Knowledge Engineering and Expert Systems

Lecture Notes on Machine Learning

Matteo Mattecci

matteucci@elet.polimi.it

Department of Electronics and Information

Politecnico di Milano

Probability for Dataminers

– Probability Basics –

Probability and Boolean Random Variables

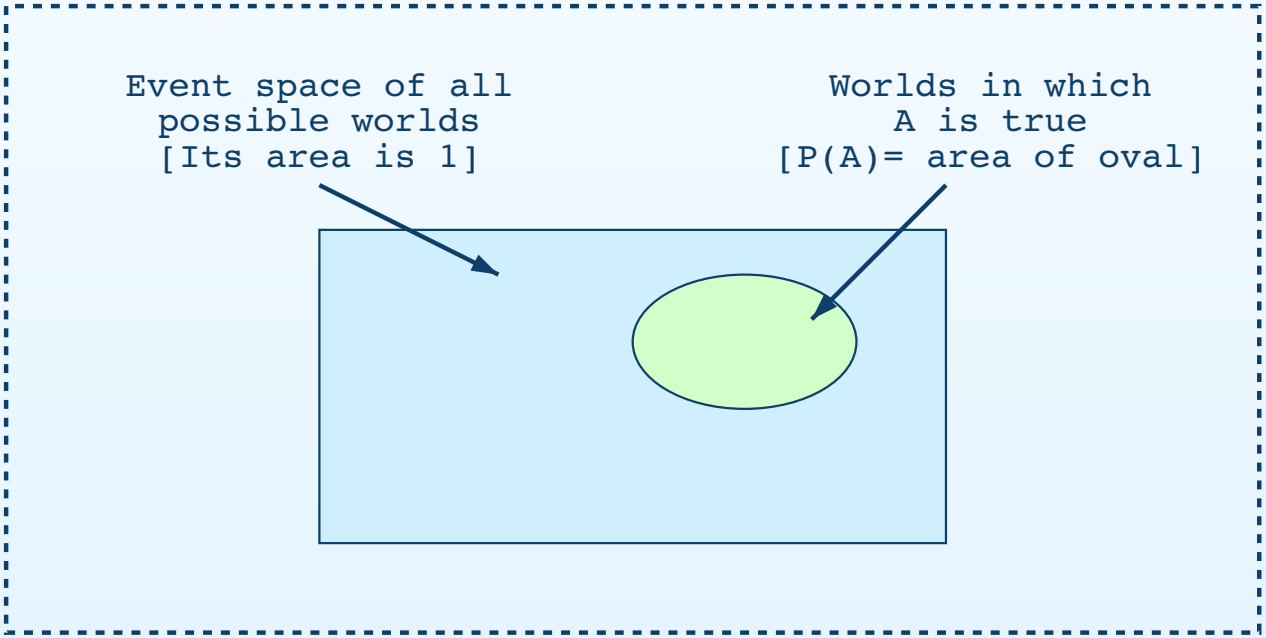
Boolean-valued random variable A is a Boolean-valued random variable if A denotes an event, and there is some degree of uncertainty as to whether A occurs.

- 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.

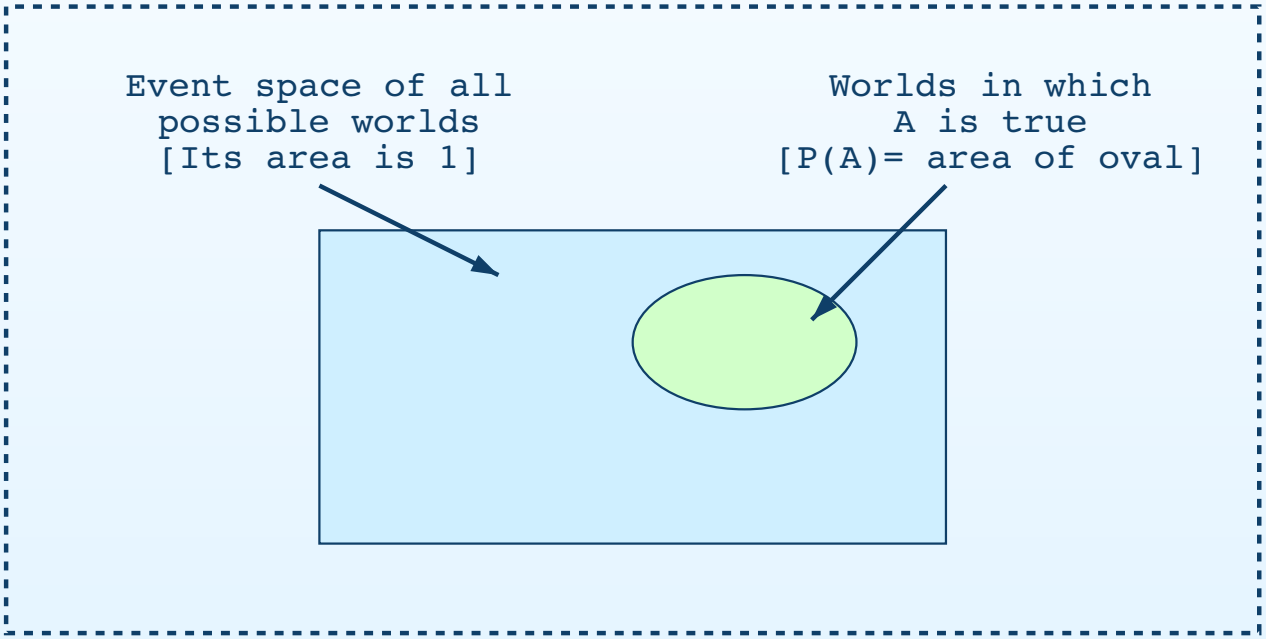
Probability of A "the fraction of possible worlds in which A is true"



