Probability overview

Probabilistic Graphical Models

Jerónimo Hernández-González

What does all this have to do with function approximation?

instead of $F : X \to Y$,

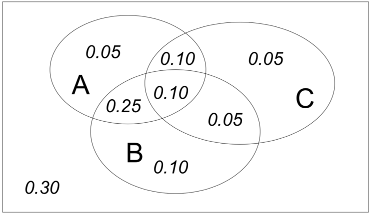
learn $P(Y|X)$

Joint distribution

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all possible combinations of values
   (M variables $\rightarrow 2^M$ combinations)
2. Say how probable each combination is
   Subscribed to the axioms of probability if sum to 1

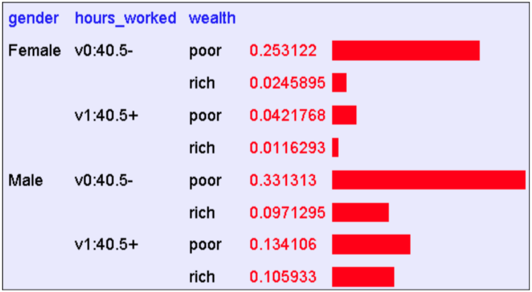
A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10



Inference with the joint distribution

You can ask for the probability of any logical expression involving your attribute

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



gender	hours_worked	wealth		
Female	v0:40.5-	poor	0.253122	
		rich	0.0245895	
	v1:40.5+	poor	0.0421768	
		rich	0.0116293	
Male	v0:40.5-	poor	0.331313	
		rich	0.0971295	
	v1:40.5+	poor	0.134106	
		rich	0.105933	

Inference with the joint distribution

You can ask for the probability of any logical expression involving a subset of attributes given another expression involving other attributes

$$P(E_1|E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)}$$

$$= \frac{\sum_{\text{r: rows matching } E_1 \text{ \& } E_2} P(r)}{\sum_{\text{o: rows matching } E_2} P(o)}$$

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Learning and the joint distribution

Suppose we want to learn the function $f :< G, H >\rightarrow W$

Equivalently, $P(W|G, H)$

Solution:

- Learn joint distribution from train data
- Calculate $P(W|G, H)$ for test data

  E.g., given a *female* patient of 39 years old:

  $\arg\max_{w \in \{rich, poor\}} P(W = w|G = female, H = 40{,}5-)$

Solution?

$P(Y|X)$ sounds like a nice alternative solution to function $F : X \to Y$

Are we done?

Main problem

Learning $P(Y|X)$ may require more data than we have

E.g., consider learning the joint distribution for 100 binary variables

- # of rows in this table?
- # of data samples to learn faithfully?
- # of rows never observed?

Facing practical problems

What to do?

1. Be smart about how to represent joint distributions
   - Bayesian networks, probabilistic graphical models
2. Be smart about how we estimate probabilities from *sparse* data
   - maximum likelihood estimates
   - maximum a posteriori estimates

Facing practical problems

What to do?

1. **Be smart about how to represent joint distributions**
   - Bayesian networks, probabilistic graphical models
2. Be smart about how we estimate probabilities from *sparse* data
   - maximum likelihood estimates
   - maximum a posteriori estimates

Conditional independence

A *qualitative* relationship between random variables

Let A, B, C be disjoint subsets of $V = \{1, ..., v\}$. We say that $\boldsymbol{X}_A$ is independent from $\boldsymbol{X}_B$ given $\boldsymbol{X}_C$ if and only if for all $(\boldsymbol{x}_A, \boldsymbol{x}_B, \boldsymbol{x}_C)$ we have that $p(\boldsymbol{x}_A|\boldsymbol{x}_B, \boldsymbol{x}_C) = p(\boldsymbol{x}_A|\boldsymbol{x}_C)$.

- Denoted by $X_A \perp\!\!\!\perp X_B|X_C$
- $p(\boldsymbol{x}_A|\boldsymbol{x}_B, \boldsymbol{x}_C) = p(\boldsymbol{x}_A|\boldsymbol{x}_C)$: Knowing/observing/fixing $\boldsymbol{x}_C$, the value $\boldsymbol{x}_B$ does not modify the probability of $\boldsymbol{x}_A$
- Exercise: Prove that
  $X_A \perp\!\!\!\perp X_B|X_C \Rightarrow p(\boldsymbol{x}_A, \boldsymbol{x}_B|\boldsymbol{x}_C) = p(\boldsymbol{x}_A|\boldsymbol{x}_C) \cdot p(\boldsymbol{x}_B|\boldsymbol{x}_C)$

Conditional independence

- Allow to simplify the factorization given by the chain rule
- Choose an appropriate ordering that allows to apply the independence over a conditional distribution

Example

- $\boldsymbol{X} = X_1, ..., X_5$
- $3 \perp\!\!\!\perp 4 | 1, 5$
- Ordering: $1, 4, 5, 3, 2$

$$
\begin{aligned}
p(\boldsymbol{X}) &= p(\boldsymbol{X}_{1,4,5})p(X_3|\boldsymbol{X}_{1,4,5})p(X_2|\boldsymbol{X}_{1,3,4,5}) \\
&= p(\boldsymbol{X}_{1,4,5})p(X_3|\boldsymbol{X}_{1,5})p(X_2|\boldsymbol{X}_{1,3,4,5})
\end{aligned}
$$

Counting the parameters

Marginal distribution

Let A and B be a partition of V. Then,

$$\forall \boldsymbol{x}_A, \boldsymbol{x}_B, p(\boldsymbol{x}_A) = \sum_{\boldsymbol{x}_B} p(\boldsymbol{x}_A, \boldsymbol{x}_B)$$

