# *Methods for Intelligent Systems*

## *Lecture Notes on Machine Learning*

Matteo Matteucci

matteucci@elet.polimi.it

Department of Electronics and Information

Politecnico di Milano



# *Probability for Dataminers*

## *– Probability Basics –*

## Probability and Boolean Random Variables

**Boolean-valued random variable** $A$ is a Boolean-valued random variable if $A$ denotes an event, and there is some degree of uncertainty as to whether $A$ occurs.

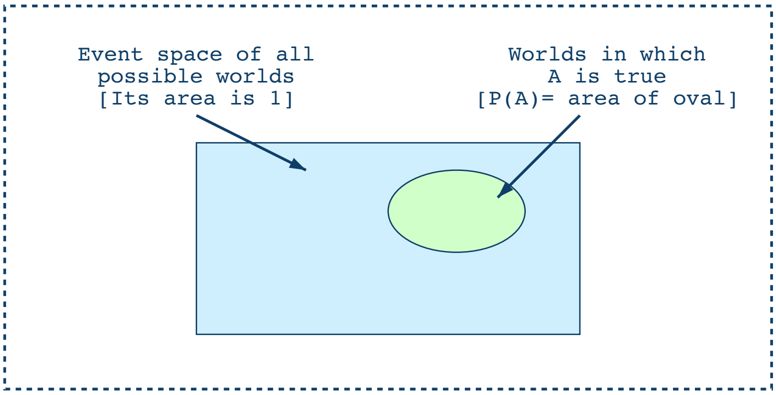
- Examples
  - A = The US president in 2023 will be male
  - A = You wake up tomorrow with a headache
  - A = You like the "Gladiator"



## Probability and Boolean Random Variables

**Boolean-valued random variable** $A$ is a Boolean-valued random variable if $A$ denotes an event, and there is some degree of uncertainty as to whether $A$ occurs.

**Probability of** $A$ "the fraction of possible worlds in which A is true"

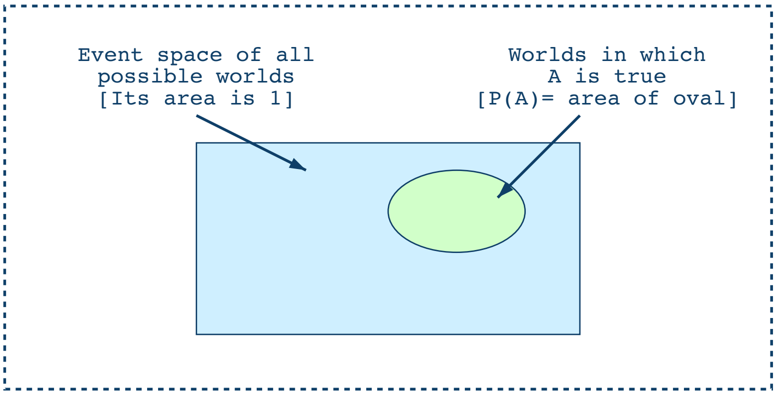



Note: this is one of the possible definitions. We won't go into the philosophy of it!

## Probability Axioms

Define the whole set of possible worlds with the label `true` and the empty set with `false`:

- $0 \leq P(A) \leq 1$
- $P(A = \texttt{true}) = 1; P(A = \texttt{false}) = 0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

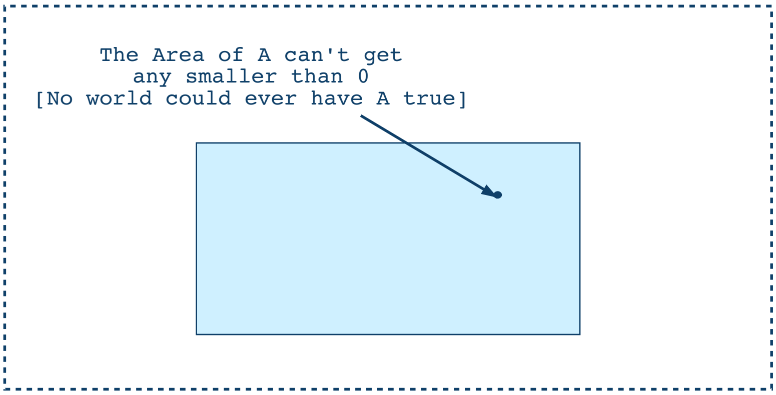





## Probability Axioms

Define the whole set of possible worlds with the label `true` and the empty set with `false`:

