BAYESIAN LEARNING

CAN BE SUMMARIZED BY THE FOLLOWING:

$$P(x|D) = \sum_{h \in H} P(x|D,h)\,P(h|D) = \sum_{h \in H} P(x|h) \cdot P(h|D)$$

- $P(x|D)$: PREDICTIONS ON NEW DATA x (GIVEN WHAT WE LEARNED FROM THE DATASET D)
- $P(x|h)$: PROBABILITY OF OBSERVING x UNDER THE HYPOTHESIS h
- $P(h|D)$: PROBABILITY OF HYPOTHESIS h ACCORDING TO OUR DATA
  - IT IS CALLED BAYESIAN BECAUSE THIS TERM OBTAINED APPLYING BAYES THEOREM

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

- $P(h|D)$: POSTERIOR: PROBABILITY OF HYPOTHESIS h AFTER OBSERVING THE DATA
- $P(D|h)$: LIKELIHOOD: PROB OF OBSERVING THE DATASET UNDER THE HYPOTHESIS h
- $P(h)$: PRIOR: PROBABILITY OF HYPOTHESIS h BEFORE OBSERVING THE DATA SET

WITH RESPECT TO THE MAXIMUM LIKELIHOOD HERE WE ARE USING ALL THE HYPOTHESES IN THE HYPOTHESES SPACE H.

WE CAN APPROXIMATE BAYESIAN INFERENCE USING A SINGLE HYPOTHESIS TO MAKE PREDICTIONS

- **MAXIMUM A POSTERIORI**: APPROXIMATE $P(x|D)$ WITH $P(x|h_{MAP})$

  WHERE

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

- **MAXIMUM LIKELIHOOD**: IF WE FURTHER ASSUME NO PRIOR AMONG THE HYPOTHESES THEN WE GET THE FOLLOWING APPROXIMATION

  $$P(x|D) \simeq P(x|h_{ML}) \quad \text{WHERE} \quad h_{ML} = \underset{h \in H}{\operatorname{argmax}}\, P(D|h)$$

EXAMPLE OF BAYESIAN INFERENCE

ASSUME I KNOW AN INDUSTRY PRODUCES 5 TYPES OF KANDY BOX (HYPOTHESES) WHICH COMPOSITION IS KNOWN AND THEY SPECIFY ALSO THE NUMBER OF BOXES OF EACH TYPE THEY PRODUCE (PRIOR PROBABILITY DISTRIBUTION OVER MY HYPOTHESIS SPACE).

TYPES OF CANDY BOXES (HYPOTHESES)	PRIOR PROBABILITY OVER THE HYPOTHESIS SPACE
h_1 100% cherry	$P(h_1) = 0.1$
h_2 75% cherry 25% lime	$P(h_2) = 0.2$
h_3 50% cherry 50% lime	$P(h_3) = 0.4$
h_4 25% cherry 75% lime	$P(h_4) = 0.2$
h_5 100% lime	$P(h_5) = 0.1$

NOW I PICK 3 CANDIES FROM A BOX (EACH TIME I REINSERT THE CANDY SUCH THAT THE PROBABILITY DOES NOT CHANGE) AND I WANT TO USE BAYESIAN INFERENCE TO PREDICT THE NEXT CANDY GIVEN THE OBSERVED DATA. ASSUME I PICK 3 LIME CANDIES $D = \{l,l,l\}$. FIRST I NEED TO COMPUTE THE POSTERIOR DISTRIBUTION OF MY HYPOTHESES AND THEN I CAN USE THE POSTERIORS TO MAKE NEW PREDICTIONS

$$P(h_1 \mid D=\{l,l,l\}) = \alpha \cdot P(D=\{l,l,l\} \mid h_1) \cdot P(h_1) = \alpha \cdot 0^3 \cdot 0.1 = 0$$

