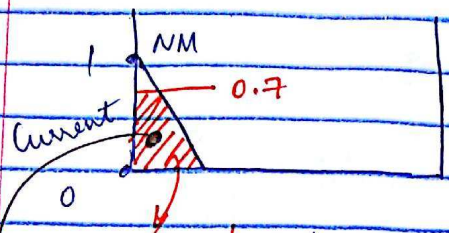
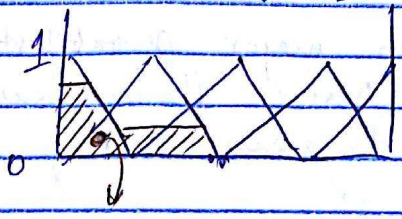


answer! but we need a number. ∴ take a centroid (represents a number)

Centre of area / CoG

Current = -6.3 mA

if we have two rules:



Current = -3.2 mA.

March 9th, 2017

Imperfection

Vague → Fuzzy Inference (Fuzzy Logic)

Uncertain → Bayesian inference (Probability Theory)

U = Universe of discourse

A = set (event) ⊂ U

x = variable ∈ A

P(A)

Probability that an ill-known variable x ranging on U hits the well-known set A.

$\mu_A(x)$

Membership of well-known variable x ranging on U hits the ill-known set A.

before it happens → event → after it happened

Eg. prob that next person entering room is wearing red.

Eg. intensity of red?

Probability based on Measure Theory

* Domain = 2^U
* Based on Boolean Algebra

Fuzzy comes from Set theory logic

* Domain = [0, 1]
* Cannot be Boolean Algebra.

Two major probabilistic methods
① Naive Bayes Learning
② Bayesian Belief Method.

Bayes Theorem
(What is the probability that my hypothesis (solution) is the right solution given the training data (D))

$$P(h/D) = \frac{P(D/h)\,P(h)}{P(D)}$$

$$\text{posterior} = \frac{\text{prior} * \text{class likelihood}}{\text{evidence}}$$

$h \in H$
all possible hypotheses/solutions

Bayes Rule Example

Meningitis can cause the patient to have a stiff neck. 50% of the time. Also, $P(\text{Meningitis}) = \frac{1}{50,000}$

$P(\text{Meningitis}/S) = P(M/S) = P(S/M) * P(M)/P(S)$
(S → stiff neck)

$P(S/M) = 0.5$
$P(M) = 1/50,000$
$P(S) = 1/20$
$P(M/S) = \frac{0.5 \times 1/50,000}{1/20} = 0.0002$

Choosing Hypotheses
Maximum a posteriori hypothesis

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

Basic Rules for Probabilities.

- Product rule $P(A \wedge B) = P(A/B) \cdot P(B)$
  $= P(B/A) \cdot P(A)$
- Sum rule $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Theorem of Total probabilities
  If events $A_1, A_2, \ldots A_n$ are mutually exclusive with $\sum_{i=1}^{n} P(A_i) = 1$, then $P(B) = \sum_{i=1}^{n} P(B/A_i) \, P(A_i)$

Brute Force MAP is very slow: (intractable)

$h_{MAP} = \underset{h \in H}{\operatorname{argmax}} \; P(h/D)$

If the size of the problem < 32 bits, then brute force works.

Minimum Description Length (MDL)

Use the length (h) + length (misclassifications).

$v_j \in V$ (set of classification cases) Eg. $\{\oplus, \ominus\}$

$\underset{v_j \in V, \, h_i \in H}{\operatorname{argmax}} \sum P(v_j/h_i) \, P(h_i/D)$

$P(h_1/D) = 0.4$	$P(\ominus/h_1) = 0$	$P(\oplus/h_1) = 1$
$P(h_2/D) = 0.3$	$P(\ominus/h_2) = 1$	$P(\oplus/h_2) = 0$
$P(h_3/D) = 0.3$	$P(\ominus/h_3) = 1$	$P(\oplus/h_3) = 0$

(Prob of "no" classification given h_1 solution)

(Optimal Bayes)

$\sum_{h_i \in H} P(\oplus|h_i)\, P(h_i|D) = 0.4$ $\qquad$ $\sum_{h_i \in H} P(\ominus|h_i)\, P(h_i|D) = 0.6$

The decision is $\ominus$

(Practical)

Practical Naive Bayes Classifier (NBC)

NBC is, along with Decision Trees and ANNs, among the most commonly used ML methods.

When to use? ① Moderate → large training data.

② Attributes are conditionally independant given classification.

("naively" assumes that conditions are independant while they aren't.)

Successful applications of NBC

① Diagnosis.

② Classify text document.

Assume the target function $f: X \to Y$, where each instance X is described by attributes $\langle a_1, a_2 \ldots a_n \rangle$. The most probable value of $f(x)$ is $V_{NB} = \underset{v_j \in V}{\text{argmax}}\ P(v_j \mid a_1, a_2 \ldots a_n)$.

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

Naive Bayes Assumption:

attributes are conditionally independant. This means:

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

This is a naive assumption

NBC $\quad V_{NB} = \underset{v_j \in V}{\text{argmax}}\ P(v_j) \prod P(a_i \mid v_j)$

Naive_Bayes_Learn (examples)

For each target value v_j

$\hat{P}(v_j) \leftarrow$ estimate $P(v_j)$

For each attribute value a_i of a

$\hat{P}(a_i|v_j) \leftarrow$ estimate $(P(a_i|v_j))$

Save this

Classify_New_Instance (x)

$$v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} \; \hat{P}(v_j) \prod \hat{P}(a_i \mid v_j)$$

|V| should be small.
(# of classification states)

Play Tennis

$P(y)\ P(sun/y)\ P(cool/y)\ P(high/y)\ P(strong/y) = 0.005$

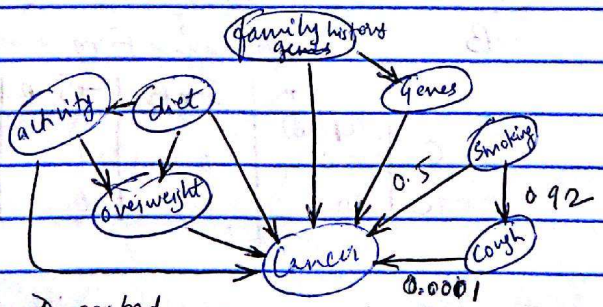
$P(n)\ P(sun/n)\ P(cool/n)\ P(high/n)\ P(strong/n) = 0.021$

↳ $V_{NB} = n.$

even if one probability is 0, then the whole probability = 0.
∴ use lots of data so that no probability is 0.
If it is 0, then we use an m-estimator to find some value to estimate a non zero value when you encounter it in N.B.

m → estimator $P(a_i \mid v_j) = \dfrac{n_c + mp}{n+m} = 0$?

March 14th, 2017



Directed Acyclic Graph (DAG)

Belief network
how to infer?

Bayesian Belief Network

- Naive Bayes is too restrictive
- The problem may prone to be intractable without such assumptions.
- Bayesian Belief Networks describe conditional independance among subsets of variables.
- allows combining the prior knowledge about (in)dependencies