Probabilities → Much of successful machine learning algorithms are based on probabilistic reasoning

Terms

Random variable → denotes something about which we are uncertain.
→ outcome of a random experiment

Ex. A = True if a randomly drawn person is female

P(A) = the fraction of possible worlds in which A is true.

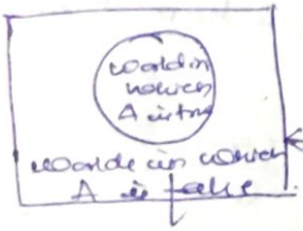
Random variable A is a function of Sample Space S (Set of all possible worlds) defined as $A : S \to \{0, 1\}$.

Sample Space → all set of possible worlds
Ex. set of students in our class

Random variable Gender : $S \to \{m, f\}$
Height : $S \to$ Reals

Event is a subset of S Ex. subset of S where G = f

visualizing A in Venn diagram.



← Sample Space of all possible worlds
Area is 1

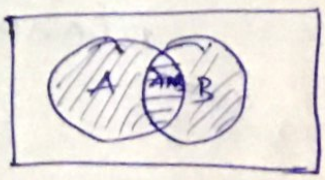
Axioms of probability

$0 <= P(A) <= 1$ Probability is between 0 to 1
false = 0
true = 1

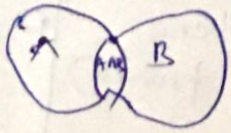
P(True) = 1

P(false) = 0

P(A or B) = P(A) + P(B) - P(A and B)



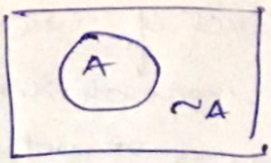
When gambling based on uncertainty formalism A, you can be exploited by an opponent



iff your uncertainty formalism A violates these axioms.

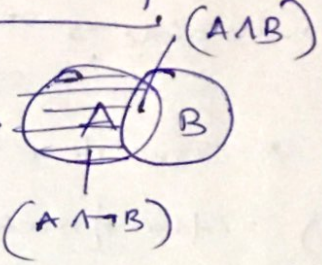
Elementary probability

$P(\sim A) + P(A) = 1$



marginal joint

$P(A) = P(A \wedge B) + P(A \wedge \sim B)$



If A & B are boolean, applying boolean logic

$A = [A \text{ and } (B \text{ or } \sim B)]$

$= [(A \text{ and } B) \text{ or } (A \text{ and } \sim B)]$

Can be rearranged as

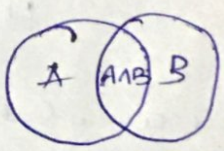
$P(A) = P(A \text{ and } B) + P(A \text{ and } \sim B) - P((A \text{ and } B) \text{ and } (A \text{ and } \sim B))$

$P(A) = P(A \text{ and } B) + P(A \text{ and } \sim B) - P(A \text{ and } B \text{ and } A \text{ and } \sim B)$

Conditional probability

$P(A/B)$ → what is the probability that A is true given B is true

$$P(A/B) = \frac{P(A \wedge B)}{P(B)}$$



Ex. what is the probability that you have blue eyes given that you are female.

Let say A is blue eyes

B is female.

what is $P(\text{blue eyes}/\text{female}) = \frac{b \wedge f}{b}$

Corollary: The chain rule.

$P(A \wedge B) = P(A/B)\,P(B) = P(B/A)\,P(A)$ Given $(A \wedge B) = (B \wedge A)$ — ①

$P(A \wedge B \wedge C) = P(A/BC)\,P(BC)$

$P(A/Bc)\,P(B/C)\,P(C)$

Hence, from ①

$$P(A/B) = \frac{P(B/A) \cdot P(A)}{P(B)}$$

Bayes' rule

we call

$P(A)$ → Prior

$P(A/B)$ → posterior

Otherforms

$$P(A/B) = \frac{P(B/A) \cdot P(A)}{P(B/A)P(A) + P(B/\sim A)\,P(\sim A)}$$ // P(B) rewritten.

$$P(A/B \wedge X) = \frac{P(B/A \wedge X)\,P(A \wedge X)}{P(B \wedge X)}$$

X could be anything.

Ex. A = you have flu
B = you just coughed.