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

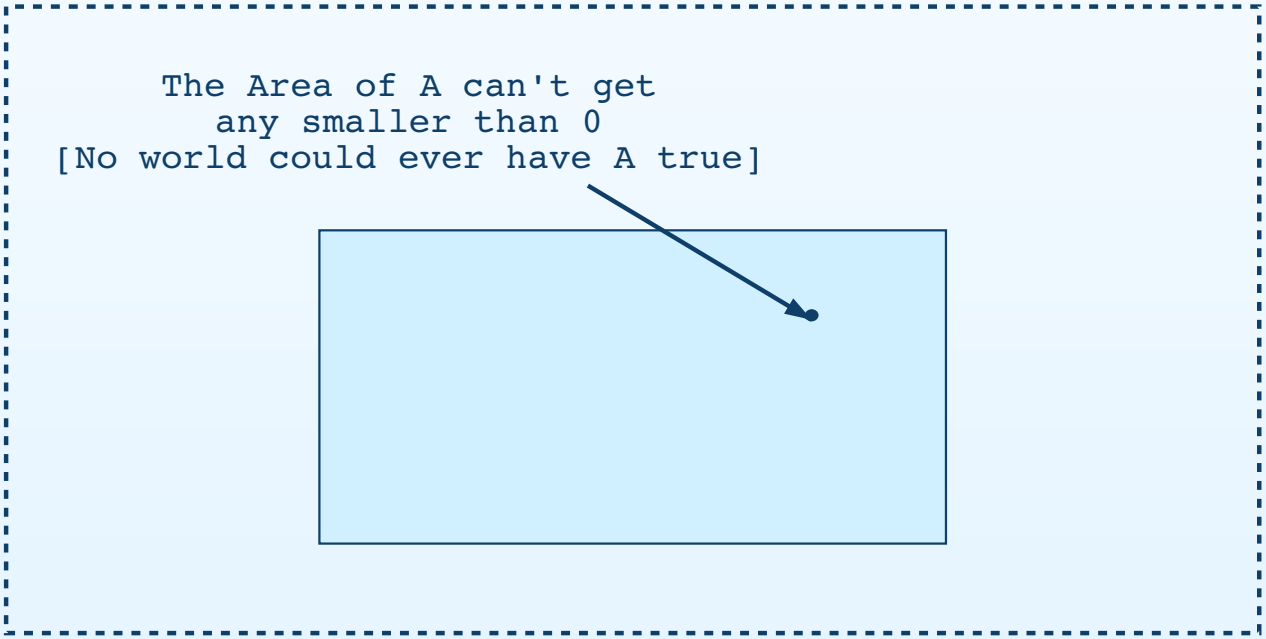
Probability Axioms

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



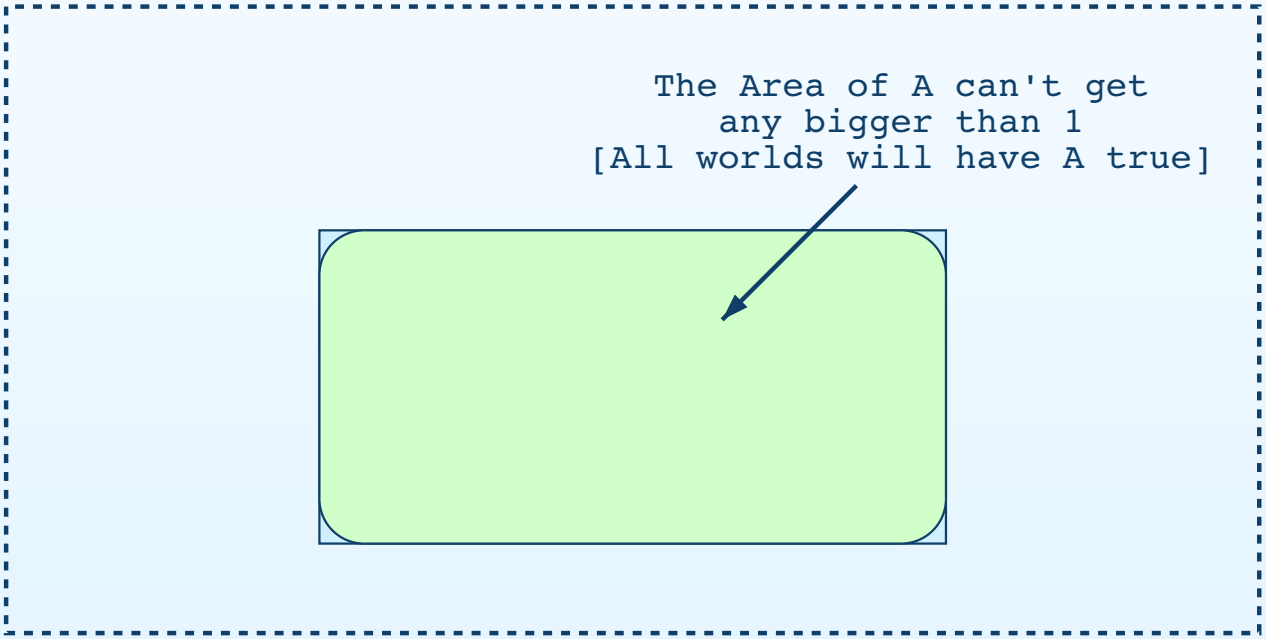
Probability Axioms

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



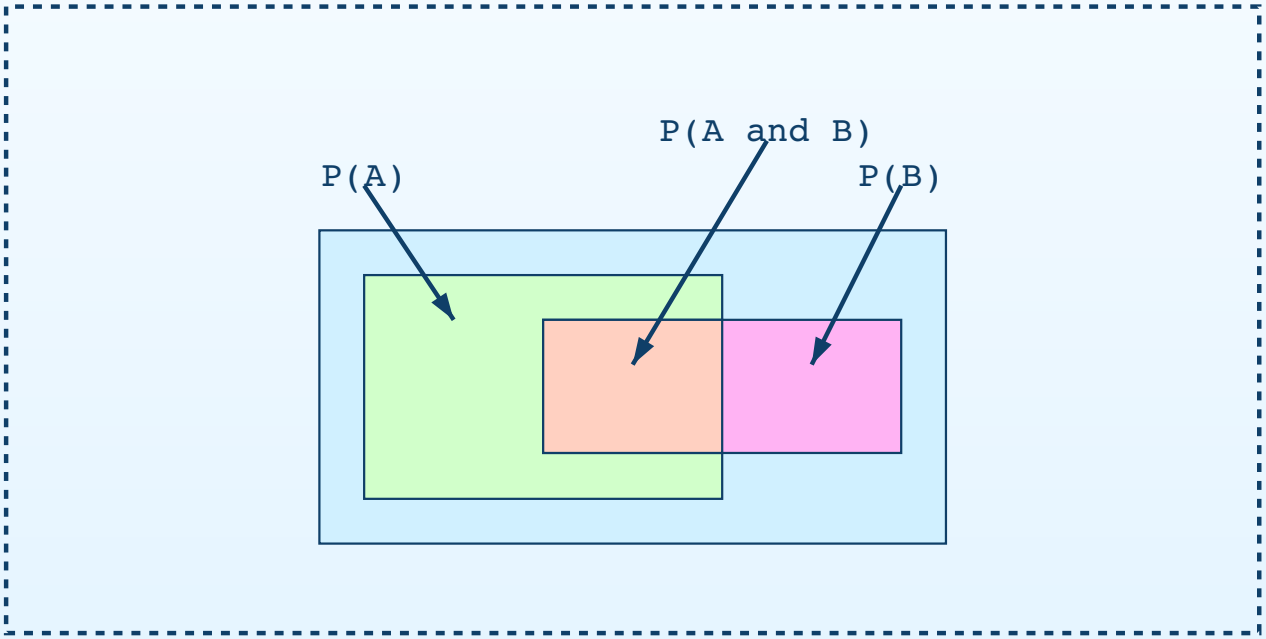
Probability Axioms

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



Probability Axioms

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



Theorems From the Axioms (I)

Using the axioms:

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

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

Theorems From the Axioms (I)

Using the axioms:

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