$$P(h_2 \mid D=\{l,l,l\}) = \alpha \cdot P(D=\{l,l,l\} \mid h_2) \cdot P(h_2) = \alpha \cdot 0.25^3 \cdot 0.2 = \alpha \cdot 0.003125 = 0.01$$

$$P(h_3 \mid D=\{l,l,l\}) = \alpha \cdot P(D=\{l,l,l\} \mid h_3) \cdot P(h_3) = \alpha \cdot 0.5^3 \cdot 0.4 = \alpha \cdot 0.05 = 0.21$$

$$P(h_4 \mid D=\{l,l,l\}) = \alpha \cdot P(D=\{l,l,l\} \mid h_3) \cdot P(h_4) = \alpha \cdot 0.75^3 \cdot 0.2 = \alpha \cdot 0.084375 = 0.36$$

$$P(h_5 \mid D=\{l,l,l\}) = \alpha \cdot P(D=\{l,l,l\} \mid h_5) \cdot P(h_4) = \alpha \cdot 1.0^3 \cdot 0.1 = \alpha \cdot 0.1 = 0.42$$

$$\alpha(0.003125 + 0.05 + 0.084375 + 0.1) = 1$$

$$\alpha = \frac{1}{0.2375} = 4.210526$$

I DO NOT COMPUTE THE DENOMINATOR AND NORMALIZE SUCH THAT THE POSTERIORS SUM TO ONE

⇓ BAYESIAN INFERENCE

$$P(l \mid D=\{l,l,l\}) = \sum_i P(l \mid h_i) \cdot P(h_i \mid D=\{l,l,l\})$$

$$= 0 \times 0 + 0.25 \times 0.01 + 0.5 \times 0.21 + 0.75 \times 0.36 + 1 \times 0.42 = 0.80$$

APPLICATIONS OF BAYESIAN INFERENCE

BAYES OPTIMAL CLASSIFIER

ASSUME I WANT TO CLASSIFY A NEW OBSERVATION x WITH A CLASS $c_s \in C$ THEN

$$c_{opt} = \underset{c_s \in C}{\operatorname{argmax}} P(c_s|x) = \underset{c_s \in C}{\operatorname{argmax}} \sum_{h \in H} P(c_s|h) \cdot P(h|x)$$

WITH THE SAME HYPOTHESIS SPACE AND PRIOR KNOWLEDGE NO OTHER CLASSIFICATION METHOD OUTPERFORMS BAYES OPTIMAL CLASSIFIER

HOWEVER IT IS OFTEN IMPRACTICAL BECAUSE

- I'VE NO ACCESS TO THE ENTIRE HYPOTHESIS SPACE OR IT IS INFINITE, SO THAT THE SUMMATION BECOMES UNTRACTABLE
- DIFFICULT TO MODEL $P(h|d)$

A SIMPLE APPROXIMATION THAT IS OFTEN USED IN PRACTICE IS

NAIVE BAYES CLASSIFIER

ASSUME A POINT x IS REPRESENTED BY A SET OF ATTRIBUTES (OR FEATURES) THAT ARE MUTUALLY INDEPENDENT GIVEN THE CATEGORY LABEL c

$x = \langle a_1 \dots, a_d \rangle$, AND $C = \{c_1 \dots, c_e\}$

$$c_{NB} = \underset{c_s \in C}{\operatorname{argmax}} P(c_s|a_1 \dots, a_d) = \underset{c_s \in C}{\operatorname{argmax}} P(a_1 \dots, a_d | c_s) \cdot P(c_s) = \underset{c_s \in C}{\operatorname{argmax}} P(c_s) \cdot \prod_{i=1}^{d} P(a_i|c_s)$$

