University of Nottingham
UK | CHINA | MALAYSIA

COMP3055 Machine Learning

Topic 7 – Bayesian Learning

Dr. Zheng LU

2018 Autumn

Probability

- The world is a very uncertain place.
- 30 years of Artificial Intelligence research danced around this fact.
- And then a few AI researchers decided to use some ideas from the eighteenth century.

Random Variables

- A variable whose possible values are outcomes of a random phenomenon.
- Denotes a quantity (event) that is uncertain.
- Maybe result of experiment (flipping a coin) or a real world observation (measuring attendance rate).
- If observe several instances of a random variable, we get different values.
- Some values occur more than others and this information is captured by a probability distribution.

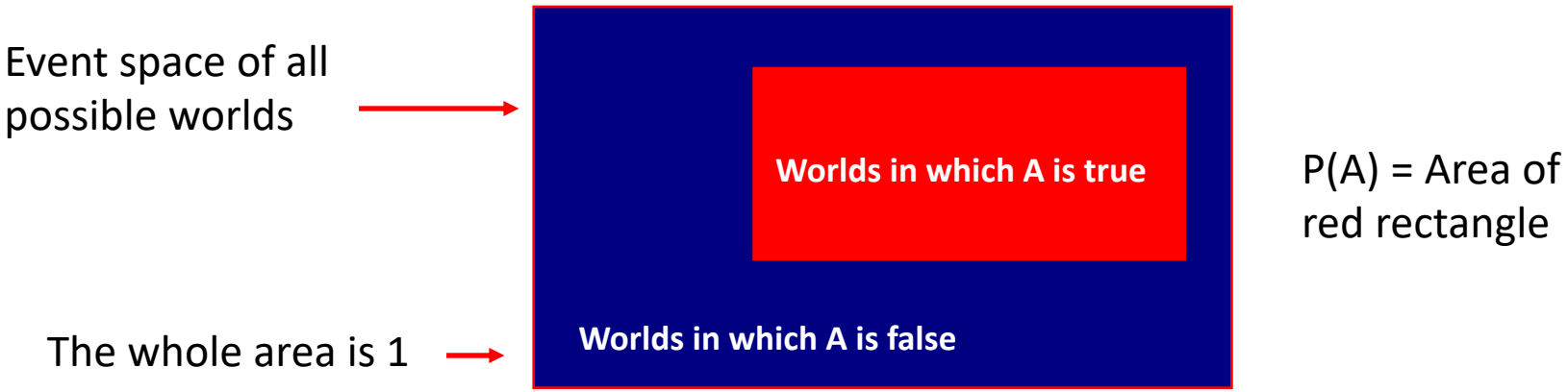
Random Variables

- A is a Boolean-valued random variable
  - if A denotes such event that there is some degree of uncertainty as to whether A occurs.

- Examples
  - A = The US president in 2023 will be male
  - A = You wake up tomorrow with a headache
  - A = You have Ebola

Probabilities

We write P(A) as
"the fraction of possible worlds in which A is true"



All probabilities between 0 and 1

$$0 <= P(A) <= 1$$

True proposition has probability 1, false has probability 0.

