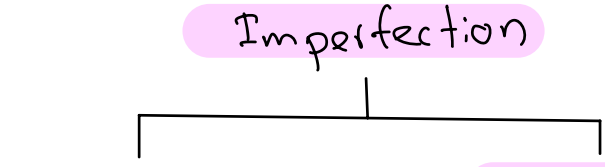
Bayesian Learning



ex: data is missing

Fuzzy logic — Probability theory

U - universe of discourse

A - set/event $\subset U$

x - random variable $\in A$

$[0,1]$ $P(A)$	$[0,1]$ $\mu_A(x)$
Probability that an ill known variable x ranging on U hits the well known set A.	Membership of well known variable x ranging on U hits the ill known set A.

before it happens → event happens → after it happended

* Will it rain tomorrow?

eg: rainy day - we know that tomorrow is raining except how high or low it is. (vague) (intensity)

* How is the intensity of rain

$P(A)$	$\mu_A(x)$
Measure theory	Set theory
Domain is 2^U	Domain is $[0,1]^U$ * not binary
Based on boolean algebra * binary	Can't be a boolean algebra

Bayes theorm

$$P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)}$$

- $P(h|D)$: posterior
- $P(D|h)$: class likelihood
- $P(h)$: prior (How much you know)
- $P(D)$: (probability of data) evidence

Example : Meningitis causes the patient to have stiff necks 50% of the time.

We know that $P(m) = 1/50000$

$P(s|m) = 0.5$ $\quad P(s) = 1/20$

$$P(m|s) = \frac{P(s/m) \cdot P(m)}{P(s)} = \frac{(0.5)\left(\frac{1}{50000}\right)}{1/20} = 0.0002$$

Basic rules

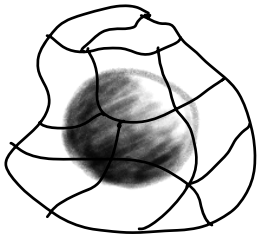
Product rule (conjunction) – $P(A \cap B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$

Sum rule (disjunction) – $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Total probability

$$P(B) = \sum_{i=1}^{n} P(B|A_i) \cdot P(A_i)$$

$A_1, A_2 \ldots A_n$ are mutually exclusive with $\sum_{i=1}^{n} P(A_i) = 1$



idea: intelligent is to find most probable hypothesis given the training data.

Let H be set of all solutions

Best hypothesis $h^* = \underset{h \in H}{\text{argmax}}\ P(h|D)$ D - evidence

$$h^* = \underset{h \in H}{\text{argmax}}\ \frac{P(D|h)\ P(h)}{P(D)}$$ ← constant

Maximum a posterion

$$h_{MAP} = \underset{h \in H}{\text{argmax}}\ P(D|h)\ P(h)$$

If all solutions have same probability we can calculate the maximum liklihood hypothesis.

$$h_{ML} = \underset{h_i \in H}{\text{argmax}}\ P(D|h_i)$$

when, $P(h_i) = P(h_j)$ for all i, j

Example for optimal classification:

Lets say we have three hypothesis,

$P(h_1|D) = 0.5$ — probability that 1st solution is correct given data D

$P(h_2|D) = 0.3$

$P(h_3|D) = 0.4$ → decision tree with 500 nodes } ex

$H = \{h_1, h_2, h_3\}$ $\quad |H| \ll 100$

new samples:

classification

$h_1(x) = \oplus$ yes

$h_2(x) = \ominus$ no

$h_3(x) = \ominus$ no

$V = \{\oplus, \ominus\}$

Bayes optimal classification

$$V_{best} = \underset{v_j \in V}{\operatorname{argmax}} \sum_{h_i \in H} P(v_j|h_i)\, P(h_i|D)$$

$P(h_i|D)$: which hypothesis is correct most of the time

$P(v_j|h_i)$: getting a certain value given that hypothesis

$P(\ominus|h_1) = 0$ $\quad P(\oplus|h_1) = 1$

$P(\ominus|h_2) = 1$ $\quad P(\oplus|h_2) = 0$

$P(\ominus|h_3) = 1$ $\quad P(\oplus|h_3) = 0$

$\sum_{h_i \in H} P(\oplus|h_i) \cdot P(h_i|D) = 0.5$

$\sum_{h_i \in H} P(\ominus|h_i) \cdot P(h_i|D) = 0.7$

} $x \in \{\ominus\}$ the best solution

But this is very expensive

Solution : Naive bayes classifier

Assumption : Attributes or features are conditionally independent. Hence,

$$P(a_1, a_2 \ldots, a_n \mid v_j) = \prod_i P(a_i \mid v_j)$$

given cancer, smoking

Variable x has features/attributes $a_1, a_2 \ldots a_n$.