FOR THE INDEPENDENCE ASSUMPTION

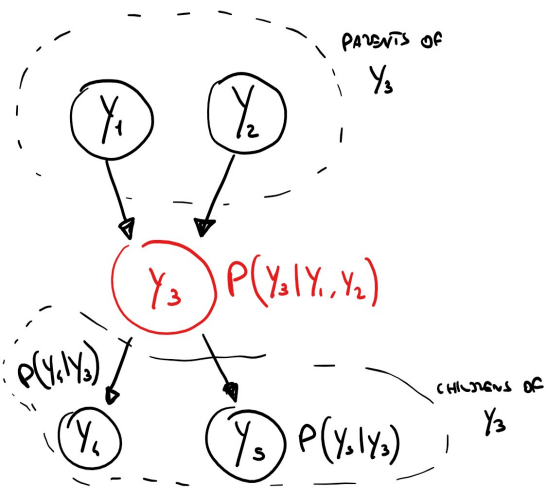
WHERE

$$P(a_i = \ell | c_s = h) = \frac{\left(\begin{array}{l}\text{NUMBER OF EXAMPLES WITH} \\ a_i = \ell \text{ THAT BELONG TO} \\ \text{CLASS } h\end{array}\right) + d}{\left(\begin{array}{l}\text{NUMBER OF EXAMPLES WITH} \\ a_i = \ell\end{array}\right) + kd}$$

THIS IS A SMOOTHING TERM, BECAUSE IF ONE TERM HERE IS ZERO THE WHOLE PROBABILITY BECOMES ZERO

BAYESIAN NETWORKS (OR BAYES NETWORKS, BAYES NETS, BELIEF NETWORKS, DECISION NETWORK)

DESCRIBES CONDITIONAL INDEPENDENCE BETWEEN SUBSETS OF RANDOM VARIABLES USING A DIRECT ACYCLIC GRAPH



ASSUMPTION: EVERY VARIABLE IS CONDITIONALLY INDEPENDENT WITH RESPECT TO ITS NON DESCENDANT, GIVEN ITS PARENTS

E.G.

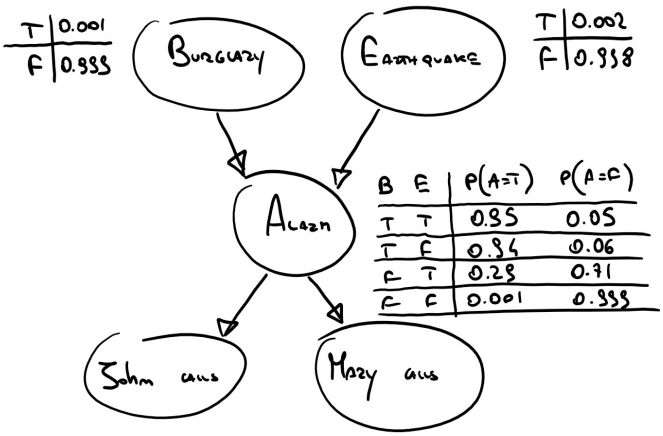
Y_5 IS INDEPENDENT W.R.T. Y_4 GIVEN Y_3

$$P(Y_5, Y_4 | Y_3) = P(Y_5|Y_3) \cdot P(Y_4|Y_3)$$

INFERENCE: GIVEN THE POSTERIOR PROBABILITY OF SOME VARIABLES WHEN OTHER VARIABLES ARE OBSERVED

$$P(X_1 = x_1, X_2 = x_2, \dots) = \prod_{i=1}^{m} P(X_i = x_i | \text{assignments to Parents}(x_i))$$

EXAMPLE THE CONDITIONAL PROBABILITIES AT EACH NODE ARE SPECIFIED THROUGH CONDITIONAL PROBABILITY TABLES (CPT)