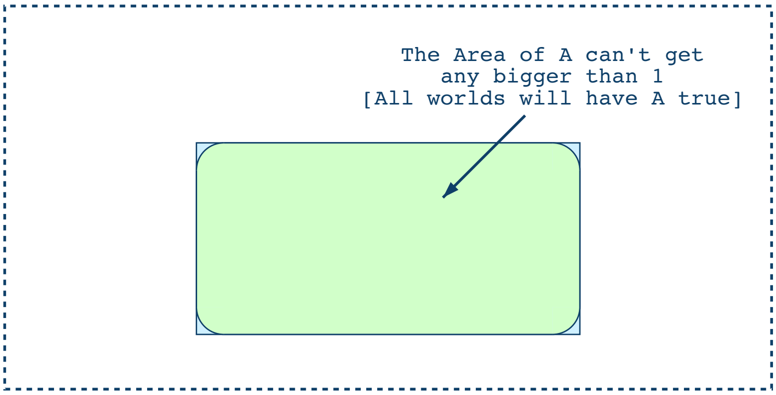
- $0 \leq P(A) \leq 1$
- $P(A = \texttt{true}) = 1; P(A = \texttt{false}) = 0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$



## Probability Axioms

Define the whole set of possible worlds with the label `true` and the empty set with `false`:

- $0 \leq P(A) \leq 1$
- $P(A = \texttt{true}) = 1; P(A = \texttt{false}) = 0$
- $P(A \lor B) = P(A) + P(B) - P(A \land B)$

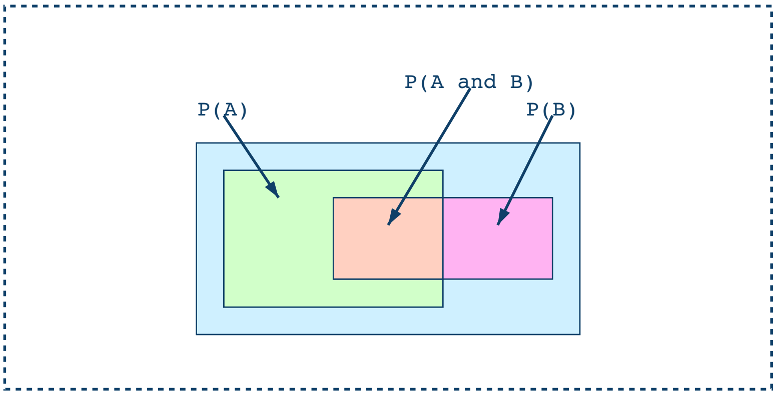





## Probability Axioms

Define the whole set of possible worlds with the label `true` and the empty set with `false`:

- $0 \leq P(A) \leq 1$
- $P(A = \texttt{true}) = 1; P(A = \texttt{false}) = 0$
- $P(A \lor B) = P(A) + P(B) - P(A \land B)$



## Theorems From the Axioms (I)

Using the axioms:

- $P(A = \texttt{true}) = 1; P(A = \texttt{false}) = 0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

Proove: $P(\sim A) = P(\bar{A}) = 1 - P(A)$

$$\begin{aligned}
\texttt{true} &= A \vee \bar{A} \\
P(\texttt{true}) &= P(A \vee \bar{A}) \\
&= P(A) + P(\bar{A}) - P(A \wedge \bar{A}) \\
&= P(A) + P(\bar{A}) - P(\texttt{false}) \\
1 &= P(A) + P(\bar{A}) - 0 \\
1 - P(A) &= P(\bar{A})
\end{aligned}$$



## Theorems From the Axioms (II)

Using the axioms:

- $P(A = \texttt{true}) = 1; P(A = \texttt{false}) = 0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

Proove: $P(A) = P(A \wedge B) + P(A \wedge \bar{B})$

$$\begin{aligned}
A &= A \wedge \texttt{true} \\
&= A \wedge (B \vee \bar{B}) \\
&= (A \wedge B) \vee (A \wedge \bar{B}) \\
P(A) &= P((A \wedge B) \vee (A \wedge \bar{B})) \\
&= P(A \wedge B) + P(A \wedge \bar{B}) - P((A \wedge B) \wedge (A \wedge \bar{B})) \\
&= P(A \wedge B) + P(A \wedge \bar{B}) - P(\texttt{false}) \\
&= P(A \wedge B) + P(A \wedge \bar{B})
\end{aligned}$$

## Multivalued Random Variables

**Multivalued random variable** $A$ is a *random variable of arity* $k$ if it can take on exactly one values out of $\{v_1, v_2, \ldots, v_k\}$.