$$P(not\ A) = P(\sim A) = 1\text{-}P(A)$$

Multivalued Random Variables

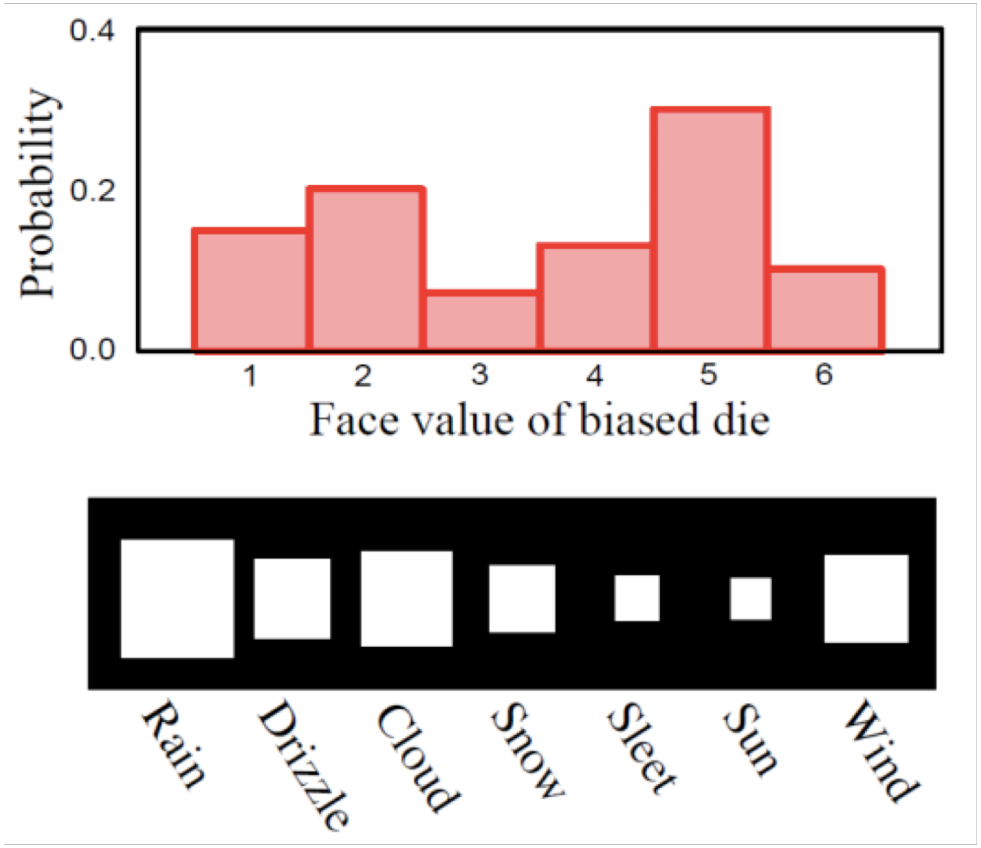
Suppose a random variable A can take on exactly one value out of $\{v_1, v_2, \ldots, v_k\}$

$$P(A = v_i \wedge A = v_j) = 0 \qquad \text{if } i \neq j$$

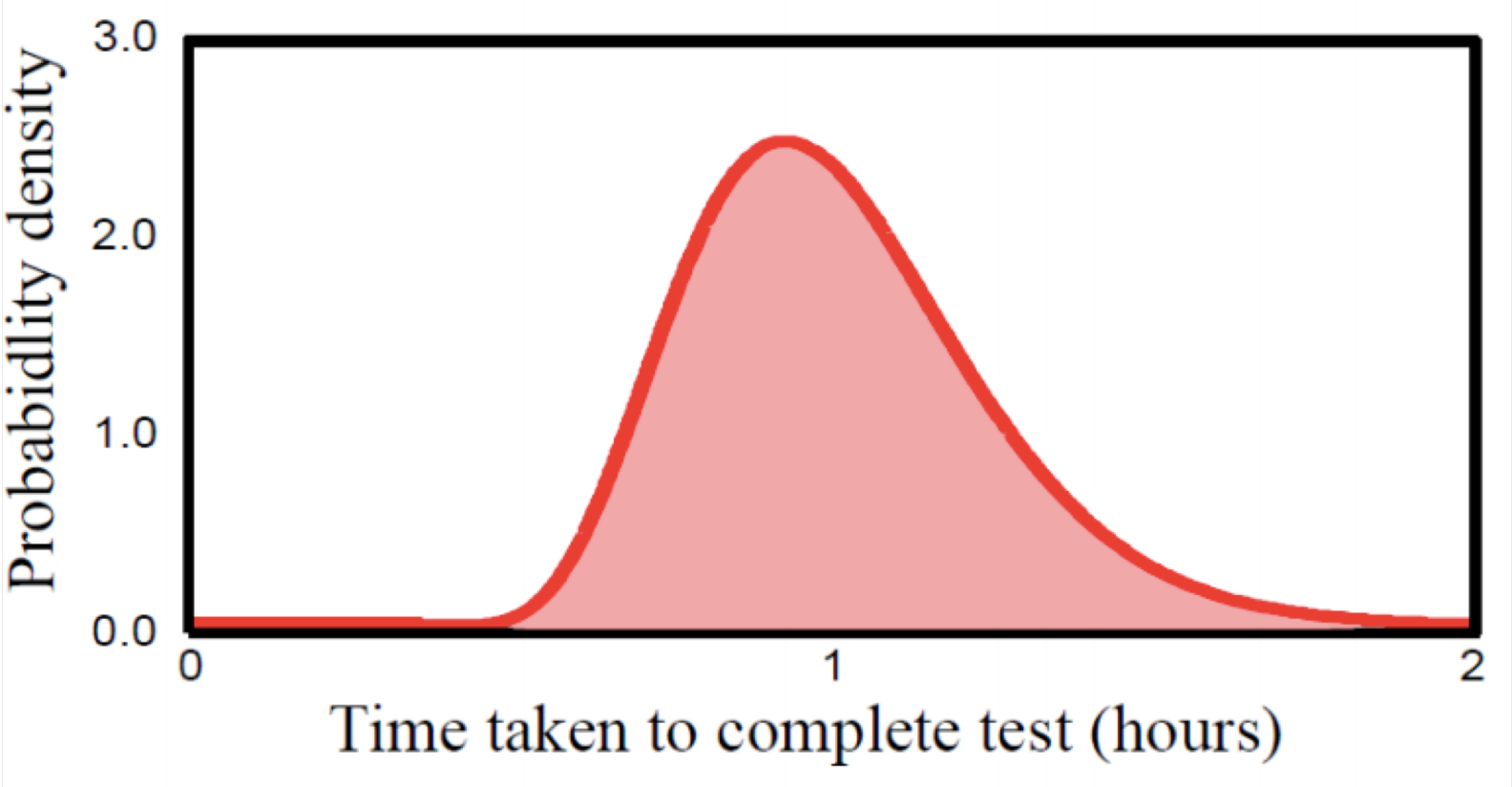
$$P(A = v_1 \vee A = v_2 \vee \cdots \vee A = v_k) = 1$$