T	0.001
F	0.999

T	0.002
F	0.998

B	E	P(A=T)	P(A=F)
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

A	P(J=T)	P(J=F)
T	0.90	0.1
F	0.05	0.95

A	P(M=T)	P(M=F)
T	0.7	0.3
F	0.4	0.6

INFERENCE WITH NO OBSERVED VARIABLES USING VARIABLE ELIMINATION

$$P(J) = \sum_A \sum_B \sum_E P(B) \cdot P(E) \cdot P(A|B,E) \cdot P(J|A)$$

$$= \sum_A \sum_B P(B) \cdot P(J|A) \left(\sum_E P(A|B,E) \cdot P(E) \right)$$

$$\lambda_1(A,B) = P(A|B,E=T) \cdot P(E=T) + P(A|B,E=F) \cdot P(E=F)$$

A	B	$\lambda_1(A,B)$
T	T	0,940
T	F	0,002
F	T	0,060
F	F	0,998

$$= \sum_A \sum_B P(B) \cdot P(J|A) \lambda_1(A,B)$$

$$= \sum_A P(J|A) \left(\sum_B \lambda_1(A,B) \cdot P(B) \right)$$

$$\lambda_2(A) = \lambda_1(A,B=T) \cdot P(B=T) + \lambda_1(A,B=F) \cdot P(B=F)$$

A	$\lambda_2(A)$
T	0,003
F	0,997

$$= \sum_A P(J|A) \cdot \lambda_2(A) = P(J|A=T) \cdot \lambda_2(A=T) + P(J|A=F) \cdot \lambda_2(A=F)$$

J	P(J)
T	0,053
F	0,947

INFERENCE WITH EVIDENCES USING VARIABLE ELIMINATION

$$P(M \mid B=\top) = \sum_A \sum_E P(B=\top) \cdot P(E) \cdot P(A \mid B=\top, E) \cdot P(M \mid A)$$

$$= \frac{1}{P(B=\top)} \sum_A 0.001 \cdot P(M \mid A) \left(\sum_E P(A \mid B=\top, E) \cdot P(E) \right)$$

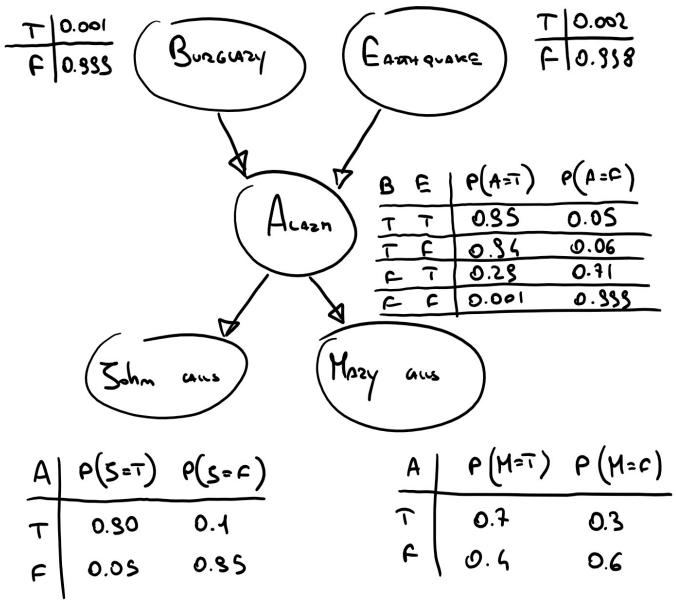
$$\lambda_1(A) = P(A \mid B=\top, E=\top) \cdot P(E=\top) + P(A \mid B=\top, E=F) \cdot P(E=F)$$

A	$\lambda_1(A)$
T	0,940
F	0,060

$$= \frac{1}{0.001} \sum_A 0.001 \; P(M \mid A) \cdot \lambda_1(A) =$$

$$= \frac{1}{0.\cancel{001}} \left(\cancel{0,001} \left(P(M \mid A=\top) \lambda_1(A=\top) + P(M \mid A=F) \lambda_1(A=F) \right) \right)$$

M	$P(M \mid B=\top)$
T	0,682
F	0,318



A	P(S=T)	P(S=F)
T	0.90	0.1
F	0.05	0.95

A	P(M=T)	P(M=F)
T	0.7	0.3
F	0.4	0.6