Lecture 8

Bayesian Network Learning

Machine Learning

Ivan Smetannikov

23.04.2021

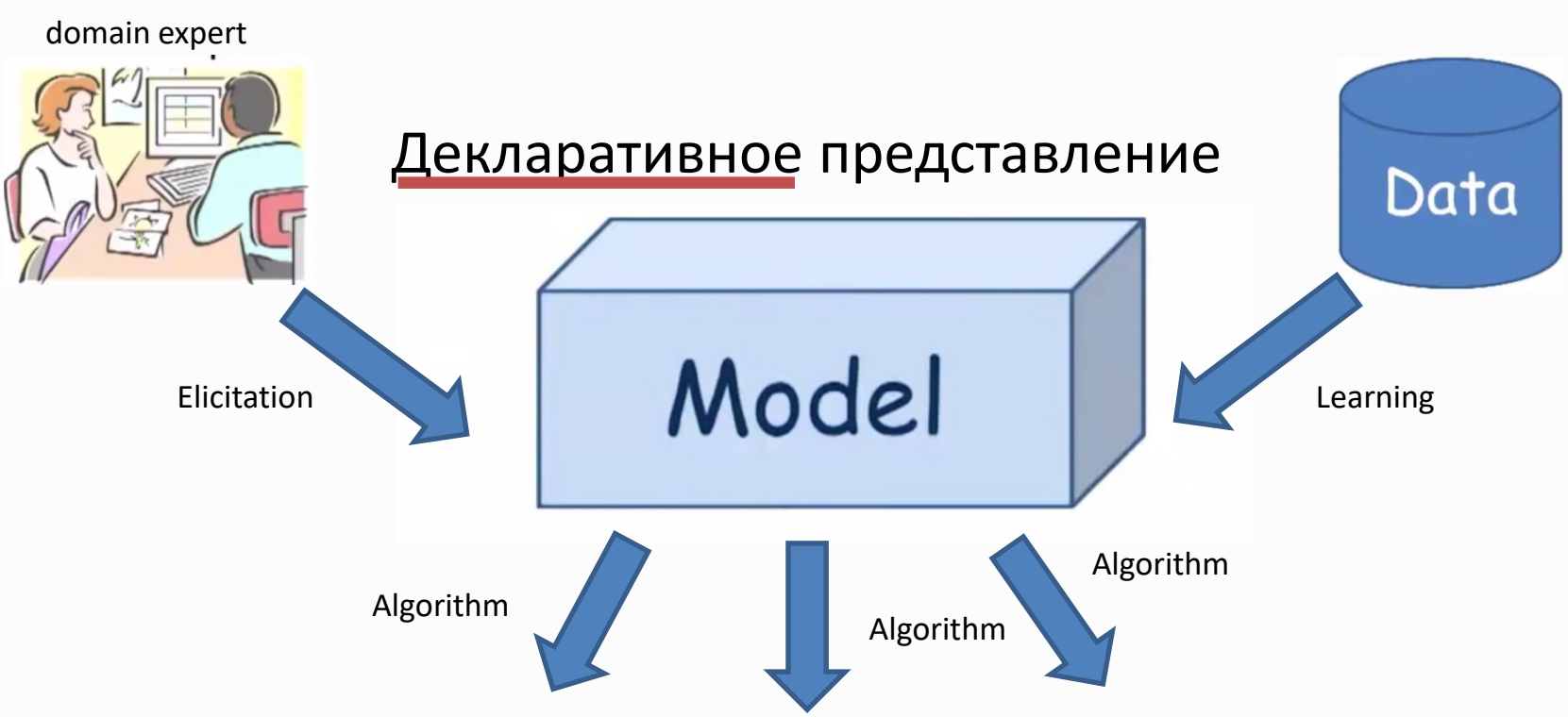
Lecture plan

- Kindly reminder
- Template models brief
- Maximum Likelihood Estimation
- Likelihood, BIC, Bayesian scores
- Learning BN structure