Assume $P(A) = 0.05$
$P(B/A) = 0.8$
$P(B/\sim A) = 0.2$

What is $P(flu/cough) = \frac{P(cough/flu) \cdot P(flu)}{P(cough/flu).P(flu) + P(cough/\sim flu)\,P(\sim flu)}$

$$= \frac{0.8 \times 0.05}{0.8 \times 0.05 + 0.2 \times 0.95} = 0.17$$

⇒ Posterior Probability

We are interested in function approximation, coming up with training algorithms that can learn functions $X \rightarrow Y$ (learning)

Email Spam. X to Y

$$f : X \rightarrow Y$$

Equivalent thing is $P(Y/X)$

$P(Spam = yes / Email)$

What is the probability distribution of probability value.

The single most important thing to learn is joint probability distribution → Assignment of probability to a collection of random variables.

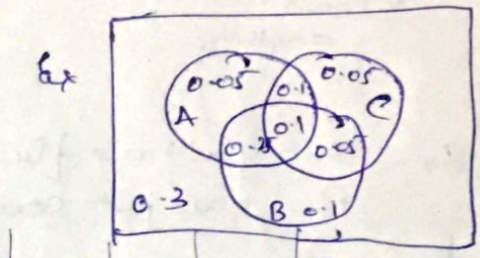
Joint distribution recipe with M variables.

(1) Make a truth table listing all combination of values
M boolean variable → 2^M rows

(2) For each combination of values, say how probable it is

(3) Probabilities must sum to 1

Ex



A	B	C	Prob
0	0	0	0.3
0	0	1	0.05
0	1	0	0.1
0	1	1	0.05
1	0	0	0.05
1	0	1	0.1
1	1	0	0.25
1	1	1	0.1

If you know joint distribution, then there is nothing that you cannot answer on probabilities, joint, conditional or any subset

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Joint probability
$P(A\,B) = P(A=1, B=1)$ (irrespective of C)
$= 0.25 + 0.1$

$P(A) = 0.05 + 0.1 + 0.25 + 0.1$

$$P(A/B) = \frac{\sum_{A \text{ and } B} P(\text{row})}{\sum_{B} P(\text{row})} = \text{Conditional probability}$$

$$= \frac{0.25 + 0.1}{0.25 + 0.1 + 0.1 + 0.05}$$

$P(\sim A/B) = 1 - P(A/B)$ for given value being same. It will sum to 1

Joint probability sounds like a solution to learning $f: X \to Y$ or $P(Y/X)$

But problems are
- There are simple boolean, but real world (2^M is huge)
- Computational values
- Also it is not easy to learn the probability distribution.

Ex.

Joint dist with 100 attributes / Reasonable in real world

rows = 2^{100} ($\approx 1000^{10}$ or 10^{30})

people on earth ($\approx 10^9$)

There are more rows but the table will have fraction of rows with

- example (≈ 0.999)

Data Sparsity is a problem.

What to do?

1. Be Smart about how we estimate probabilities from sparse data
   - → Maximum likelihood estimates
   - → Maximum a posteriori estimates
2. Be Smart about how to represent joint distribution
   - → Bayes network, Graphical models.

Being Smart about estimating probabilities

Ex. Coin flip H = 1 T = 0

α_1 times α_0 times

$P(X=1)$?

$P(X=0)$?

Algorithm for estimating the probability by flipping repeatedly

$$P(X=1) = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$

Estimating $\theta = P(X=1)$

Test A : 100 flips 51 heads (α_1) 49 tails (α_2) → lot of data.

$$\hat{P}(X=1) = \frac{\alpha_1}{\alpha_1 + \alpha_2} = \frac{51}{100} = 0.51$$

Test B : 3 flips 2 heads 1 tail → This is problem with simple universe, the true probability might be skewed.

$$\hat{P}(X=1) = \frac{2}{3} = 0.67$$

Indicative of problem with MLE with less data.

Case c: (Online learning)

Keep flipping, want single learning algorithm that gives reasonable estimate after each flip

$\theta = P(x=1)$

$\alpha_1 = \#$ of observed heads $(x=1)$

$\alpha_0 = \#$ of observed tails $(x=0)$

Given that there is also some prior → All got-minted coins have equal probability

we can add imagnary hallucinated flips of 10 H & 10 T → which could reflect the prior.

$$\frac{\alpha_1 + 10}{(\alpha_1 + 10) + (\alpha_0 + 10)}$$

At 0 coin flips, the ratio is 0.5

At 1 " , $\frac{10}{21}$

At million, 10 is negligible & as $\alpha \to \infty$ the ratio converges to $\frac{\alpha_1}{\alpha_1 + \alpha_0}$.