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:

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

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

Theorems From the Axioms (II)

Using the axioms:

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

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 *ranom 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 $i \neq j$
- $P(A = v_1 \vee A = v_2 \vee \ldots \vee A = v_k) = 1$

Multivalued Random Variables

Multivalued random variable A is a *ranom 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: $P(B \wedge [A = v_1 \vee A = v_2 \vee \ldots \vee A = v_i]) = \sum_{j=1}^{i} P(B \wedge A = v_j)$

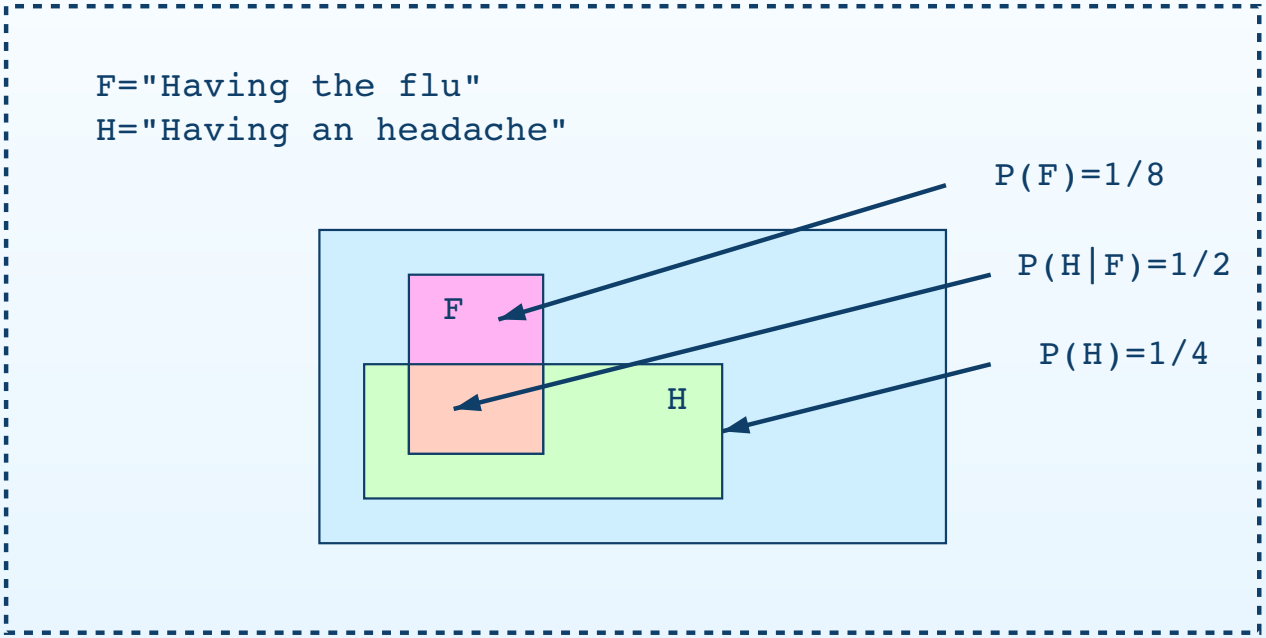
Proove: $P(B) = \sum_{j=1}^{k} P(B \wedge A = v_j)$