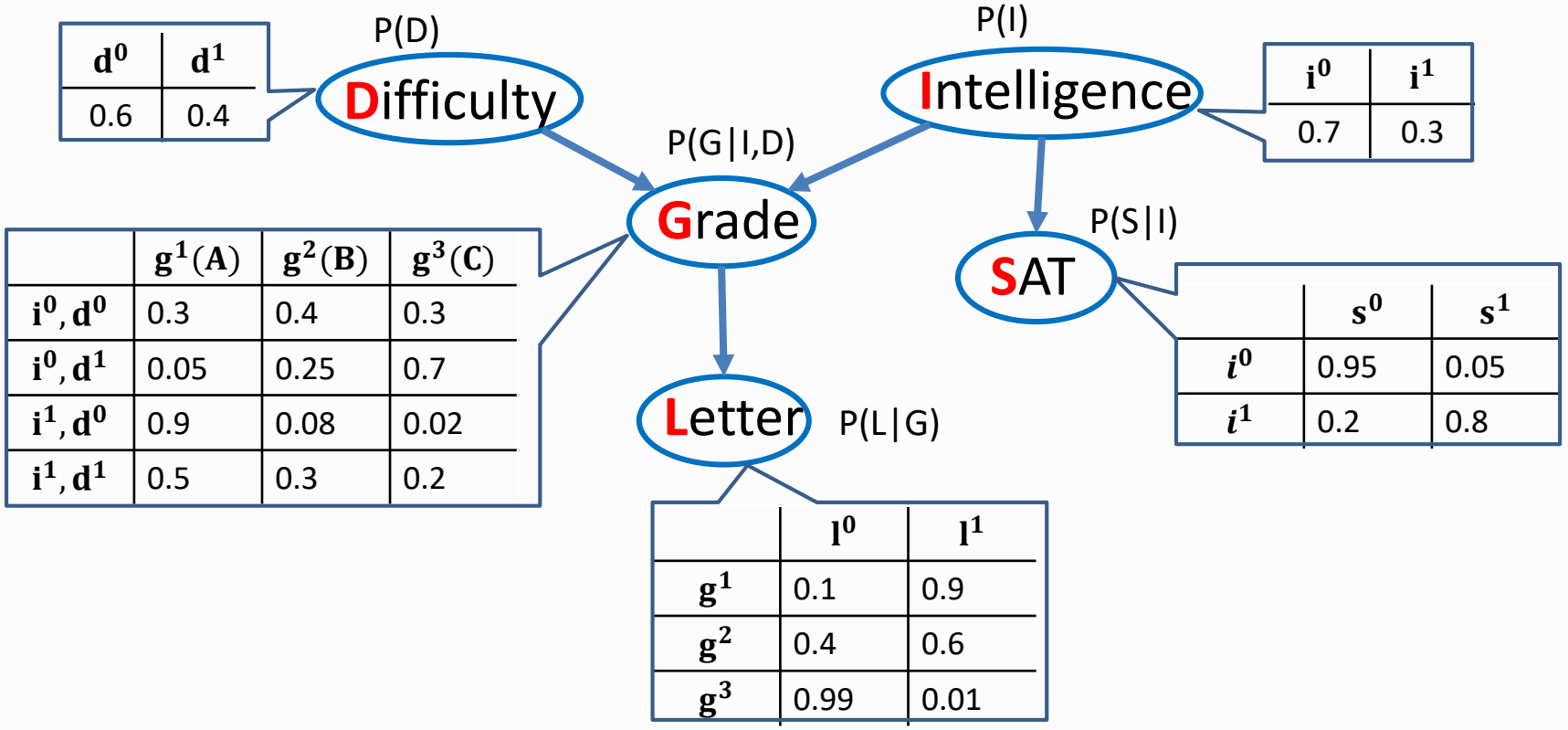
Lecture plan

- Kindly reminder
- Template models brief
- Maximum Likelihood Estimation
- Likelihood, BIC, Bayesian scores
- Learning BN structure

Модели


domain expert
Декларативное представление
Data
Model
Elicitation
Learning
Algorithm
Algorithm
Algorithm

CPD= Условное вероятностное распределение



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1(A)	g^2(B)	g^3(C)
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