But why 10, instead of 100, it would take more data to weigh in

General algorithm, $\beta_1 = \#$ hallucinated $x=1$s

$\beta_0 = \#$ hallucinated $x=0$s

The formula could be replaced as

$$\frac{\alpha_1 + \beta_1}{(\alpha_1 + \beta_1) + (\alpha_0 + \beta_0)}$$ (observed + hallucinated)

more confidence on prior, higher value of β

As the flips increase β becomes negligible → In case of sparse data, we could use the prior.

Principles for estimating probabilities

Principle 1 → Maximum likelihood

Choose parameters θ that maximize $P(\text{data}/\theta)$

Ex. $\hat{\theta}_{MLE} = \frac{\alpha_1}{\alpha_1 + \alpha_0}$

Principle 2 → Maximum a posteriori probability $= \frac{P(\text{data}/\theta) \cdot P(\theta)}{P(\text{data})}$

Choose parameters θ that maximize $P(\theta/\text{data})$ (posterior)

Ex. $\theta^{MAP} = \frac{\alpha_1 + \#\text{ hallucinated\_1s}}{(\alpha_1 + \#\text{hallucinated\_1s}) + (\alpha_0 + \#\text{ hallucinated\_0s})}$

TO maximize $P(\theta/\text{data})$, $P(\text{data})$ has nuisance effect

hence maximise the numerator.

Maximum Likelihood Estimation

$P(X=1) = \theta \quad P(X=0) = (1-\theta)$

Here say the coin flips result in $\{1, 0, 0, 1\}$ (1 → H, 0 → T)

$S = \{\theta, 1-\theta, (1-\theta), \theta\}$

$P(\text{data}/\theta) = \prod S$

$= \theta \cdot (1-\theta)(1-\theta) \cdot \theta \cdot \theta$

$= \theta^3 (1-\theta)^2$

Generalising for observed heads & tails.

$P(D/\theta) = P(\alpha_1, \alpha_0/\theta) = \theta^{\alpha_1}(1-\theta)^{\alpha_0}$

Hence maximising $\theta, \hat{\theta}$ would be in maximising the value of $P(D/\theta)$

Maximise the function or its log is going to be the same since log is monotonic.

$$\hat{\theta} = \arg\max_{\theta} \ln\left(P(D/\theta)\right)$$ ← data likelihood.

$$= \arg\max_{\theta} \ln \theta^{\alpha_1} (1-\theta)^{\alpha_0}$$

Set derivative to zero $\frac{d}{d\theta} \ln P(D/\theta) = 0$ $\quad \frac{\partial \ln \theta}{\partial \theta} = \frac{1}{\theta}$

$$\hat{\theta} = \arg\max_{\theta} \ln\left[\theta^{\alpha_1}(1-\theta)^{\alpha_0}\right]$$

$$= \frac{\partial}{\partial \theta} \alpha_1 \ln\theta + \alpha_0 \ln(1-\theta)$$

$$\alpha_1 \frac{1}{\theta} + \alpha_0 \frac{\partial \ln(1-\theta)}{\partial \theta}$$ ← use chain rule

$$\frac{\partial \ln(1-\theta)}{\partial(1-\theta)} \cdot \frac{\partial(1-\theta)}{\partial \theta}$$

$$= \frac{1}{1-\theta} \cdot -1$$

$$0 = \alpha_1 \frac{1}{\theta} - \frac{\alpha_0}{1-\theta}$$

$$\theta = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$

For a coin flip which has a bernoulli distribution (a number)

$P(x=1) = \theta$

$P(x=0) = 1-\theta$

Each flip yields boolean value for x

$x \sim \text{Bernulli } P(x) = \theta^x (1-\theta)^{(1-x)}$

HTTHH

$P(HTTHH) = \theta \cdot (1-\theta)(1-\theta)(\theta)(\theta)$

$\theta^x (1-\theta)^{x-n}$

Dataset D of independent identically distributed (i.i.d) flips
Produces α_1 ones α_0 zeroes (Binomial)

$$P(D/\theta) = P(\alpha_1, \alpha_0/\theta) = \theta^{\alpha_1}(1-\theta)^{\alpha_0}$$

$$\hat{\theta}^{MLE} = \text{argmax}_\theta P(D/\theta) = \frac{\alpha_1}{\alpha_1+\alpha_0}$$