We still have the probability axioms plus

- $P(A = v_i \wedge A = v_j) = 0$ if $i \neq j$
- $P(A = v_1 \vee A = v_2 \vee \ldots \vee A = v_k) = 1$

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

Proove: $\sum_{j=1}^{k} P(A = v_j) = 1$

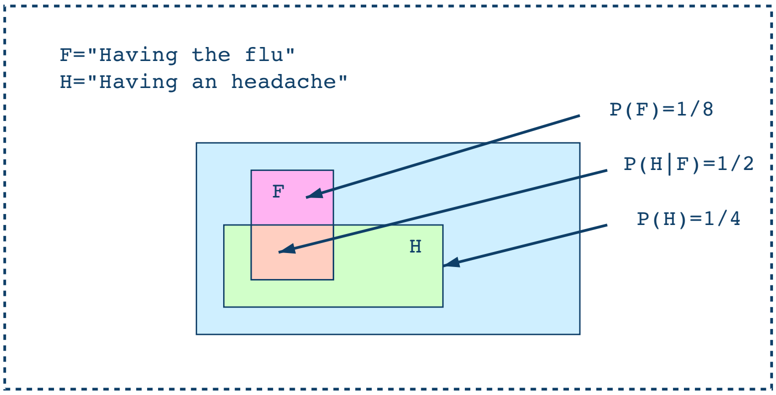
Proove:
$P(B \wedge [A = v_1 \vee A = v_2 \vee \ldots \vee A = v_i]) = \sum_{j=1}^{i} P(B \wedge A = v_j)$

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



## Conditional Probability

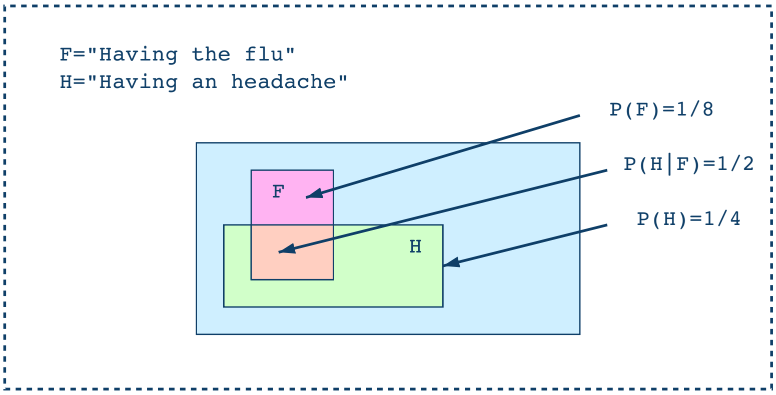
**Probability of $A$ given $B$:** "the fraction of possible worlds in which $B$ is true that also have $A$ true"




*"Sometimes I've the flu and sometimes I've a headache, but half of the times I'm with the flu I've also a headache!"*

## Conditional Probability

**Probability of $A$ given $B$:** "the fraction of possible worlds in which $B$ is true that also have $A$ true"

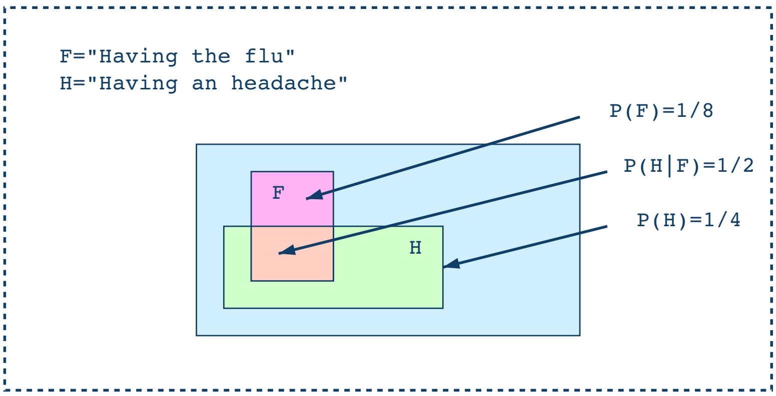



$$P(H|F) = \frac{\text{Num. of worlds with F and H}}{\text{Num. worlds with F}} = \frac{P(H \wedge F)}{P(F)}$$