Conditional Probability

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

Conditional Probability

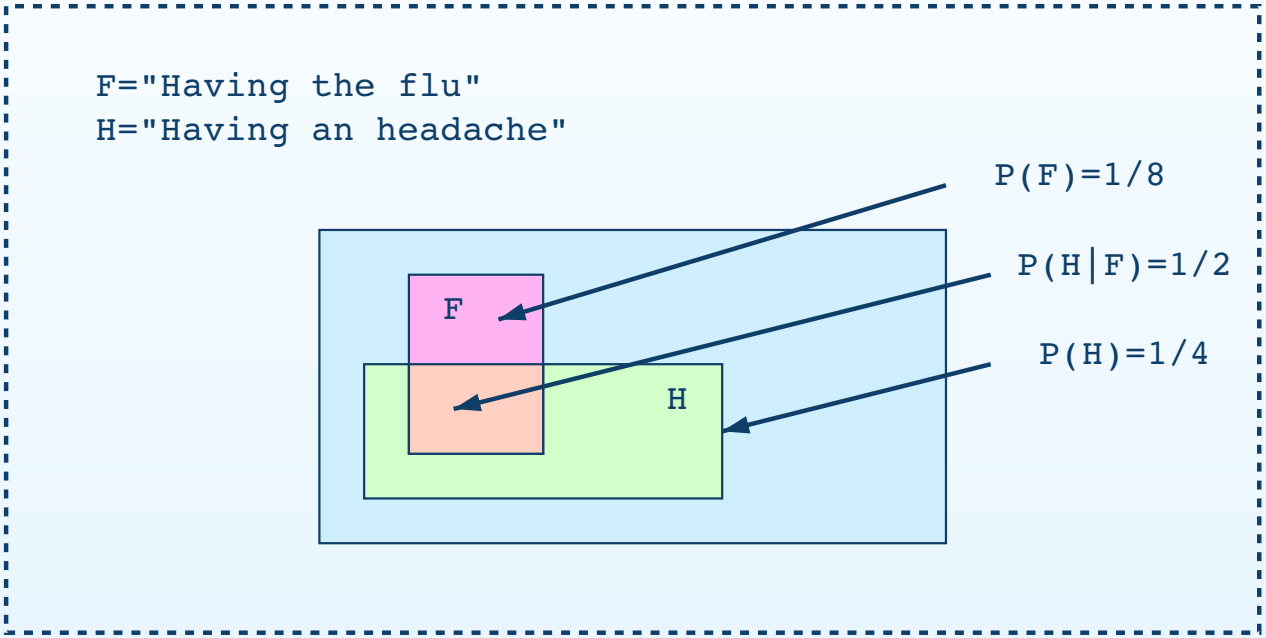
Probability of A given B: "the fraction of possible worlds in which B is true that also have A true"



"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

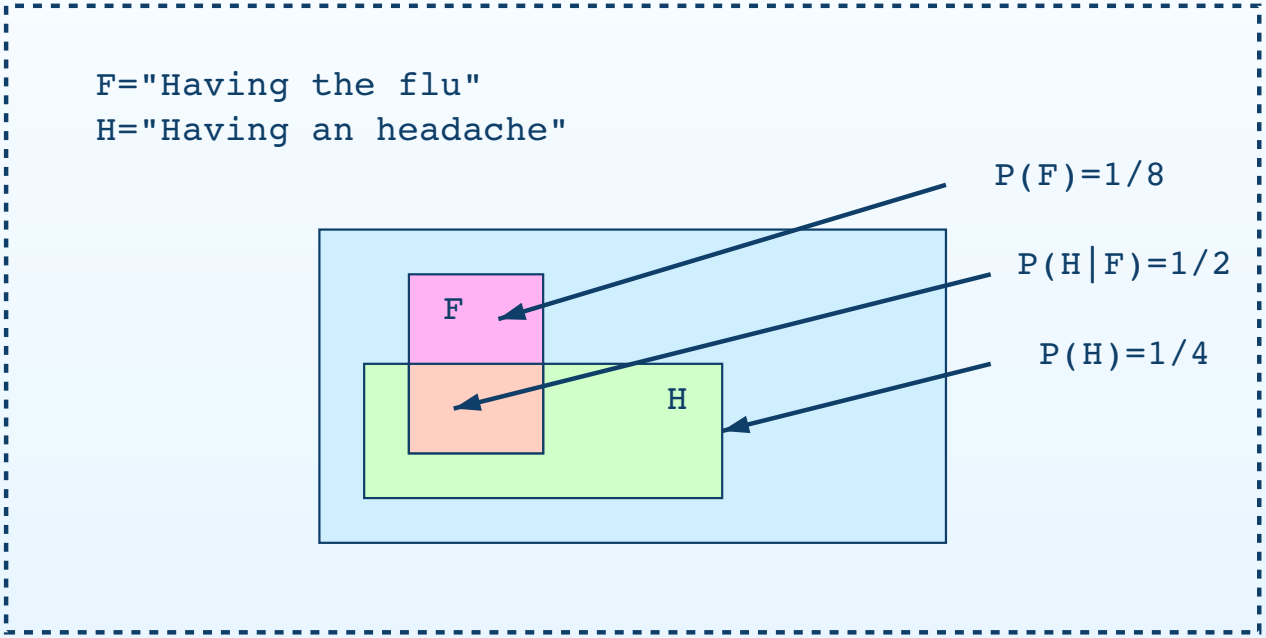
Probability of A given B: "the fraction of possible worlds in which B is true that also have A true"