Principle :-> Maximum a posteriori probability

Beta prior distribution - $P(\theta)$

$$P(\theta) = \frac{\theta^{B_H-1}(1-\theta)^{B_T-1}}{B(B_H, B_T)} \sim \text{Beta}(B_H, B_T)$$

Likelihood function : $P(D/\theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$

Posterior : $P(\theta/D) \propto P(D/\theta)\,P(\theta)$

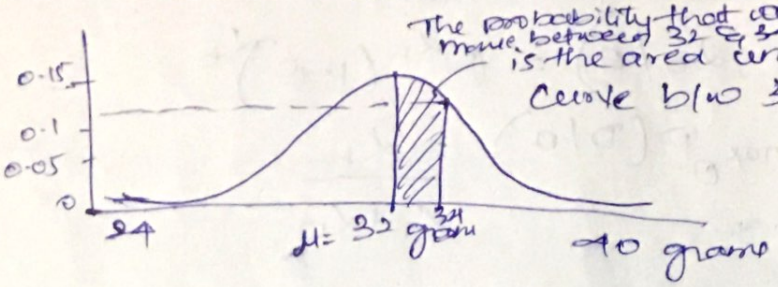
$$\propto \theta^{\alpha_H+B_H-1}(1-\theta)^{\alpha_T+B_T-1}$$

$$\hat{\theta}^{MAP} = \frac{(\alpha_H+B_H-1)}{(\alpha_H+B_H-1)+(\alpha_T+B_T-1)}$$

Terms:
Likelihood function : $P(\text{data}/\theta)$
prior : $P(\theta)$
Posterior $P(\theta/\text{data})$
Conjugate prior : $P(\theta)$ is the conjugate prior for likelihood function $P(\text{data}/\theta)$ if the forms of $P(\theta)$ & $P(\theta/\text{data})$ are the same.

Probability vs Likelihood

Let's take a normal distribution plot of mouse weight



Here if the area under curve = 0.29

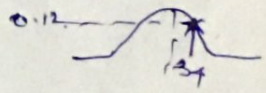
⇒ there's a 29% chance a randomly selected mouse will weigh between 32 & 34 grams.

Mathematically.

$$P\left(\text{wt } 32\text{-}34 \;\middle|\; \underset{\text{mean}}{\mu} = 32 \;\&\; \underset{\text{Standard deviation}}{\sigma} = 2.5\right) = 0.29$$

Lets say we measured a mouse with weight = 34 grams

The likelihood of weighing a 34 grams mouse is



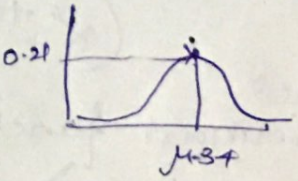
Mathematically

$$L(\text{mean} = 32 \;\&\; \text{Standard deviation} = 2.5 \mid \text{mouse weighs 34 grams})$$

$$= 0.12$$

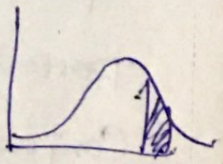
↳ this is fixed in likelihood.

If mean is shifted to 34 grams



with right being fixed we modify the shape & location of the distribution with the left side

P(data / distribution) : area under a fixed distribution



Likelihood are the y axis values for fixed data points with distributions that can be moved.

L (distribution / data)

Bayes Rule

$$P(Y/X) = \frac{P(X/Y)\,P(Y)}{P(X)}$$

which is shorthand for

$$(\forall i,j)\ P(Y=y_i/X=x_j) = \frac{P(X=x_j/Y=y_i)\,P(Y=y_i)}{P(X=x_j)}$$

Conditional Independence

$$(\forall i,j,k)\ P(X=x_i/Y=y_j, Z=z_k) = P(X=x_i/Z=z_k)$$

X is conditionally independent of Y given Z if the probability distribution governing X is independent of the value of Y, given the value of Z.

$$P(X/Y,Z) = P(X/Z)$$

Ex. $P(\text{Thunder}/\text{Rain}, \text{Lightning}) = P(\text{Thunder}/\text{Lightning})$

Naive Bayes uses assumption that the X_i are conditionally independent, given Y Ex $P(X_1/X_2\,Y) = P(X_1/Y)$

