Using Stochastic Methods to Setup High Precision Experiments

Kristina Veljkovic, Petar Kochovski, Vlado Stankovski[1]

[1]Faculty of Computer and Information Science, University of Ljubljana, Slovenia

- Introduction
- Model definition (Bayesian network, Markov decision process)
- Re-evaluation of the probabilities
- Ranking of the options
- Example
- Concluding remarks

- We introduce a novel approach for setting up scientific experiments that are guided by Bayesian network and Markov decision processes.
- Data analytics is experimentation-driven and puts the users' feedback at the centre of the process.
- The goal of our work is to define a probabilistic model of data analytics that helps the experimenter at each step of the experiment design.

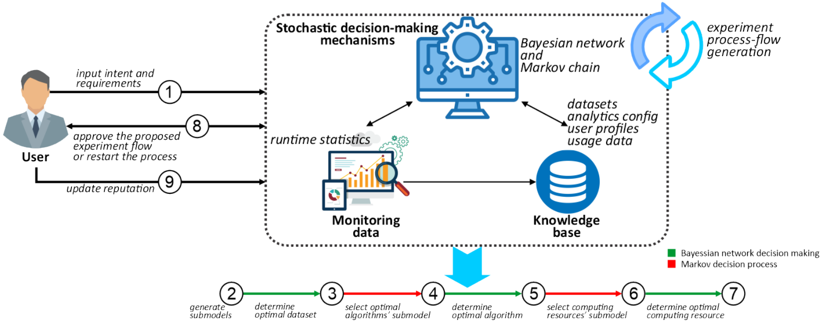


FIGURE: Outline of the baseline scenario.

CAUSAL DEPENDENCIES BETWEEN DATASETS, METHODS AND DEPLOYMENT

$$\underset{\text{Datasets } D}{\boxed{d_1, d_2, \ldots d_b}} \xrightarrow{P(m_j|d_i)} \underset{\text{Methods } M}{\boxed{m_1, m_2 \ldots m_c}} \xrightarrow{P(o_k|d_i m_j)} \underset{\text{Deployment } O}{\boxed{o_1, o_2, \ldots o_d}}$$

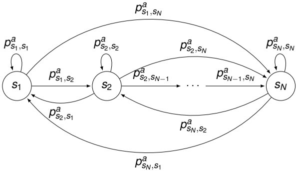
- Set of possible methods corresponds to given intent and satisfies the hard constraints for methods, while set of possible deployment options satisfies the hard constraints for deployment.
- Initial probabilities

$$p_i = P(D = d_i) = \frac{1}{b}, \quad 1 \leq i \leq b,$$

$$q_j = P(M = m_j|D = d_i) = \frac{1}{c}, \quad 1 \leq j \leq c$$

$$r_k = P(O = o_k|D = d_i, M = m_j) = \frac{1}{d}, \quad 1 \leq k \leq d.$$

INTERNAL DEPENDENCIES BETWEEN DATASETS, BETWEEN METHODS AND BETWEEN DEPLOYMENT OPTIONS



- Transition probability from state s at the step t to state s' at the step $t+1$ made due to an action a

$$p^a_{s,s'} = P(S_{t+1} = s' | S_t = s, A_t = a),\ s, s' \in S,$$

MDP	**State space $\mathbf{s_1, s_2, \dots s_N}$**	**Action**	**Initial $\mathbf{p^a_{s,s'}}$**
Datasets	$d_1, d_2, \dots d_b$	Selection of a dataset	$\frac{1}{b}$
Methods	$m_1, m_2, \dots m_c$	Selection of a method	$\frac{1}{c}$
Deployment	$o_1, o_2, \dots o_d$	Selection of a deployment	$\frac{1}{d}$

- Let $U_1, U_2, \ldots, U_q$ denote users with the expertise scores $E_1, E_2, \ldots E_q$, respectively, where $0 \leq E_l \leq 1, 1 \leq l \leq q$.

$$\hat{p}_i = \frac{1 + \sum_{l=1}^{q} E_l I_l(D = d_i)}{b + \sum_{l=1}^{q} E_l}, 1 \leq i \leq b,$$

$$\hat{q}_j = \frac{1 + \sum_{l=1}^{q} E_l I_l(M = m_j | D = d_i)}{c + \sum_{l=1}^{q} E_l}, 1 \leq j \leq c, 1 \leq i \leq b$$

$$\hat{r}_k = \frac{1 + \sum_{l=1}^{q} E_l I_l(O = o_k | D = d_i, M = m_j)}{d + \sum_{l=1}^{q} E_l},$$

$1 \leq k \leq d, 1 \leq i \leq b, 1 \leq j \leq c.$

- The transition probability $p^a_{s_i,s_j}$, $1 \le i,j \le N$ is estimated from data with
$$\hat{p}^a_{s_i,s_j} = \frac{n_{i,j}}{\sum\limits_{j=1}^{N} n_{i,j}},$$
where $n_{i,j}$ is the number of times transition from state s_i to state s_j is made and $\sum\limits_{j=1}^{N} n_{i,j}$ total number of all transitions from s_i to N states.