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

Probabilistic Inference

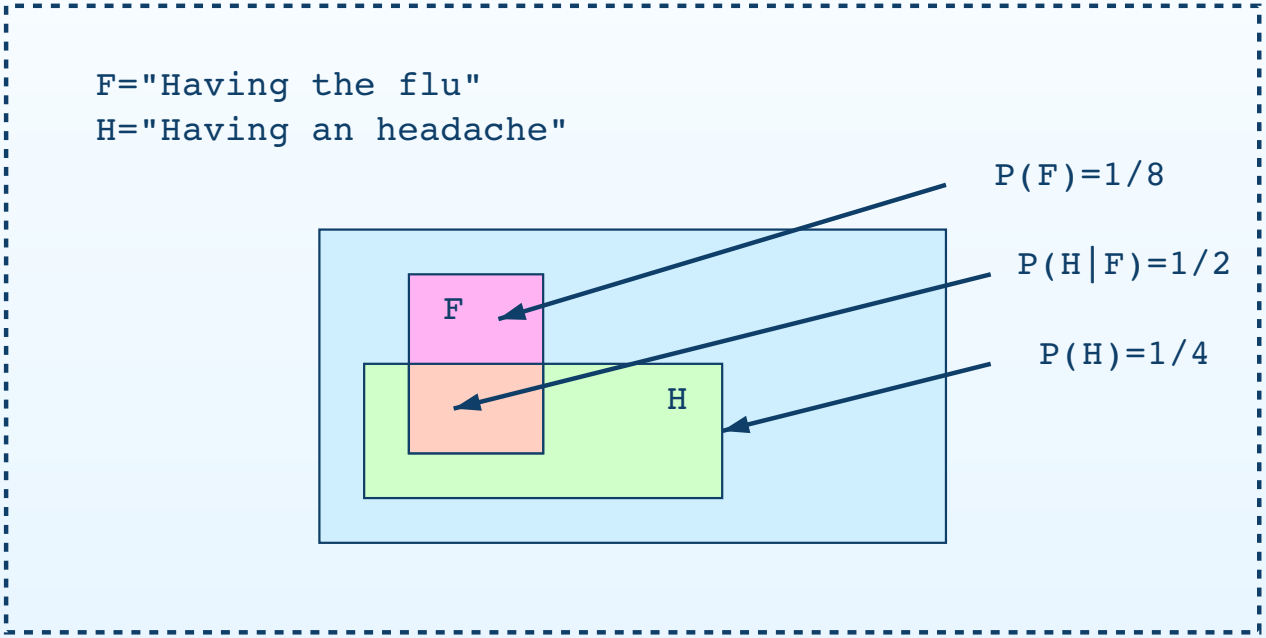
One day you wake up with a headache and you think: *"Half of the flus are associated with headaches so I must have* 50% *chance of getting the flu".*



Is this reasoning correct?

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".*



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

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

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

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:

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

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 defi nition:

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

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

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 defi nition:

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

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

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

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 defi nition:

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

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

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 defi nition:

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

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

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

Unsupervised Learning

– Density Estimation –

The world is a very unceratin place!

Thus there have been attempts to use different methodologies for dealing with world uncertainty:

- Probability theory
- Fuzzy logic
- Three-valued logic
- Dempster-Shafer
- Non-monotonic reasoning

The world is a very unceratin place!

Thus there have been attempts to use different methodologies for dealing with world uncertainty:

- Probability theory
- Fuzzy logic
- Three-valued logic
- Dempster-Shafer
- Non-monotonic reasoning

In the next we'll focus on **Probabilistic Modelling**

“Why a probabilistic approach?”

A probabilistic **model of the data** can be used to:

- Make inference about missing inputs
- Generate prediction/fantasies/imagery
- Make decisions which minimise expected loss
- Communicate the data in an effi cient way

“Why a probabilistic approach?”

A probabilistic **model of the data** can be used to:

- Make inference about missing inputs
- Generate prediction/fantasies/imagery
- Make decisions which minimise expected loss
- Communicate the data in an effi cient way

Statistical modeling is equivalent to other views of learning:

- Information theoretic: fi nding compact representations of the data
- Physical analogies: minimising free energy of a corresponding statistical mechanical system

The Joint Distribution

How to make a joint distribution of M variables:

1. Make a truth table listing all combination of values
2. For each combination state/compute how probable it is
3. Check that all probabilities sum up to 1

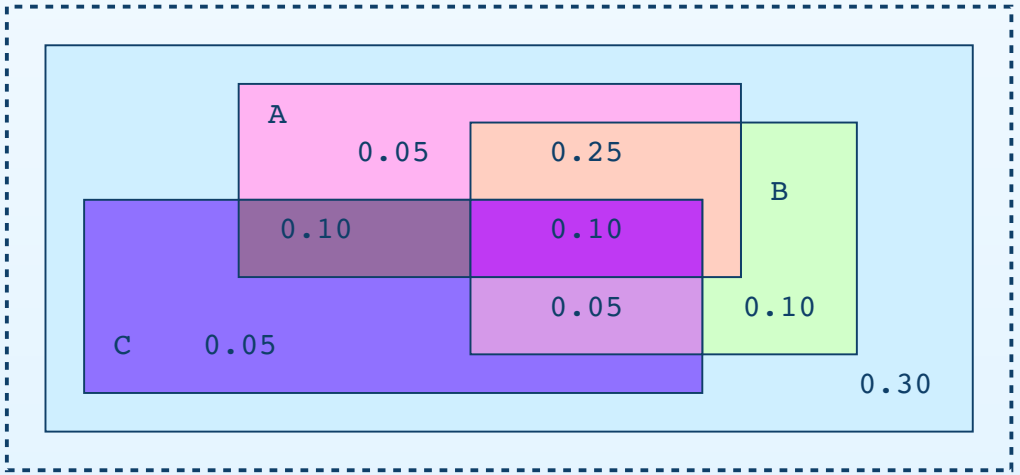
The Joint Distribution

How to make a joint distribution of M variables:

1. Make a truth table listing all combination of values
2. For each combination state/compute how probable it is
3. Check that all probabilities sum up to 1