Discrete vs Continuous Random Variable

Depends on whether outcome values are discrete or continuous



Continuous Random Variables

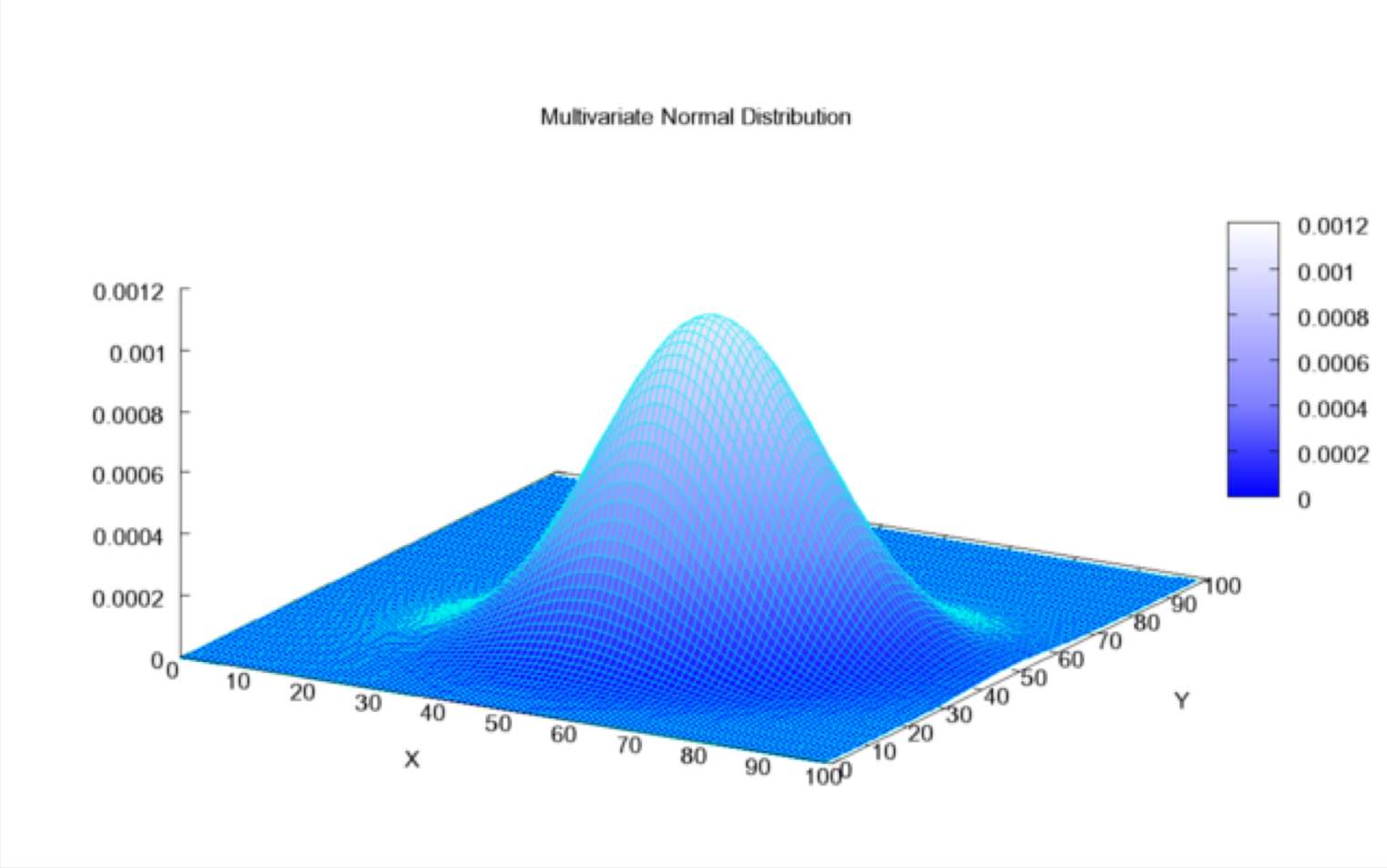


Note: Instead of summing over all possible outcome in discrete random variable, integrating the probability over all outcome should be 1 in continuous random variable.

Joint Probability

- Consider two random variables A and B.
- If we observe multiple paired instances, some combinations of outcomes are more likely than others.
- This is captured by joint probability $P(A \wedge B)$.

Joint Probability



Marginalization

$$P(A = v_1 \vee A = v_2 \vee \cdots \vee A = v_i) = \sum_{j=1}^{i} P(A = v_j)$$

$$P(B \wedge [A = v_1 \vee A = v_2 \vee \cdots \vee A = v_i]) = \sum_{j=1}^{i} P(B \wedge A = v_j)$$

$$P(B) = \sum_{j=1}^{k} P(B \wedge A = v_j)$$