- Probability associated with each path of the Bayesian network is calculated as

$$P(O = o_k|D = d_i, M = m_j)P(M = m_j|D = d_i)P(D = d_i),$$
$$1 \leq i \leq b, 1 \leq j \leq c, 1 \leq k \leq d.$$

- Paths are ranked by their probabilities and the path with the highest probability is offered to a new user as the best choice.

- Utility function is defined as

$$u(s) = R_a(s, s') + \gamma \max_{a \in A} \sum_{s' \in S} P(s'|s, a)u(s'),$$

  where $R_a(s, s')$ is the expected reward received after transitioning from a state s to a state s', $P(s'|s, a)u(s')$ are the future discounted rewards and γ is a discount factor, $0 \leq \gamma \leq 1$.
- Utility function provides the ranking score for each state of Markov decision process.

- *Example in the domain of stomatology* The datasets are of the same context with the following description:
  1. the intent is described as the analysis of the effects of two factors,
  2. the relevant variables' specification as the hard constraints for methods: one dependent continuous variable and two categorical independent variables with repeated measures on one of them,
  3. the geographical location (Frankfurt) as the hard constraint for deployment.

- There are
  1. 2 datasets $\{d_1, d_2\}$,
  2. 4 mixed ANOVA methods (corresponding to the intent and satisfying the constraints for methods) $\{m_1, m_2, m_3, m_4\}$ which represent **parametric ANOVA method, non-parametric ANOVA for trimmed means, non-parametric ANOVA bootstrap t-method and non-parametric Brunner-Langer mixed ANOVA**, respectively, and
  3. 3 suitable deployment options (satisfying hard constraints for deployment) $\{o_1, o_2, o_3\}$ which represent **Google Cloud Computing deployment options n2-standard-2, n2-standard-4 and n2-standard-16**, respectively.
- There is a total of 24 possible paths of Bayesian network.

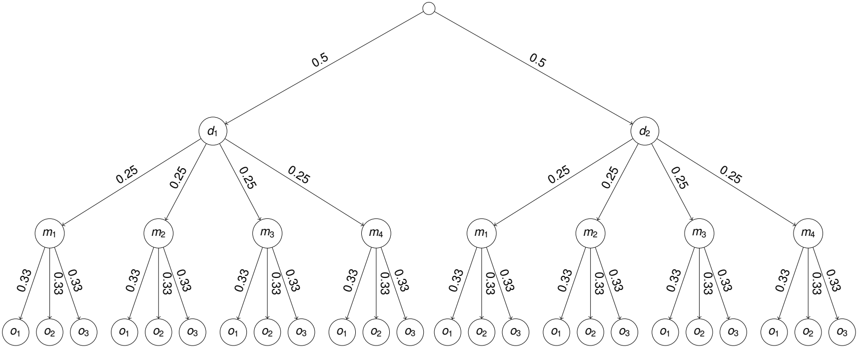


FIGURE: Initial Bayesian network

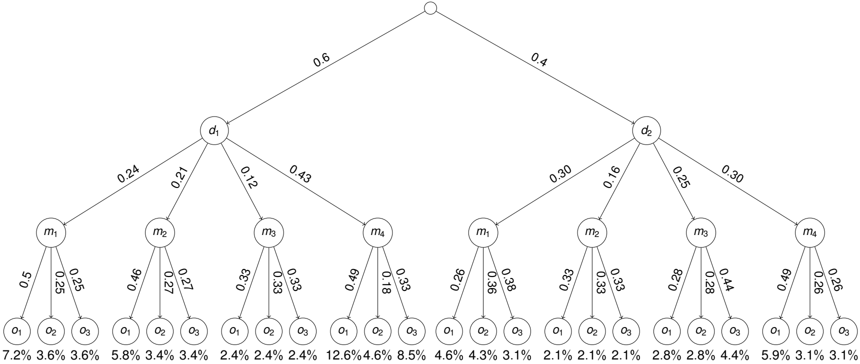


FIGURE: Bayesian network model with re-evaluated probabilities

- Markov decision process gives the insight into user's behaviour, while Bayesian network provides the best path (dataset, method and deployment option) for a given intent and hard constraints.
- Information from the Markov decision process will be used in the re-evaluation of the probabilities of Bayesian network.
- Our probabilistic model enables us to incorporate the expert's knowledge and experience into data analytics.