CS 480 – Introduction to Artificial Intelligence

Topic: Bayesian Networks Parameter Estimation

Mustafa Bilgic

http://www.cs.iit.edu/~mbilgic

https://twitter.com/bilgicm

Bayesian network parameter estimation

- Given:
  - A set of random variables, V_i
    - E.g., age, gender, cholesterol level, etc.
  - A Bayesian network structure over these variables
    - E.g., a doctor can point out the most important correlations and causations
  - Data
    - E.g., existing patient records, where some or all V_i are known
- Goal:
  - Estimate the parameters needed for the Bayesian network, i.e., $P(V_i \mid \text{parentsOf}(V_i))$

KNOWN BAYESIAN NETWORK STRUCTURE

- In this class, we assume the structure is given
- How reasonable is this assumption?
  - In some domains, the expert might provide a reasonable structure to start with
- There are many methods that learn the structure of the Bayesian network from data
  - Those topics are covered in the CS583 – Probabilistic Graphical Models course in detail

Parameter estimation for BNs

- Assume the network structure is given over variables V_i
- Let d_j be a fully observed instance
  - $d_j = \langle V_1\text{=t}, V_2\text{=f}, \ldots, V_n\text{=t} \rangle$
- The data $\mathcal{D}$ consists of fully observed instances
  - $\mathcal{D} = \{d_1, d_2, \ldots, d_m\}$
- Estimate the network parameters $\mathrm{P}(V_i \mid \text{parents}(V_i))$
- Two approaches
  1. Maximum likelihood estimation
  2. Bayesian estimation

SIMPLEST CASE – ONE VARIABLE

- Imagine we have a thumbtack
- Flip it, and it comes as heads or tails

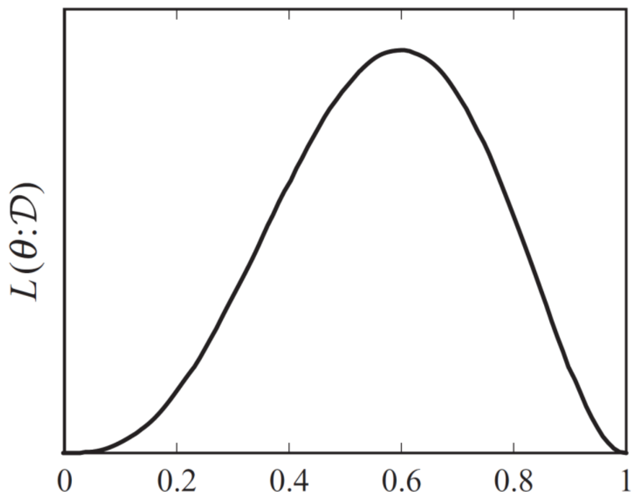
- P(Heads) = θ, P(Tails) = 1- θ
- Assume we flip it 100 times and it comes head 30 times
- What is θ?

THUMBTACK TOSSES

- Assume we have a set of thumbtack tosses
  - $\mathcal{D} = \{d_1, d_2, \ldots, d_{100}\}$
- Assume we have 30 heads and 70 tails
- P(Heads) = θ, P(Tails) = 1- θ
- θ can be any number between 0 and 1
- We have an infinite number of choices
  - θ=0, …, θ=0.3, …, θ=0.5, …, θ=1
- We want to formulate an objective function f(θ: $\mathcal{D}$), where, given 30 heads and 70 tails, f(θ: $\mathcal{D}$) achieves its maximum when θ=0.3
  - Any ideas?

LIKELIHOOD

- What is the probability, or *likelihood*, of seeing the sequence H, T, T, H, H?
  - $\theta * (1-\theta) * (1-\theta) * \theta * \theta = \theta^3 (1-\theta)^2$



When is $L(\theta : \mathcal{D})$ maximum?

Likelihood/Log-Likelihood

- Number of heads = k, number of tails = m-k
- Likelihood: $L(\theta:\mathcal{D}) = \theta^k(1-\theta)^{m-k}$
- Log-likelihood: $l(\theta:\mathcal{D}) = k\log\theta+(m-k)\log(1-\theta)$
- Note that $L(\theta:\mathcal{D})$ achieves its maximum for θ that maximizes $l(\theta:\mathcal{D})$
- Find θ that maximizes the log-likelihood
- Take derivate of $l(\theta:\mathcal{D})$ w.r.t. θ and set it to zero

LET'S SEE A FEW EXAMPLES

- Simple structure
  - $X \rightarrow Y$
- General structure
  - The key is that the parameters for each variable can be optimized independently
  - Examples

Bayesian estimation

- Assume we flip a coin 10 times and we get 4 Heads, 6 Tails
  - What is $P(C=H)$?
- What if we repeat the flips 10M times and we get 4M Heads and 6M Tails?
- Bayesian estimation will let us encode our *prior knowledge*

To cut it short, (I mean really short)

- We'll encode our prior knowledge as a set of "imaginary" counts
- For example, we will assume that we have already seen α heads and β tails
- Assume we flip a coin 10 times and we get 4 Heads, 6 Tails
  - $P(C=heads) = (4 + \alpha) / (10 + \alpha + \beta)$
  - $\alpha = 0, \beta = 0$; $P(C=h) = 4/10 = 0.4$
  - $\alpha = 1, \beta = 1$; $P(C=h) = 5/12 = 0.417$
  - $\alpha = 10, \beta = 10$; $P(C=h) = 14/30 = 0.467$
  - $\alpha = 100, \beta = 100$; $P(C=h) = 104/210 = 0.495$
- Assume we flip a coin 1000 times and we get 400 Heads, 600 Tails
  - $P(C=heads) = (400 + \alpha) / (1000 + \alpha + \beta)$
  - $\alpha = 0, \beta = 0$; $P(C=h) = 400/1000 = 0.4$
  - $\alpha = 1, \beta = 1$; $P(C=h) = 401/1002 = 0.4002$
  - $\alpha = 10, \beta = 10$; $P(C=h) = 410/1020 = 0.402$
  - $\alpha = 100, \beta = 100$; $P(C=h) = 500/1200 = 0.417$

Imaginary counts

- Note that imaginary counts can be applied to any categorical variable, not necessarily just binary variables
- Also helps with dealing zero probabilities
- When all imaginary counts are 1, this is called Laplace smoothing
  - E.g, $\alpha = 1$, $\beta = 1$
- Let’s see some examples