)p(\omega 2) = 0.1\times 0.4 = 0.04$$

Normal

Naïve Bayes

Given the title of an article, decide whether it is about sports or not

the training set is:

Text	Category
A great game	Sports
The election was over	Not Sports
Very clean match	Sports
A clean but forgettable game	Sports
It was a close election	Not sports

Is the <A very close game> about sports or not?

Naïve Bayes

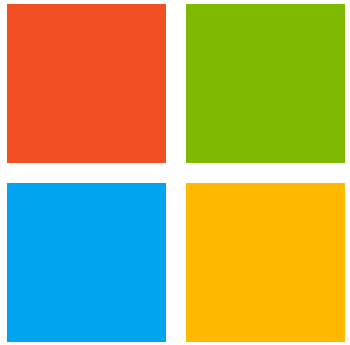
P(a very close game|Sports)
= P(a|Sports)×P(very|Sports)×P(close|Sports)×P(game|Sports)

P(a very close game|Not Sports)
= P(a|NS)×P(very|NS)×P(close|NS)×P(game|NS)

Word	P(word | Sports)	P(word | Not Sports)
a	$\frac{2+1}{11+14}$	$\frac{1+1}{9+14}$
very	$\frac{1+1}{11+14}$	$\frac{0+1}{9+14}$
close	$\frac{0+1}{11+14}$	$\frac{1+1}{9+14}$
game	$\frac{2+1}{11+14}$	$\frac{0+1}{9+14}$

Why Naive?

We assume that every word in a sentence has nothing to do with other words. This means that we do not look at the whole sentence, but the single word.



Microsoft

同济微软俱乐部
Tongji Microsoft Student Club

微软学生俱乐部

扫一扫二维码，加入该群。

关注TJMSC微信公众号