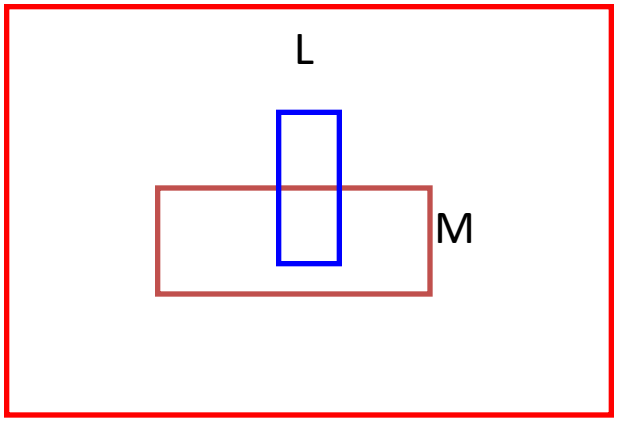
Marginalization: We can recover probability distribution of any variable in a joint distribution by summing over the other variables

Conditional Probability

$P(A|B)$ = Fraction of worlds in which B is true that also have A true.

M = "Miss 9am lecture"
L = "Go to bed late"

P(M) = 1/4
P(L) = 1/10
P(M|L) = 1/2



"Missing 9am class is rare and going to bed late is rarer. But if you go to bed late, there is a 50-50 chance you will miss 9am lecture."

Conditional Probability

P(A|B) = Fraction of worlds in which B is true that also have A true.

M = "Miss 9am lecture"
L = "Go to bed late"

P(M|L) = Fraction of late sleeper worlds in which also miss 9am lecture

P(M) = 1/4
P(L) = 1/10
P(M|L) = 1/2

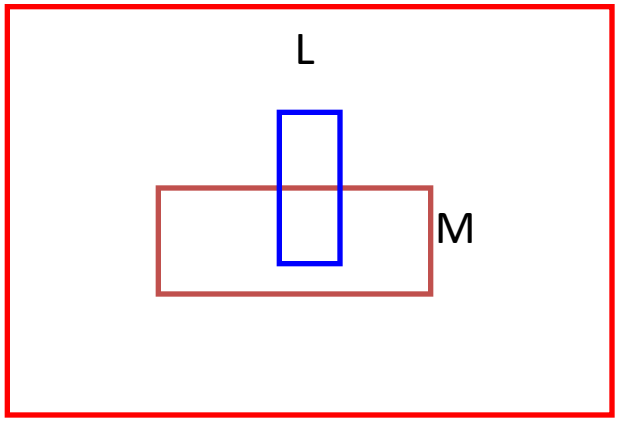
= #worlds of late sleeper and miss 9am class