Most probable value of f(x) :

$$V_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} \; P(v_j \mid a_1, a_2 \ldots a_n)$$

$$= \underset{v_j \in V}{\operatorname{argmax}} \; \frac{P(a_1, a_2 \ldots v_n \mid v_j)\, P(v_j)}{P(a_1, a_2, a_2 \ldots a_n)}$$

→ Probability of data

$$= \underset{v_j \in V}{\operatorname{argmax}} \; P(a_1, a_2 \ldots a_n \mid v_j) \cdot P(v_j)$$

cannot calculate

Navie bias classifier

$$V_{NB} = \underset{v_j \in V}{\operatorname{argmax}} \; P(v_j) \cdot \prod_{i=1}^{n} P(a_i \mid v_j)$$

But the conditional independence is often violated. Nonetheless, it works.

Using training data, estimate $P(v_j)$ and $P(a_i \mid v_j)$

Potential problem: $\hat{P}(a_i|v_j) = 0$

↳ No value is assigned to the target v_j

A better estimate:

$$\hat{P}(a_i|v_j) = \frac{n_c + mp}{n + m}$$

n_c = # of cases with $v = v_j$ and $a = a_i$

n = # of cases with $v = v_j$

m = weight assigned to prior

p = prior estimated based on data

Example: Car theft (by Eric Meisner)

Naive Bayes Classifier example

Eric Meisner

November 22, 2003

1 The Classifier

The Bayes Naive classifier selects the most likely classification V_{nb} given the attribute values $a_1, a_2, \ldots a_n$. This results in:

$$V_{nb} = \text{argmax}_{v_j \in V} P(v_j) \prod P(a_i|v_j) \quad (1)$$

We generally estimate $P(a_i|v_j)$ using m-estimates:

$$P(a_i|v_j) = \frac{n_c + mp}{n + m} \quad (2)$$

where:

n =	the number of training examples for which $v = v_j$
n_c =	number of examples for which $v = v_j$ and $a = a_i$
p =	a priori estimate for $P(a_i|v_j)$
m =	the equivalent sample size

2 Car theft Example

Attributes are Color , Type , Origin, and the subject, stolen can be either yes or no.

2.1 data set

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

2.2 Training example

We want to classify a Red Domestic SUV. Note there is no example of a Red Domestic SUV in our data set. Looking back at equation (2) we can see how to compute this. We need to calculate the probabilities

`P(Red|Yes), P(SUV|Yes), P(Domestic|Yes) ,` $P(a_i, v_j)$

`P(Red|No) , P(SUV|No), and P(Domestic|No)`

Will a "Red Domestic SUV" be stolen?

(This is not in the table)

P(Red|Y), P(SUV|Y), P(Domestic|Y)

P(Red|N), P(SUV|N), P(Domestic|N)

features are conditionally idependent

$n = 5$, $p = 0.5$, $m = 3$ (empirically)

$n_c = ?$

	Y	N
Red	3	2
SUV	1	3
Domestic	2	3

m - estimator

$$P(a_i|v_j) = \frac{n_c + mp}{n + m}$$

$$P(red|Y) = \frac{3 + 3\times.5}{5+3} = 0.56$$

$$P(red|N) = \frac{2 + 3\times0.5}{5+3} = 0.44$$

$$P(SUV|Y) = \frac{1 + 3\times0.5}{5+3} = 0.31$$

$$P(SUV|N) = \frac{3 + 3\times0.5}{5+3} = 0.56$$

$$P(dom|Y) = \frac{2 + 3\times0.5}{5+3} = 0.44$$

$$P(dom|N) = \frac{3 + 3\times0.5}{5+3} = 0.56$$

$$P(Y) = 0.5 \qquad P(N) = 0.5$$

$$V = Y \; :: = P(Y) \times P(red|y) \times P(suv|y) \times P(dom|y)$$

$$P(a_i|v_j) = \frac{n_c + mp}{n + m} \qquad = 0.5 \times 0.56 \times 0.31 \times 0.44 = 0.037$$