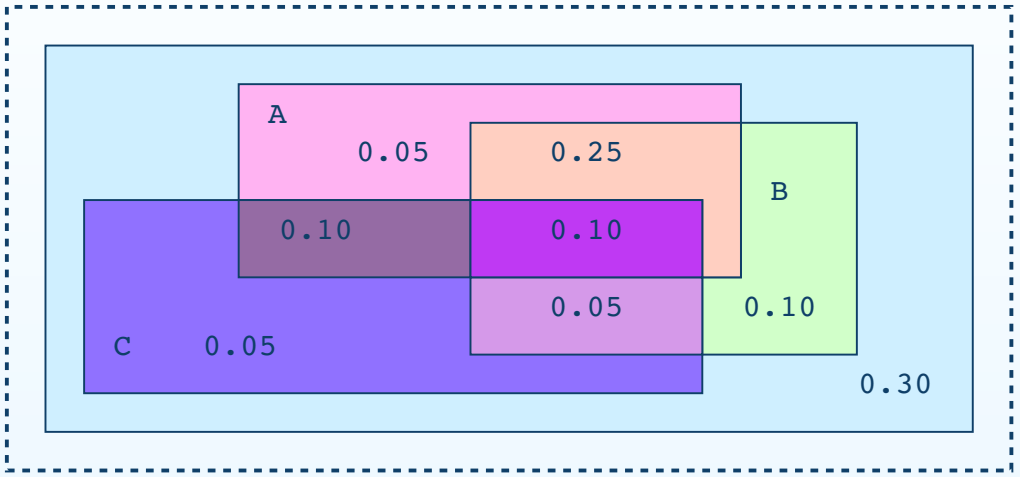
Example with 3 boolean variables A, B and C.

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10



Using the Joint Distribution (I)

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10

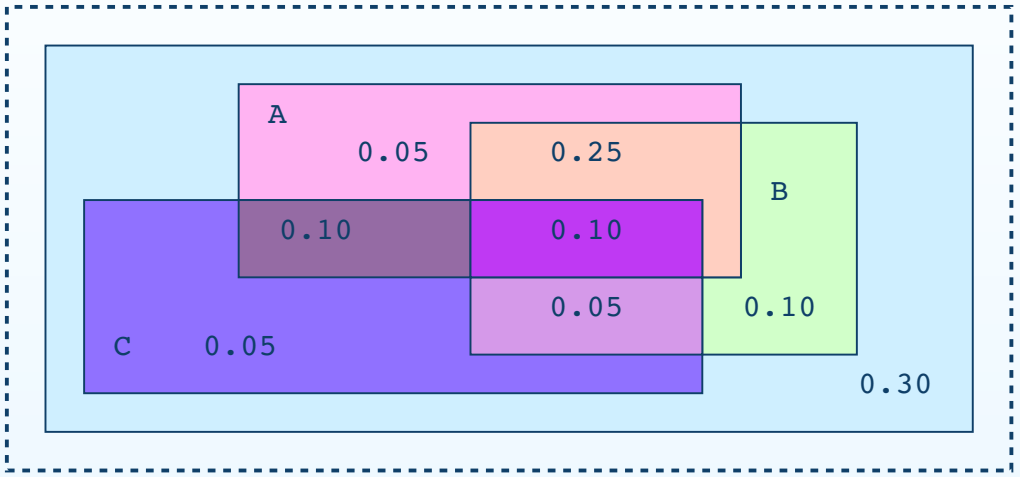


You can use it to compute the **probability of logical expression**:

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

Using the Joint Distribution (I)

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10

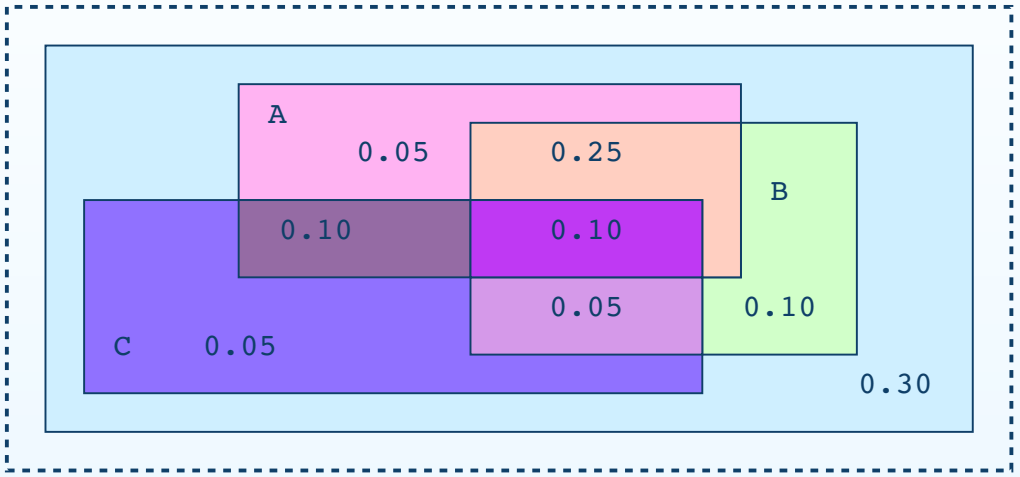


You can use it to compute the **probability of logical expression**:

- $P(A) = 0.05 + 0.10 + 0.25 + 0.10 = 0.6$
- $P(A \wedge B) = 0.25 + 0.10 = 0.35$
- $P(\bar{A} \vee C) = 0.30 + 0.05 + 0.10 + 0.05 + 0.05 + 0.25$

Using the Joint Distribution (II)

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10

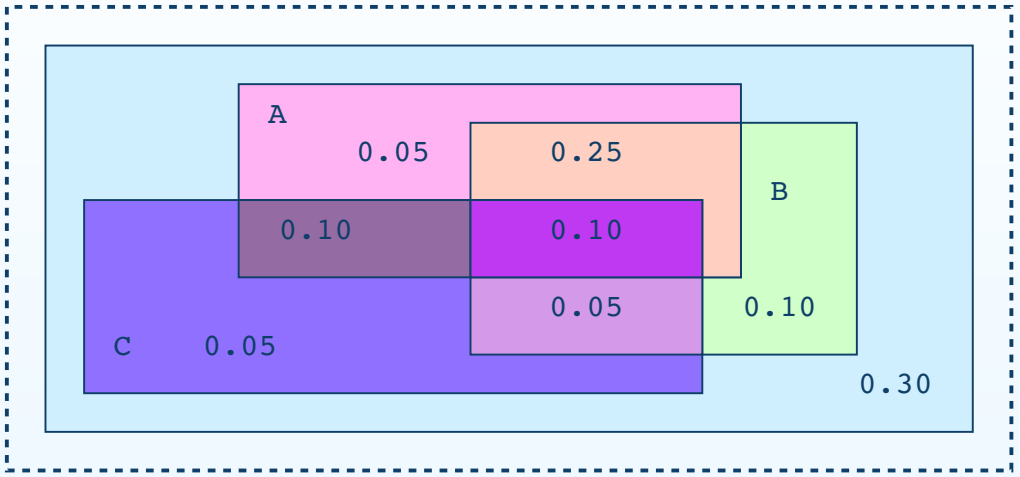


Now you can use it for **making inference**:

$$P(E_1|E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{row matching } E_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{row matching } E_2} P(\text{row})}$$