Given this assumption then

$$P(X_1, X_2/Y) = P(X_1/X_2\,Y)\,P(X_2/Y)$$
$$P(X_1/Y)\,P(X_2/Y)$$

$$P(X_1 \ldots X_n/Y) = P(X_1/Y) \ldots P(X_n/Y)$$
$$\prod_i P(X_i/Y)$$

without conditional independence assumption $2\cdot(2^n-1)+1$
with conditional $2n+1$

Naive Bayes classifier

Supervised learning Algorithm used for classification
Based on Bayes' theorem, $P(A/B) = \frac{P(B/A) \cdot P(A)}{P(B)}$

P(A) = Prior probability (Probability of event before event B)
P(A/B) = Posterior probability (probability of event after event B is true)

Aim & Dataset : [Document]

Text	Category	
"A great game"	Sports	# words in Sports "
"The election was over"	Non-Sports	
"Very Clean Match"	Sports	# words in non sports "
"A clean but forgettable game"	Sports	# unique words "
"It was a close election"	Non Sports	

Classify >> "A very close game" Category? Sports / NonSports

Concept

P(Sports / "A very close game") : ?
P(Non-Sports / "A very close game") } whoever wins is the answer

How to find probability

P("A very close game") = P(A) × P(very) × P(close) × P(game)

Then $P(\text{"A very close game"}/\text{Sports}) = P(A/\text{Sports}) \times P(\text{very}/\text{Sports}) \times P(\text{close}/\text{Sports}) \times P(\text{game}/\text{Sports}) \times P(\text{Sports})$

$P(\text{"A very close game"}/\text{Non-Sports}) = P(A/\text{Non-Sports}) \times P(\text{very}/\text{Non Sports}) \times P(\text{Non sports}) \ldots$

Calculating probabilities $P(Sports) = 3/5$ $P(Non\text{-}sports) = 2/5$

$P(A/Sports) = 2/11$ (2 times in Sports), $P(very/Sports) = 1/11$, $P(close/Sports) = ? \Rightarrow 0/11$

and hence $P(\text{"A very close game"}/Sports) = 0 \Rightarrow$ this is incorrect

Solution

Laplace Smoothing

$$\theta_j = \frac{x_i + \alpha}{N + \alpha d}$$

$\alpha > 0$, mainly $\alpha = 1$ (always)

$d \Rightarrow$ no. of unique words or distinct words.

In other words

$$P(word) = \frac{\text{word count} + 1}{\text{Total no. of words} + \text{No. of unique words}}.$$

with laplace smoothing $P(close/Sports) = \frac{0+1}{11+14}$

word	$P(word/Sports)$	$P(word/Non\text{-}Sports)$
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	$\frac{2+1}{11+14}$	$\frac{0+1}{9+14}$

$P(a|\text{sports}) \times P(\text{very}|\text{sports}) \times P(\text{close}|\text{sports}) \times P(\text{game}|\text{sports}) \times P(\text{sports})$ (0.6)

$= 2.76 \times 10^{-5} = 0.0000276$

$P(a|\text{non-sports}) \times P(\text{very}|\text{non-sports}) \times P(\text{close}|\text{non-sports}) \times P(\text{game}|\text{non-sports})$

$\times P(\text{Not Sports})$

$= \frac{2}{23} \times \frac{1}{23} \times \frac{2}{23} \times \frac{1}{23} \times 0.4$

$= 0.572 \times 10^{-5}$

$P(\text{Sentence}|\text{Sports}) > P(\text{Sentence}|\text{non sports})$

$\Rightarrow$ Sports Category

Naive Bayes in a nutshell

$$P(Y=y_k|X_1 \dots X_n) = \frac{P(Y=y_k)\, P(X_1 \dots X_n | Y=y_k)}{\sum_j P(Y=y_j)\, P(X_1 \dots X_n | Y=y_j)}$$

Assuming conditional independence among X_i's

$$P(Y=y_k|X_1 \dots X_n) = \frac{P(Y=y_k) \prod_j P(X_i|Y=y_k)}{\sum_j P(Y=y_j) \prod P(X_i|Y=y_j)}$$

$\Rightarrow$ denominator independent of y_k

For a new data

$$y^{new} \leftarrow \underset{y_k}{\text{argmax}}\ P(Y=y_k) \prod_i P(X_i^{new}|Y=y_k)$$

Classifying Text documents

Turn words in document into bag of words.

Y discrete valued Eg Spam or not

$X = \langle X_1, X_2, \dots X_n \rangle$ = document

X_i is boolean [1 word in document, 0 otherwise]