#worlds late sleeper

= Area of "M and L" region

Area of "L" region

= $\frac{P(M \wedge L)}{P(L)}$



Definition of Conditional Probability

Conditional probability definition

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

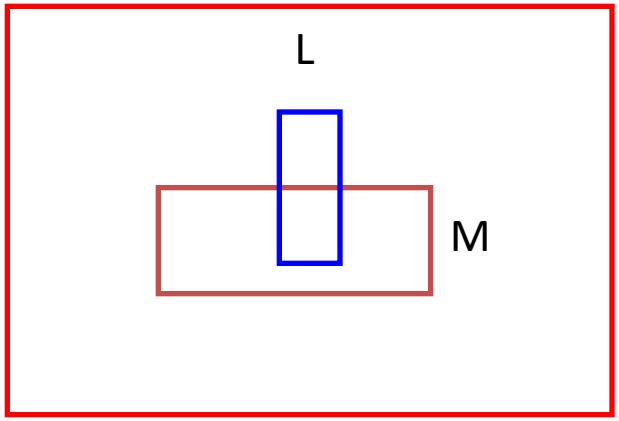
Hence, the Chain Rule

$$P(A \wedge B) = P(A|B)P(B)$$

Probabilistic Inference

M = "Miss 9am lecture"
L = "Go to bed late"

P(M) = 1/4
P(L) = 1/10
P(M|L) = 1/2



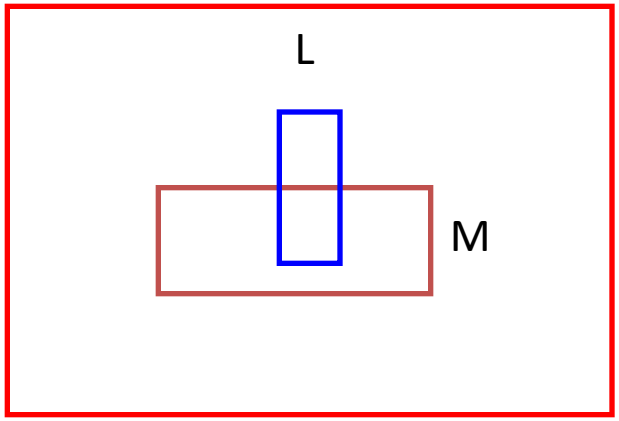
One day you miss my 9am class.
I think, "50% of late sleepers miss my 9am lecture, so this student (you) must have a 50-50 chance of going to bed late last night."

Is this thought reasonable?

Probabilistic Inference

M = "Miss 9am lecture"
L = "Go to bed late"

P(M) = 1/4
P(L) = 1/10
P(M|L) = 1/2



One day you miss my 9am class.
I think, "50% of late sleepers miss my 9am lecture, so this student (you) must have a 50-50 chance of going to bed late last night." ← **P(L|M)**

Is this thought reasonable?

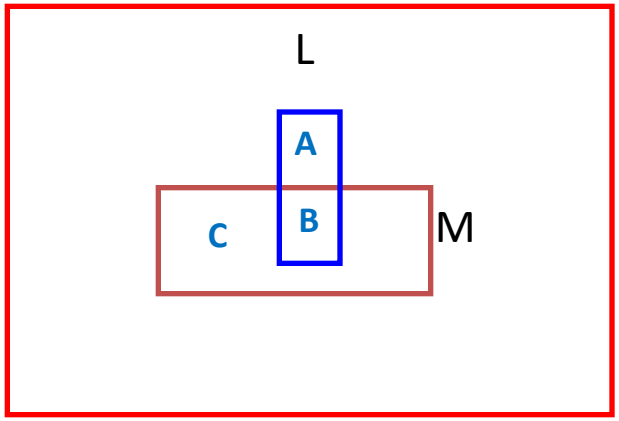
$$P(M \wedge L) = P(M|L)P(L)$$
$$= \frac{1}{2} * \frac{1}{10} = \frac{1}{20}$$

$$P(L|M) = \frac{P(L \wedge M)}{P(M)}$$
$$= \frac{1/20}{1/4} = \frac{1}{5}$$

Probabilistic Inference

M = "Miss 9am lecture"
L = "Go to bed late"

P(M) = 1/4
P(L) = 1/10
P(M|L) = 1/2



Area wise we have

P(L)=A+B

P(M)=B+C

P(M|L)=B/(A+B)

P(L|M)=B/(B+C) = P(M|L) P(L)/P(M)