Using the Joint Distribution (II)

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10



Now you can use it for **making inference**:

- $P(A|B) = (0.25 + 0.10)/(0.10 + 0.05 + 0.25 + 0.10) = 0.35/0.50 = 0.70$
- $P(C|A \wedge B) = (0.10)/(0.25 + 0.10) = 0.10/0.35 = 0.285$
- $P(\bar{A}|C) = (0.05 + 0.05)/(0.05 + 0.05 + 0.10 + 0.10) = 0.10/0.30 = 0.333$

Setting up a Joint Distribution

Now we know what they are and how to use them, but where do Joint Distributions come from?

- Expert Humans
- Simpler probabilistic facts and some algebra
  - Suppose you knew

$$\begin{aligned} P(A) &= 0.7 \\ P(B|A) &= 0.2 \\ P(B|\sim A) &= 0.1 \\ P(C|A \wedge B) &= 0.1 \end{aligned}$$

$$\begin{aligned} P(C|A \wedge \sim B) &= 0.8 \\ P(C|\sim A \wedge B) &= 0.3 \\ P(C|\sim A \wedge \sim B) &= 0.1 \end{aligned}$$

Then you can automatically compute the JD using the chain rule

- **Learn them from data!**

Joint Distribution Estimator

A Density Estimator learns a mapping from a set of attributes to a probability distribution over the attributes space

Joint Distribution Estimator

A Density Estimator learns a mapping from a set of attributes to a probability distribution over the attributes space

Our Joint Distribution learner is our first example of something called Density Estimation

- Build a Joint Distribution table for your attributes in which the probabilities are unspecified
- The fill in each row with

$$\hat{P}(\text{row}) = \frac{\text{records matching row}}{\text{total number of records}}$$

Joint Distribution Estimator

A Density Estimator learns a mapping from a set of attributes to a probability distribution over the attributes space

Our Joint Distribution learner is our first example of something called Density Estimation

- Build a Joint Distribution table for your attributes in which the probabilities are unspecified
- The fill in each row with

$$\hat{P}(\text{row}) = \frac{\text{records matching row}}{\text{total number of records}}$$

How can we evaluate it?

Evaluating a Density Estimator

We can use **likelihood** for evaluating density estimation:

- Given a record $\mathbf{x}$, a density estimator M tells you how likely it is
$$\hat{P}(\mathbf{x}|M)$$
- Given a dataset with R records, a density estimator can tell you how likely the dataset is under the assumption that all records were **independently** generated from the density estimator's joint distribution
$$\hat{P}(\text{dataset}) = \hat{P}(\mathbf{x}_1 \wedge \mathbf{x}_2 \wedge \ldots \wedge \mathbf{x}_R|M) = \prod_{k=1}^{R} \hat{P}(\mathbf{x}_k|M)$$

Evaluating a Density Estimator

We can use **likelihood** for evaluating density estimation:

- Given a record $\mathbf{x}$, a density estimator M tells you how likely it is
$$\hat{P}(\mathbf{x}|M)$$
- Given a dataset with R records, a density estimator can tell you how likely the dataset is under the assumption that all records were **independently** generated from the density estimator's joint distribution
$$\hat{P}(\text{dataset}) = \hat{P}(\mathbf{x}_1 \wedge \mathbf{x}_2 \wedge \ldots \wedge \mathbf{x}_R|M) = \prod_{k=1}^{R} \hat{P}(\mathbf{x}_k|M)$$

Since likelihood can get too small we usually use **log-likelihood**:

$$\log \hat{P}(\text{dataset}) = \log \prod_{k=1}^{R} \hat{P}(\mathbf{x}_k|M) = \sum_{k=1}^{R} \log \hat{P}(\mathbf{x}_k|M)$$

Joint Distribution Summary

Now we have a way to learn a Joint Density estimator from data

- Joint Density estimators can do many **good** things
  - Can sort the records by probability, and thus spot weird records (e.g., anomaly/outliers detection)
  - Can do inference: $P(E_1|E_2)$ (e.g., Automatic Doctor, Help Desk)
  - Can be used for Bayes Classifi ers (see later)

Joint Distribution Summary

Now we have a way to learn a Joint Density estimator from data

- Joint Density estimators can do many **good** things
  - Can sort the records by probability, and thus spot weird records (e.g., anomaly/outliers detection)
  - Can do inference: $P(E_1|E_2)$ (e.g., Automatic Doctor, Help Desk)
  - Can be used for Bayes Classifi ers (see later)
- Joint Density estimators can **badly** overfi t!
  - Joint Estimator just mirrors the training data
  - Suppose you see a **new dataset**, its likelihood is going to be:

$$\log \hat{P}(\text{new dataset}|M) = \sum_{k=1}^{R} \log \hat{P}(\mathbf{x}_k|M) = -\infty$$
$$if\ \exists k:\ \hat{P}(\mathbf{x}_k|M) = 0$$

Joint Distribution Summary

Now we have a way to learn a Joint Density estimator from data

- Joint Density estimators can do many **good** things
  - Can sort the records by probability, and thus spot weird records (e.g., anomaly/outliers detection)
  - Can do inference: $P(E_1|E_2)$ (e.g., Automatic Doctor, Help Desk)
  - Can be used for Bayes Classifi ers (see later)
- Joint Density estimators can **badly** overfi t!
  - Joint Estimator just mirrors the training data
  - Suppose you see a **new dataset**, its likelihood is going to be:

$$\log \hat{P}(\text{new dataset}|M) = \sum_{k=1}^{R} \log \hat{P}(\mathbf{x}_k|M) = -\infty$$
$$if\ \exists k:\ \hat{P}(\mathbf{x}_k|M) = 0$$

We need something which generalizes! $\rightarrow$ **Naïve Density Estimator**

Naïve Density Estimator

The naïve model assumes that each attribute is distributed independently of any of the other attributes.