## Probabilistic Inference

One day you wake up with a headache and you think: *"Half of the flus are associated with headaches so I must have* $50\%$ *chance of getting the flu".*




Is this reasoning correct?

$$P(F|H) = \frac{P(F \wedge H)}{P(H)} = \frac{P(H \wedge F)}{P(H)} = \frac{P(H|F) * P(F)}{P(H)} = \frac{1/2 * 1/8}{1/4} = 1/4$$

## Theorems that we used (and will use)

In doing the previous inference we have used two famous theorems:

- Chain rule

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

- Bayes theorem

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

We can have more general formulae:

- $P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|\bar{A})P(\bar{A})}$
- $P(A|B \wedge X) = \frac{P(B|A \wedge X)P(A \wedge X)}{P(B \wedge X)}$
- $P(A = v_i|B) = \frac{P(B|A=v_i)P(A=v_i)}{\sum_{k=1}^{n_A} P(B|A=v_k)P(A=v_k)}$



## Independent Variables

**Independent variables:** Assume $A$ and $B$ are boolean random variables; $A$ and $B$ are independent (denote it with $A \perp B$) if and only if:

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

Using the definition:

- $P(A|B) = P(A)$

Proove: $P(A \wedge B) = P(A)P(B)$

$$\begin{aligned} P(A \wedge B) &= P(A|B)P(B) \\ &= P(A)P(B) \end{aligned}$$

Proove: $P(B|A) = P(B)$

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

# *Probability for Dataminers*

## *– Information Gain –*



## Information and Bits

Your mission, if you decide to accept it, will be:

*“Transmit a set of independent random samples of $X$ over a binary serial link.”*

1. Starring at $X$ for a while, you notice that it has olny four possible values: A, B, C, D
2. You decide to transmit the data encoding each reading with two bits:

$$A = 00, B = 01, C = 10, D = 11.$$

Mission Accomplished!

## Information and “Fewer Bits”

Your mission, if you decide to accept it, will be:

*“The previous code uses $2$ bits for symbol. Knowing that the probabilities are not equal: P(X=A)=1/2, P(X=B)=1/4, P(X=C)=1/8, P(X=D)=1/8, invent a coding for your transmission that only uses $1.75$ bits on average per symbol.”*

1. You decide to transmit the data encoding each reading with a different number of bits:

$$A = 0, B = 10, C = 110, D = 111.$$

Mission Accomplished!



## Information and Entropy

Suppose $X$ can have one of $m$ values with probability

$$P(X = V_1) = p_1, \ldots, P(X = V_m) = p_m.$$

What’s the smallest possible number of bits, on average, per symbol, needed to transmit a stream of symbols drawn from $X$’s distribution?

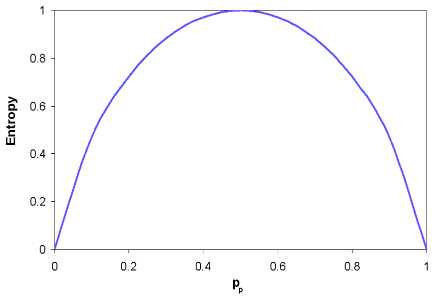
$$\begin{aligned} H(X) &= -p_1 \log_2 p_1 - p_2 \log_2 p_2 - \ldots - p_m \log_2 p_m \\ &= -\sum_{j=1}^{m} p_j \log_2 p_j = \textit{Entropy} \text{ of } X \end{aligned}$$

“Good idea! But what is entropy anyway?”