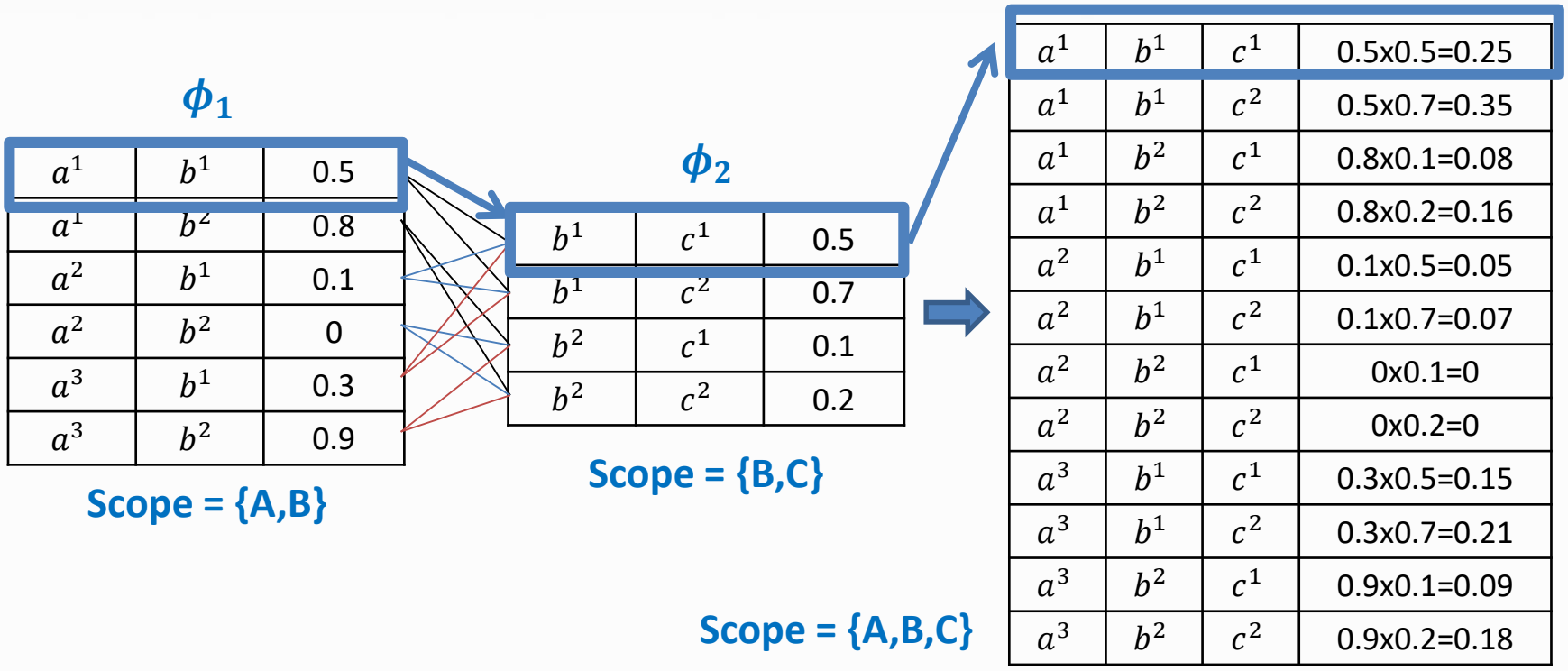
P факторизуется над G

- Пусть G граф на случайных величинах $X_1,\ldots, X_n$.
- P факторизуется над G если