$$V = N \; : = P(N) \times P(red|N) * P(suv|N) * P(dom|N)$$

$$= 0.56 \times 0.44 \times 0.56 \times 0.56 = 0.069$$

Swarm Intelligence

- Population based stochastic optimization

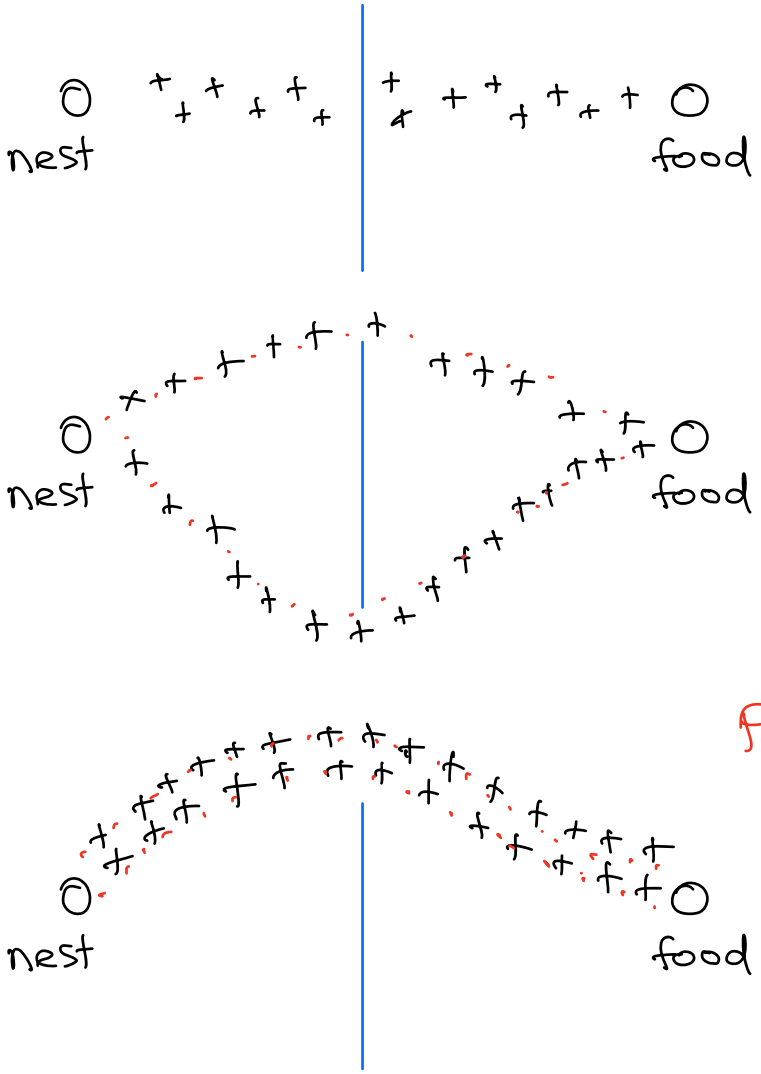
example: Ant colony optimization (ACO)

Particle swarm optimization (PSO)

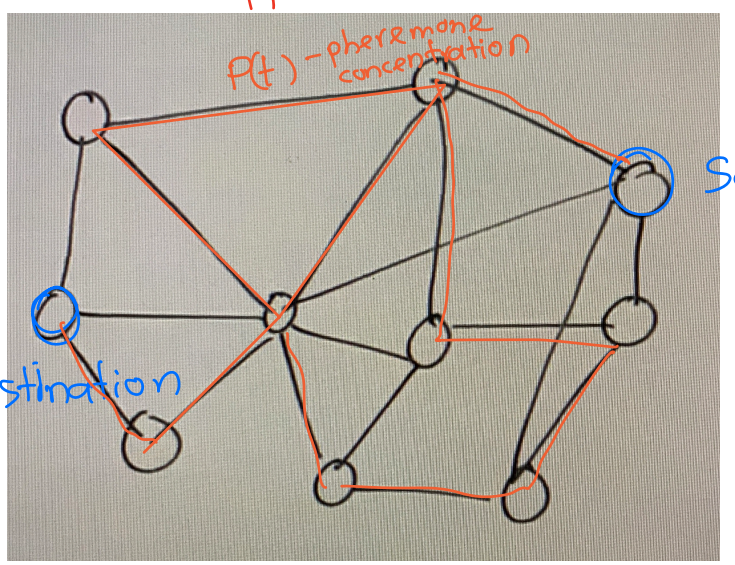
ACO - current ant population 10^{16}

- concept of stigmergy:

Indirect communication used by social insects to coordinate their activities



ACO applied on TSP



* artificial ants are agents moving from city to city on a TSP graph.
* cities connected via pheromone-rich edges are preferable.

Progress of Optimization

① Local trail update
② Global trail update

when all ants complete a tour, the ant that made the shortest tour modifies the edges of its tour by adding pheremone inversly propotional to the tour length.