Bayes Rule

- **Bayes, Thomas (1763)** An essay towards solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society of London, **53:370-418**

- **Bayes Rule**

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

Bayesian Learning

$$p(h \mid x) = \frac{P(x \mid h)P(h)}{P(x)}$$

Understanding Bayes' rule

x = data

h = hypothesis (model)

- rearranging

$p(h \mid x)P(x) = P(x \mid h)P(h)$

$P(x,h) = P(x,h)$

the same joint probability on both sides

$P(h)$: prior belief (probability of hypothesis *h* before seeing any data)

$P(x \mid h)$: likelihood (probability of the data if the hypothesis *h* is true)

$P(x) = \sum_h P(x \mid h)P(h)$: data evidence (marginal probability of the data)

$P(h \mid x)$: posterior (probability of hypothesis *h* after having seen the data *d*)

Choosing Hypotheses

- Generally, we want the most probable hypothesis(class label) given the observed data
  - Maximum a posteriori (**MAP**) hypothesis
  - Maximum likelihood (**ML**) hypothesis

Maximum A Posteriori (MAP)

- Maximum a posteriori (**MAP**) hypothesis

$$p(h \mid x) = \frac{P(x \mid h)P(h)}{P(x)}$$

$$h_{MAP} = \arg\max_{h \in H} p(h \mid x) = \arg\max_{h \in H} \frac{P(x \mid h)P(h)}{P(x)} = \arg\max_{h \in H} P(x \mid h)P(h)$$

Note $P(x)$ is independent of h, hence can be ignored.

Maximum Likelihood (ML)

$$h_{MAP} = \arg\max_{h \in H} P(x \mid h) P(h)$$

- Assuming that each hypothesis in H is equally probable, i.e., $P(h_i) = P(h_j)$, for all i and j, then we can drop $P(h)$ in MAP. $P(x|h)$ is often called the likelihood of data x given h. Any hypothesis that maximizes $P(x|h)$ is called the maximum likelihood hypothesis

$$h_{ML} = \arg\max_{h \in H} P(x \mid h)$$

An Illustrating Example

Classifying days according to whether someone will play tennis.

Each day is described by the attributes, Outlook, Temperature, Humidity and Wind.

Based on the training data in the table, classify the following instance

Outlook =	sunny
Temperature =	cool
Humidity =	high
Wind =	strong

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	*No*
Day2	Sunny	Hot	High	Strong	*No*
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	*No*
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	*No*
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	*No*

An Illustrating Example

Training sample pairs (X, D)

$X = (x_1, x_2, \ldots, x_n)$ is the feature vector representing the instance.

$D = (d_1, d_2, \ldots d_m)$ is the desired (target) output of the classifier

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	*No*
Day2	Sunny	Hot	High	Strong	*No*
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	*No*
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	*No*
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	*No*

An Illustrating Example

Training sample pairs (X, D)

$X = (x_1, x_2, \ldots, x_n)$ is the feature vector representing the instance.

$n = 4$

x_1= *outlook* = {*sunny, overcast, rain*}

x_2= *temperature* = {*hot, mild, cool*}

x_3= *humidity* = {*high, normal*}

x_4= *wind* = {*weak, strong*}

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	*No*
Day2	Sunny	Hot	High	Strong	*No*
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	*No*
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	*No*
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	*No*

An Illustrating Example

Training sample pairs (X, D)

$D = (d_1, d_2, \ldots d_m)$ is the desired (target) output of the classifier

$m = 1$

$d = Play\ Tennis = \{yes, no\}$

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

Bayesian Classifier

- The Bayesian approach to classifying a new instance X is to assign it to the most probable target value Y (MAP classifier)

$$
\begin{aligned}
Y &= \arg\max_{d_i \in d} p(d_i \mid X) \\
&= \arg\max_{d_i \in d} p(d_i \mid x_1, x_2, x_3, x_4) \\
&= \arg\max_{d_i \in d} \frac{p(x_1, x_2, x_3, x_4 \mid d_i) P(d_i)}{p(x_1, x_2, x_3, x_4)} \\
&= \arg\max_{d_i \in d} p(x_1, x_2, x_3, x_4 \mid d_i) P(d_i)
\end{aligned}
$$

Bayesian Classifier

$$Y = \arg\max_{d_i \in d} p(x_1, x_2, x_3, x_4 \mid d_i) P(d_i)$$

$P(d_i)$ is easy to calculate: simply counting how many times each target value d_i occurs in the training set.

$$P(d = yes) = 9/14$$

$$P(d = no) = 5/14$$

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	*No*
Day2	Sunny	Hot	High	Strong	*No*
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	*No*
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	*No*
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	*No*

Bayesian Classifier

$$Y = \arg\max_{d_i \in d} p(x_1, x_2, x_3, x_4 \mid d_i) P(d_i)$$

$P(x_1, x_2, x_3, x_4|d_i)$ is much more difficult to estimate.