$$P(X_1,\ldots, X_n) = \prod_i P(X_i | Par_G(X_i))$$

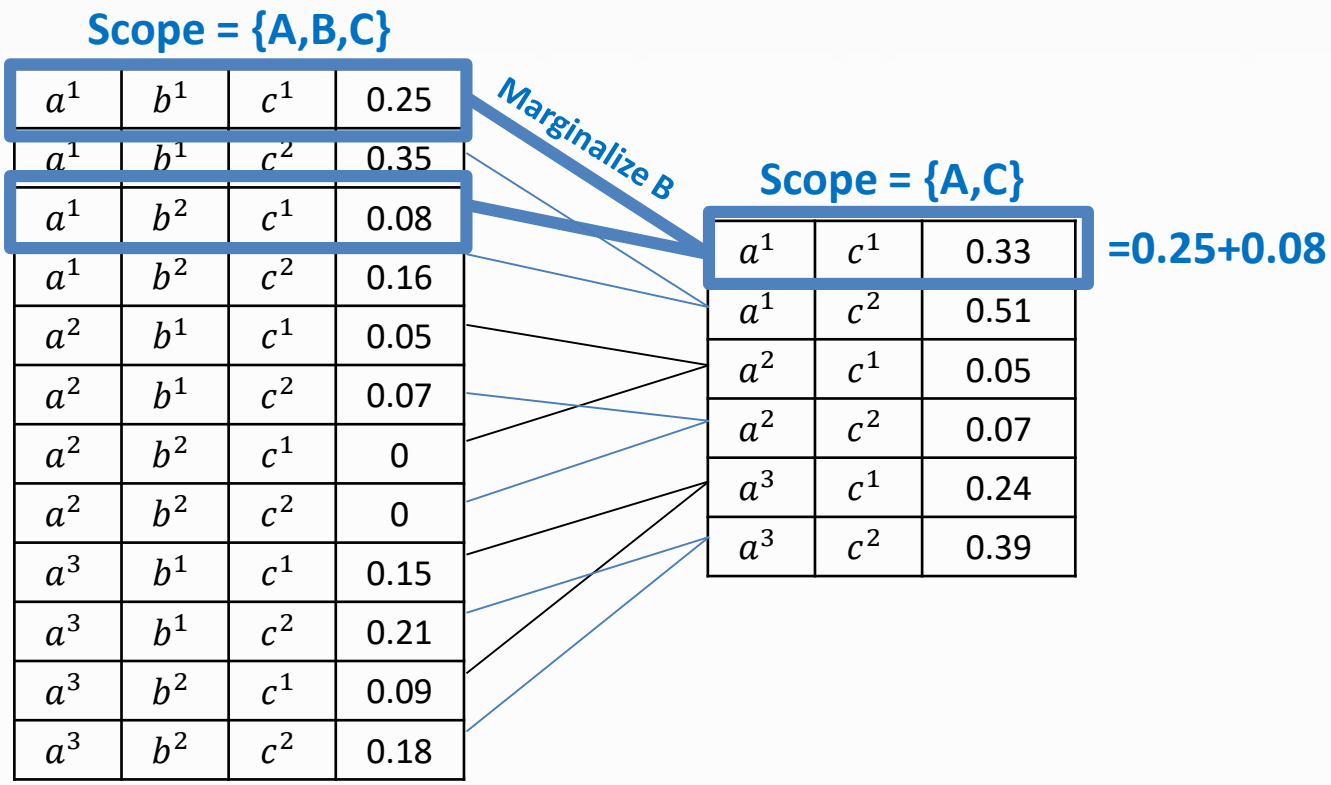
chain rule (цепное правило)

Произведение факторов



a^1	b^1	c^1	0.5x0.5=0.25
a^1	b^1	c^2	0.5x0.7=0.35
a^1	b^2	c^1	0.8x0.1=0.08
a^1	b^2	c^2	0.8x0.2=0.16
a^2	b^1	c^1	0.1x0.5=0.05
a^2	b^1	c^2	0.1x0.7=0.07
a^2	b^2	c^1	0x0.1=0
a^2	b^2	c^2	0x0.2=0
a^3	b^1	c^1	0.3x0.5=0.15
a^3	b^1	c^2	0.3x0.7=0.21
a^3	b^2	c^1	0.9x0.1=0.09
a^3	b^2	c^2	0.9x0.2=0.18

Маргинализация факторов



Редукция факторов

a^1	b^1	c^1	0.25
a^1	b^1	c^2	0.35
a^1	b^2	c^1	0.08
a^1	b^2	c^2	0.16
a^2	b^1	c^1	0.05
a^2	b^1	c^2	0.07
a^2	b^2	c^1	0
a^2	b^2	c^2	0
a^3	b^1	c^1	0.15
a^3	b^1	c^2	0.21
a^3	b^2	c^1	0.09
a^3	b^2	c^2	0.18