- Number of free parameters: $\prod_{i \in A} r_i - 1 = r_A - 1$

Counting the parameters

Conditional distribution

Let A and B be a partition of V

$$\forall \boldsymbol{x}_A, p(\boldsymbol{x}_A | \boldsymbol{x}_B) = \frac{p(\boldsymbol{x}_A, \boldsymbol{x}_B)}{p(\boldsymbol{x}_B)} = \frac{p(\boldsymbol{x}_A, \boldsymbol{x}_B)}{\sum_{\boldsymbol{x}_A} p(\boldsymbol{x}_A, \boldsymbol{x}_B)}$$

- Family of distribution:
  A marginal distribution for each value assignment $\boldsymbol{X}_B = \boldsymbol{x}_B$
  $\sum_{\boldsymbol{x}_A} p(\boldsymbol{x}_A | \boldsymbol{x}_B) = 1$
- Number of free parameters:
  $(\prod_{i \in A} r_i - 1) \cdot (\prod_{j \in B} r_j) = (r_A - 1) \cdot r_B$

Counting the parameters

Conditional independences reduce the number of (free) parameters

- $X_A \perp\!\!\!\perp X_B$
  - $\forall \boldsymbol{x}$: $p(\boldsymbol{x}_A, \boldsymbol{x}_B) = p(\boldsymbol{x}_A)p(\boldsymbol{x}_B)$
  - From $r_A \cdot r_B - 1$ to $r_A - 1 + r_B - 1$

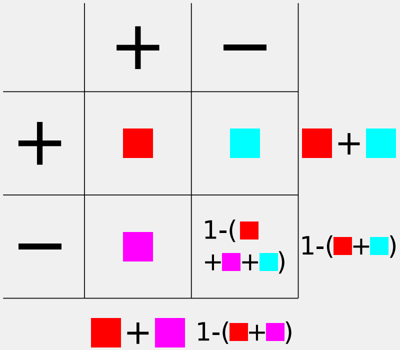
    E.g., $r_A = 3$, $r_B = 4$
    From $3 \times 4 - 1$ to $2 + 3$

- $X_A \perp\!\!\!\perp X_B | X_C$
  - $\forall \boldsymbol{x}$: $p(\boldsymbol{x}_A, \boldsymbol{x}_B | \boldsymbol{x}_C) = p(\boldsymbol{x}_A | \boldsymbol{x}_C)p(\boldsymbol{x}_B | \boldsymbol{x}_C)$
  - From $(r_A \cdot r_B - 1) \cdot r_C$ to $(r_A - 1 + r_B - 1) \cdot r_C$

    E.g., $r_A = 3$, $r_B = 4$, $r_C = 3$
    From $(3 \times 4 - 1) \times 3$ to $(2 + 3) * 3$

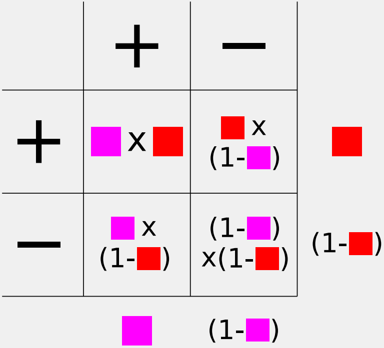
Counting the parameters, without independence

$p(A, B)$: three parameters



Counting the parameters, with independence

$p(A, B)$: two parameters



Conditional independence

Discarding parameters

The complexity of a statistical model can be understood as its flexibility for learning or its requirement of memory

Conditional independences:

- Reduce the complexity (i.e., # parameters) of the statistical model represented by the (simplified) chain rule
- Allow for avoiding to model (irrelevant) parameters associated to soft conditional dependences
- Help to deal with the trade-off between the complexity of the statistical model and amount of train data
  This has crucial implications in statistical models, e.g., overfitting

Exercise

Gobierno de Canarias | Portal de Noticias | Datos COVID-19

MENÚ

Actualidad sanitaria, Portada, Salud Pública, Sanidad 5 de noviembre de 2021

El 84 por ciento de la población canaria de más de 12 años está ya inmunizado contra la COVID-19

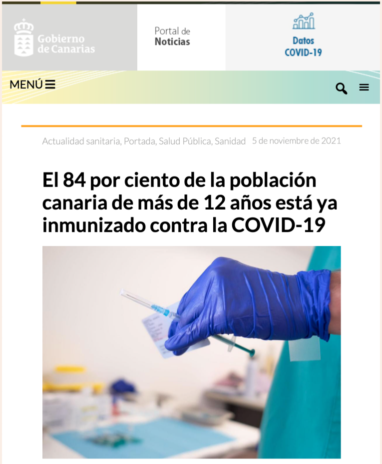
SOCIEDAD ED+

El 57% de los ingresados por coronavirus en Canarias está sin vacunar y no sufre patologías previas

Yanira Martín

- Can you tell which is the probability of ending up in the hospital if you get COVID19 and you are not vaccinated?

Exercise


Gobierno de Canarias
Portal de Noticias
Datos COVID-19
MENÚ
Actualidad sanitaria, Portada, Salud Pública, Sanidad 5 de noviembre de 2021

El 84 por ciento de la población canaria de más de 12 años está ya inmunizado contra la COVID-19


SOCIEDAD
ED+

El 57% de los ingresados por coronavirus en Canarias está sin vacunar y no sufre patologías previas

Yanira Martín

- Can you tell which is the probability of ending up in the hospital if you get COVID19 and you are not vaccinated?
- Can say anything?

Probability overview

Probabilistic Graphical Models

Jerónimo Hernández-González