- Let $\mathbf{x}[i]$ denote the i^{th} fi eld of record $\mathbf{x}$.
- The Naïve Density Estimator says that

$$\mathbf{x}[i] \perp \{\mathbf{x}[1], \mathbf{x}[2], \ldots, \mathbf{x}[i-1], \mathbf{x}[i+1], \ldots, \mathbf{x}[M]\}$$

Naïve Density Estimator

The naïve model assumes that each attribute is distributed independently of any of the other attributes.

- Let $\mathbf{x}[i]$ denote the i^{th} fi eld of record $\mathbf{x}$.
- The Naïve Density Estimator says that

$$\mathbf{x}[i] \perp \{\mathbf{x}[1], \mathbf{x}[2], \ldots, \mathbf{x}[i-1], \mathbf{x}[i+1], \ldots, \mathbf{x}[M]\}$$

Example: suppose to randomly shake a green dice and a red dice

- Dataset 1: A = red value, B = green value
- Dataset 2: A = red value, B = sum of values
- Dataset 3: A = sum of values, B = difference of values

Which of these datasets violates the naïve assumption?

Naïve Density Estimator

The naïve model assumes that each attribute is distributed independently of any of the other attributes.

- Let $\mathbf{x}[i]$ denote the i^{th} fi eld of record $\mathbf{x}$.
- The Naïve Density Estimator says that

$$\mathbf{x}[i] \perp \{\mathbf{x}[1], \mathbf{x}[2], \ldots, \mathbf{x}[i-1], \mathbf{x}[i+1], \ldots, \mathbf{x}[M]\}$$

From a Naïve Distribution you can compute the Joint Distribution:

- Suppose A, B, C, D are independently distributed

$$P(A \wedge \bar{B} \wedge C \wedge \bar{D}) \quad = \quad ?$$

Naïve Density Estimator

The naïve model assumes that each attribute is distributed independently of any of the other attributes.

- Let $\mathbf{x}[i]$ denote the i^{th} fi eld of record $\mathbf{x}$.
- The Naïve Density Estimator says that

$$\mathbf{x}[i] \perp \{\mathbf{x}[1], \mathbf{x}[2], \ldots, \mathbf{x}[i-1], \mathbf{x}[i+1], \ldots, \mathbf{x}[M]\}$$

From a Naïve Distribution you can compute the Joint Distribution:

- Suppose A, B, C, D are independently distributed

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

Learning a Naïve Density Estimator

Suppose $\mathbf{x}[1], \mathbf{x}[2], \ldots, \mathbf{x}[M]$ are independently distributed

- Once we have the Naïve Distribution, we can construct any row of the implied Joint Distribution on demand

$$P(\mathbf{x}[1] = u_1, \mathbf{x}[2] = u_2, \ldots, \mathbf{x}[M] = u_M) = \prod_{k=1}^{M} P(\mathbf{x}[k] = u_k)$$

- We can do any inference!

Learning a Naïve Density Estimator

Suppose $\mathbf{x}[1], \mathbf{x}[2], \ldots, \mathbf{x}[M]$ are independently distributed

- Once we have the Naïve Distribution, we can construct any row of the implied Joint Distribution on demand

$$P(\mathbf{x}[1] = u_1, \mathbf{x}[2] = u_2, \ldots, \mathbf{x}[M] = u_M) = \prod_{k=1}^{M} P(\mathbf{x}[k] = u_k)$$

- We can do any inference!

But how do we learn a Naïve Density Estimator?

Learning a Naïve Density Estimator

Suppose $\mathbf{x}[1], \mathbf{x}[2], \ldots, \mathbf{x}[M]$ are independently distributed

- Once we have the Naïve Distribution, we can construct any row of the implied Joint Distribution on demand

$$P(\mathbf{x}[1] = u_1, \mathbf{x}[2] = u_2, \ldots, \mathbf{x}[M] = u_M) = \prod_{k=1}^{M} P(\mathbf{x}[k] = u_k)$$

- We can do any inference!

But how do we learn a Naïve Density Estimator?

$$\hat{P}(\mathbf{x}[i] = u) = \frac{\text{number of record for which } \mathbf{x}[i] = u}{\text{total number of records}}$$

Joint Density vs. Naïve Density

What we got so far?

- Joint Distribution Estimator
  - Can model anything
  - Given 100 records and more than 6 Boolean attributes will perform poorly
  - Can easily overfi t the data
- Naïve Distribution Estimator
  - Can model only very boring distributions
  - Given 100 records and 10,000 multivalued attributes will be fi ne
  - Quite robust to overfi tting

Joint Density vs. Naïve Density

What we got so far?

- Joint Distribution Estimator
  - Can model anything
  - Given 100 records and more than 6 Boolean attributes will perform poorly
  - Can easily overfi t the data
- Naïve Distribution Estimator
  - Can model only very boring distributions
  - Given 100 records and 10,000 multivalued attributes will be fi ne
  - Quite robust to overfi tting

So far we have two simple density estimators, in other lectures we'll see vastly more impressive ones (Mixture Models, Bayesian Networks, Density Trees, Kernel Densities and many more).