Reduce to the context c^1

a^1	b^1	c^1	0.25
a^1	b^2	c^1	0.08
a^2	b^1	c^1	0.05
a^2	b^2	c^1	0
a^3	b^1	c^1	0.15
a^3	b^2	c^1	0.09

Scope = {A,B}

Цепное правило это Legal Distribution: ∑ P = 1

$$\sum_{D,I,G,S,L} P(D,I,G,S,L) = \sum_{D,I,G,S,L} P(D)P(I)P(G|I,D)P(S|I)P(L|G)$$

chain rule

$$= \sum_{D,I,G,S} P(D)P(I)P(G|I,D)P(S|I) \sum_{L} P(L|G) \quad =1$$

$$= \sum_{D,I,G,S} P(D)P(I)P(G|I,D)P(S|I)$$

$$= \sum_{D,I,G} P(D)P(I)P(G|I,D) \sum_{S} P(S|I) = \sum_{D,I} P(D)P(I) \sum_{G} P(G|I,D)$$

$$= \cdots$$

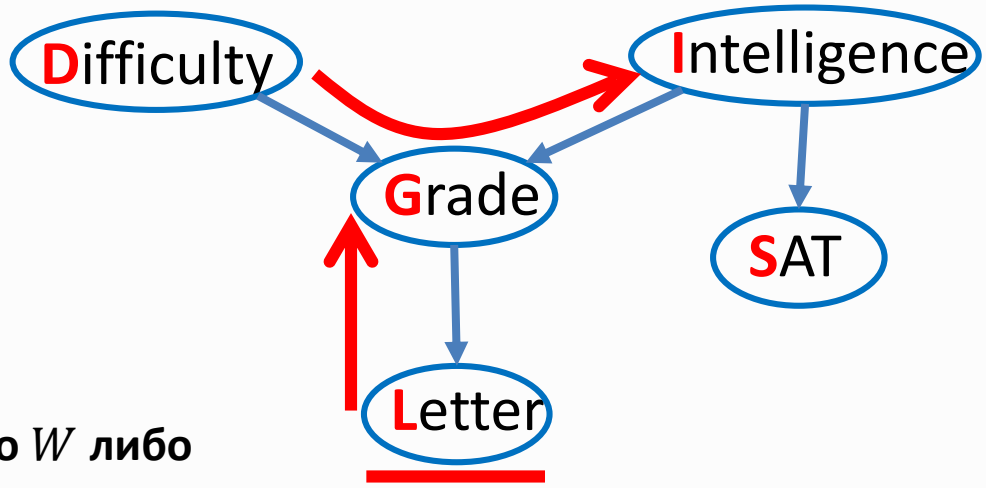
=1

Поток вероятностей: когда X влияет на Y при наличии наблюдения Z?

- $X \rightarrow Y$ ✓
- $X \leftarrow Y$ ✓
- $X \rightarrow W \rightarrow Y$ ✓ if $W \notin Z$
- $X \leftarrow W \leftarrow Y$ ✓ if $W \notin Z$
- $X \leftarrow W \rightarrow Y$ ✓ if $W \notin Z$
- $X \rightarrow W \leftarrow Y$ ✗ if $W \notin Z$, ✓ if $W \in Z$

W и все его потомки не в Z

либо W либо один из его потомков в Z



Активные пути

Путь $X_1 - \cdots - X_k$ называется активным для данного Z если:

- для любой v-structure $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$ получается что X_i или один из его потомков $\in Z$

активирует v-structure

- никакой другой X_i не находится в Z

не формирует v-structure

Lecture plan

- Kindly reminder
- Template models brief
- Maximum Likelihood Estimation
- Likelihood, BIC, Bayesian scores
- Learning BN structure