In this simple example, there are 3x3x2x2x2 = 72 possible terms.

To obtain a reliable estimate, we need to see each terms many times.

Hence, we need a very, very large training set! (which in most cases is impossible to get).

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	*No*
Day2	Sunny	Hot	High	Strong	*No*
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	*No*
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	*No*
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	*No*

Naïve Bayes Classifier

Naïve Bayes classifier is based on the simplifying assumption that the attribute values are conditionally independent given the target value.

This means, we have

$$P(x_1, x_2, \cdots, x_n \mid d_i) = \prod_i P(x_i \mid d_i)$$

Naïve Bayes Classifier

$$Y = \arg\max_{d_i \in d} P(d_i) \prod_{k=1}^{4} P(x_k \mid d_i)$$

Back to the Example

Naïve Bayes Classifier

$$Y = \arg\max_{d_i \in d} \prod_{k=1}^{4} P(x_k \mid d_i) P(d_i)$$

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	*No*
Day2	Sunny	Hot	High	Strong	*No*
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	*No*
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	*No*
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	*No*

$$Y = \arg\max_{d_i \in \{yes,no\}} P(d_i) P(x_1 = suny \mid d_i) P(x_2 = cool \mid d_i) P(x_3 = high \mid d_i) P(x_4 = strong \mid d_i)$$

Back to the Example

$$Y = \arg\max_{d_i \in \{yes,no\}} P(d_i)P(x_1 = suny \mid d_i)P(x_2 = cool \mid d_i)P(x_3 = high \mid d_i)P(x_4 = strong \mid d_i)$$

$P(d=yes)=9/14 = 0.64$

$P(d=no)=5/14 = 0.36$

$P(x_1=sunny|yes)=2/9$
$P(x_1=sunny|no)=3/5$

$P(x_2=cool|yes)=$
$P(x_2=cool|no)=$

$P(x_3=high|yes)=$
$P(x_3=high|no)=$

$P(x_4=strong|yes)=3/9$
$P(x_4=strong|no)=3/5$

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

Back to the Example

$$Y = \arg\max_{d_i \in \{yes,no\}} P(d_i)P(x_1 = suny \mid d_i)P(x_2 = cool \mid d_i)P(x_3 = high \mid d_i)P(x_4 = strong \mid d_i)$$

$$P(yes)P(x_1 = suny \mid yes)P(x_2 = cool \mid yes)P(x_3 = high \mid yes)P(x_4 = strong \mid yes) = 0.0053$$

$$P(no)P(x_1 = suny \mid no)P(x_2 = cool \mid no)P(x_3 = high \mid no)P(x_4 = strong \mid no) = 0.0206$$

$$Y = Play\ Tennis = no$$

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	*No*
Day2	Sunny	Hot	High	Strong	*No*
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	*No*
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	*No*
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	*No*

Estimating Probabilities

- So far, we estimate the probabilities by the fraction of times the event is observed to occur over the entire opportunities

- In the above example, we estimated

  $P(wind=strong|play\ tennis=no)=Nc/N$,
  where $N = 5$ is the total number of training samples for which $play\ tennis = no$, and Nc is the number of these for which wind=strong

- What happens if $Nc = 0$?

Estimating Probabilities

- When Nc is small, however, such approach provides poor estimation. To avoid this difficulty, we can adopt the **m-estimate** of probability

$$\frac{N_C + mP}{N + m}$$

where P is the prior estimate of the probability we wish to estimate, m is a constant called the equivalent sample size.

A typical method for choosing P in the absence of other information is to assume uniform priors: If an attribute has k possible values we set $P=1/K$.

For example, $P(\ wind = strong \mid play\ tennis = no)$, we note wind has two possible values, so uniform priors means $P = ½$

Another Illustrative Example

- **Car theft Example**
  - Attributes are Color , Type , Origin, and the subject, stolen can be either yes or no.