## Entropy: “What is it anyway?”

Simple Case:

- $X$ has $2$ values $\oplus$ and $\ominus$
- $p_\oplus$ probability of $\oplus$
- $p_\ominus = 1 - p_\oplus$ probability of $\ominus$




$$H(X) = -p_\ominus \log_2 p_\ominus - p_\oplus \log_2 p_\oplus$$

Entropy measures “disorder” or “uniformity in distribution”

1. *High Entropy*: $X$ is very “disordered” $\longrightarrow$ “interesting”
2. *Low Entropy*: $X$ is very “ordered” $\longrightarrow$ “boring”



## Useful Facts on Logarithms

Just for you to know it might be useful to review a couple of formulas to be used in calculation:

- $\ln x \times y = \ln x + \ln y$
- $\ln \frac{x}{y} = \ln x - \ln y$
- $\ln x^y = y \times \ln x$
- $\log_2 x = \frac{\ln x}{\ln 2} = \frac{\log_{10} x}{\log_{10} 2}$
- $\log_a x = \frac{}{\log_b a}$
- $\log_2 0 = -\infty$ (the formula is no good for a probability of 0)

Now we can practice with a simple example!

## Specific Conditional Entropy

| X | Y |
|---|---|
| Math | Yes |
| History | No |
| CS | Yes |
| Math | No |
| Math | No |
| CS | Yes |
| Hystory | No |
| Math | Yes |

Suppose we are interested in predicting output $Y$ from input $X$ where

- $X$ = University subject
- $Y$ = Likes the movie "Gladiator"

From this data we can estimate

- P(Y = Yes) = 0.5
- P(X = Math) = 0.5
- P(Y = Yes | X = History) = 0

Definition of Specific Conditional Entropy:

- H(Y|X=v): *the entropy of $Y$ only for those records in which $X$ has value $v$*
  - H(Y|X=Math) = 1
  - H(Y|X=History) = 0



## Conditional Entropy

| X | Y |
|---|---|
| Math | Yes |
| History | No |
| CS | Yes |
| Math | No |
| Math | No |
| CS | Yes |
| Hystory | No |
| Math | Yes |

Definition of Conditional Entropy H(Y|X):

- *The average $Y$ specific conditional entropy*
- *Expected number of bits to transmit Y if both sides will know the value of X*
- $\sum_j P(X = v_j) H(Y|X = v_j)$

Definition of Conditional Entropy H(Y|X):

- $\sum_j P(X = v_j) H(Y|X = v_j)$

| $v_j$ | $P(X = v_j)$ | $H(Y\|X = v_j)$ |
|---|---|---|
| Math | $0.5$ | $1$ |
| Hystory | $0.25$ | $0$ |
| CS | $0.25$ | $0$ |

H(Y|X) = ?

## Information Gain

| X | Y |
|---|---|
| Math | Yes |
| History | No |
| CS | Yes |
| Math | No |
| Math | No |
| CS | Yes |
| Hystory | No |
| Math | Yes |

*I must transmit Y on a binary serial line.*
*How many bits on average would it save me if both ends of the line knew X?*

$$
\begin{aligned}
IG(Y|X) &= H(Y) - H(Y|X) \\
&= 1 - 0.5 = 0.5
\end{aligned}
$$

Information Gain measures the “information” provided by $X$ to predict $Y$

This IS about Machine Learning!



## Relative Information Gain

| X | Y |
|---|---|
| Math | Yes |
| History | No |
| CS | Yes |
| Math | No |
| Math | No |
| CS | Yes |
| Hystory | No |
| Math | Yes |

*I must transmit Y on a binary serial line.*
*What fraction of the bits on average would it save me if both ends of the line knew X?*

$$
\begin{aligned}
RIG(Y|X) &= (H(Y) - H(Y|X))/H(Y) \\
&= (1 - 0.5)/1 = 0.5
\end{aligned}
$$

Well, we’ll find soon Information Gain and Relative Information gain talking about supervised learning with Decision Trees ...

## Why is Information Gain Useful?

Your mission, if you decide to accept it, will be:

*"Predict whether someone is going live past 80 years."*

From historical data you might find:

- IG(LongLife | HairColor) = 0.01
- IG(LongLife | Smoker) = 0.2
- IG(LongLife | Gender) = 0.25
- IG(LongLife | LastDigitOfSSN) = 0.00001

What you should look at?