Lecture 3

Cezar Ionescu

Bayesian networks

Joint probability distributions

The mathematical model of probability consists of a triple `(Ω, Event, p)` as in the previous lecture. Very frequently, however, the set of possible outcomes `Ω` has the structure of a *cartesian product*, i.e., a tuple of possible results. For instance, the diagnostic problem we discussed last time could also have been modelled by taking `Omega = (H, R)`, i.e., the set of tuples `{(healthy, ⊖), (healthy, ⊕), (ill, ⊖), (ill, ⊕)`. The event "healthy person" would then be described by `Healthy = {(healthy, ⊖), (healthy, ⊕)}`, etc.

Exercise: Define the other events of interest: "ill", "tested positive", "tested negative".

We can now construct new probability spaces on each component of `Ω`. For example:

```
pₕ : ℙ(H) → [0, 1]
pₕ (healthy) = p({(healthy, ⊖), (healthy, ⊕)})
             = p (healthy, ⊖) + p (healthy, ⊕)
```

Remark: in general, if `Ω = (Ω₁, Ω₂, ..., Ωₙ)` and `ωᵢ ∈ Ωᵢ`, we will use `p(ω_i)` to denote

```
p(ωᵢ) = p({(x₁, x₂, ..., xₙ) | x₁ ∈ Ω₁, ..., xₙ ∈ Ωₙ, xᵢ = ωᵢ})
```

Remark: recall our convention that for `x ∈ Ω` `p(x)` is an abbreviation for `p({x})`. Above, we have taken this convention one step further, by using `p(x, y)` to denote `p({(x, y)})`, and `p(x)` to denote `p({(x, y) | y ∈ Y})`. We shall continue using this form of abbreviation throughout, as is standard usage.

The "complete" probability measure `p` is called the **joint probability distribution**, the probability measures over the individual components (or over tuples of individual components) are called **marginal probability distributions**.

Remark: it is trivial to derive the marginal probability distributions, given the joint probability distribution. However, even if we have all the marginal distributions, we cannot derive the joint distribution from them.

By definition, the joint distribution is sufficient to answer any query about the probabilities of events in `Ω`. The problem is that it requires exponential storage space with respect to the number of components. For example, if Ω consists of a tuple of n boolean components, then the joint probability distribution can be represented as a table with 2^n rows.

Let $\Omega = (\Omega_1, \Omega_2, ..., \Omega_n)$ and $(\omega_1, \omega_2, ..., \omega_n) \in \Omega$. Then

```
p(ω₁, ω₂, ..., ωₙ)  =  p(ω₁ ∩  ω₂ ∩  ... ∩ ωₙ)                   -- recall convention above!
                    =  p(ω₁ | ω₂ ∩  ... ∩ ωₙ)*p(ω₂ ∩  ... ∩ ωₙ)  -- Bayes' theorem
```

It is easy to see that, therefore:

```
p(ω₁, ω₂, ..., ωₙ)  =  p(ω₁ |  ω₂ ∩  ... ∩ ωₙ) *
                       p(ω₂ | ω₃ ∩ ... ∩ ωₙ) * ...
                       p(ωₙ)
```

Independence, conditional independence

Bayesian networks

Bayesian networks are more efficient representations of joint probability distributions.

Homework: compute the following joint probability (Nilsson, 19.4, page 340).

Minimum description length (MDL) principle

```
hₘₐₚ = argmaxₕ p(d | h) * p(h)
iff
hₘₐₚ = argmaxₕ log₂ (p(d | h) * p(h))
iff
hₘₐₚ = argmaxₕ log₂ p(d | h) + log₂ p(h)
iff
hₘₐₚ = argminₕ -log₂ p(d | h) - log₂ p(h)
iff
hₘₐₚ = argminₕ length(encode(d | h)) + length(encode(h))
```