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

Another Illustrative Example

- **Car theft Example**
  - We want to classify a Red Domestic SUV.
  - Note there is no example of a Red Domestic SUV in our 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

Another Illustrative Example

- We need to estimate

$$P(x_i \mid d_j) = \frac{N_C + mP}{N + m}$$

N = the number of training examples for which $d = d_j$
Nc = number of examples for which $d = d_j$ and $x = x_i$
P = a priori estimate for $P(x_i|d_j)$
m = the equivalent sample size

Another Illustrative Example

- To classify a Red, Domestic, SUV, we need to estimate

$$Y = \arg\max_{d_i \in \{yes, no\}} P(d_i)P(x_1 = RED \mid d_i)P(x_2 = SUV \mid d_i)P(x_3 = Domestic \mid d_i)$$

Another Illustrative Example

Yes:

Red:
N = 5
Nc= 3
P = .5
m = 3

SUV:
N = 5
Nc = 1
P = .5
m = 3

Domestic:
N = 5
Nc = 2
P = .5
m = 3

No:

Red:
N = 5
Nc = 2
P = .5
m = 3

SUV:
N = 5
Nc = 3
P = .5
m = 3

Domestic:
N = 5
Nc = 3
P = .5
m =3

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

Another Illustrative Example

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

$$P(Red|Yes) = \frac{3 + 3 * .5}{5 + 3} = .56 \qquad P(Red|No) = \frac{2 + 3 * .5}{5 + 3} = .43$$

$$P(SUV|Yes) = \frac{1 + 3 * .5}{5 + 3} = .31 \qquad P(SUV|No) = \frac{3 + 3 * .5}{5 + 3} = .56$$

$$P(Domestic|Yes) = \frac{2 + 3 * .5}{5 + 3} = .43 \qquad P(Domestic|No) = \frac{3 + 3 * .5}{5 + 3} = .56$$

Another Illustrative Example

- To classify a Red, Domestic, SUV, we need to estimate

$$Y = \arg\max_{d_i \in \{yes,no\}} P(d_i)P(x_1 = RED \mid d_i)P(x_2 = SUV \mid d_i)P(x_3 = Domestic \mid d_i)$$

$$P(yes)P(x_1 = RED \mid yes)P(x_2 = SUV \mid yes)P(x_3 = Domestic \mid yes)$$
$$= 0.5 * 0.56 * 0.31 * 0.43 = 0.037$$

$$P(no)P(x_1 = RED \mid no)P(x_2 = SUV \mid no)P(x_3 = Domestic \mid no)$$
$$= 0.5 * 0.43 * 0.56 * 0.56 = 0.069$$

$$Y = no$$

Naïve Bayesian Classifier

- What about using Naïve Bayesian Classifier for our handwritten digit recognition problem?
  - Each pixel is an x_i. There will be 784 x's.
  - Digit label is d_k. Note there will be 10 possible *d's*.
  - $P(d_k)$ can be calculated by counting number of training images for the digit, divided to total number of training images.
  - $P(x_i|d_k)$ can be calculated by counting number of images for a given digit, given pixel position, and given an intensity value, divided by number of training images with that digit.

Naïve Bayesian Classifier

$P(x_i|d_k)$ can be calculated by counting number of images for a given digit, given pixel position, and given an intensity value, divided by number of training images with that digit.

For example,

$$P(x_1 = 255|d = 0)$$
$$= \frac{Number\ of\ images\ whose\ first\ pixel\ value\ is\ 255\ and\ contain\ digit\ 0}{Total\ number\ of\ images\ contain\ digit\ 0}$$

Naïve Bayesian Classifier

- For a given input image X and given digit label d_k, calculate $P(d_k)$ and all $P(x_i|d_k)$
- For each digit label d_k, calculate
  $P(d_k|X) = P(d_k)P(x_1 = 0|d_k)P(x_2 = 255|d_k) \dots P(x_{784} = 0)$
- Choose the digit label k that give the max value according to above calculation.

0	255	0	0	0	0	0	0	0	0
0	0	0	0	255	255	0	0	0	0
0	0	0	255	0	0	255	0	0	0
0	0	0	0	0	0	255	0	0	0
0	0	0	0	0	255	0	0	0	0
0	0	0	0	255	0	0	0	0	0
0	0	0	255	0	0	0	0	0	0
0	0	0	255	0	0	0	0	0	0
0	0	0	255	0	0	0	0	0	0
0	0	0	255	255	255	255	255	255	0

Input image X

Further Readings

Chapter 6, T. M. Mitchell, Machine Learning, McGraw-Hill International Edition, 1997