Марковское предположение

цепное правило для вероятностей

$$P(X^{(0:T)}) = P(X^{(0)}) \prod_{t=0}^{T-1} P(X^{(t+1)} | X^{(0:t)})$$

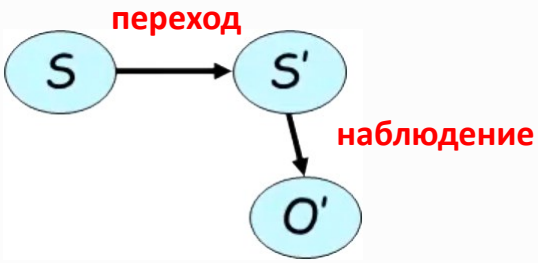
время идет вперед

состояние t+1 состояния 0…t

$$(X^{(t+1)} \perp X^{(0:t-1)} | (X^{(t)})$$

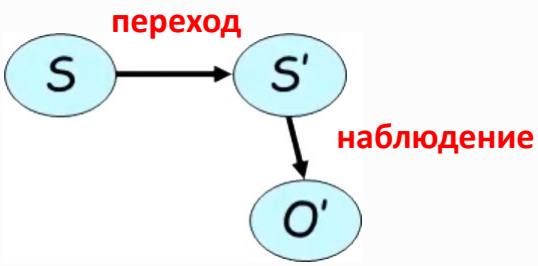
следующий шаг прошлое настоящее

Скрытые марковские модели

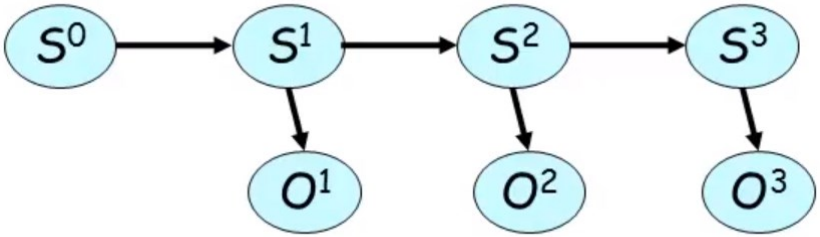


Скрытые марковские модели

2TBN

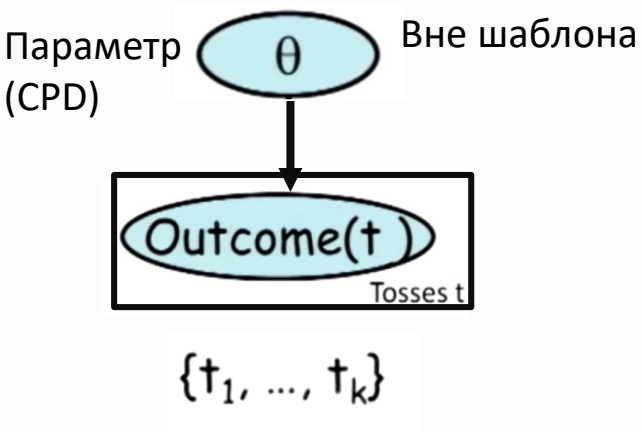


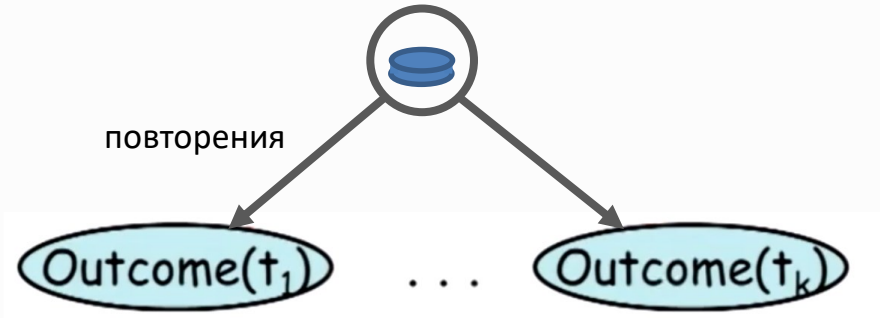
Развернутая сеть



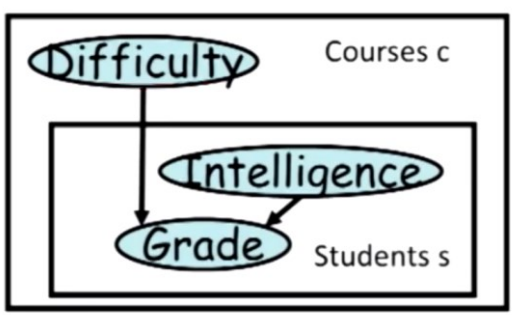
Повторение моделей

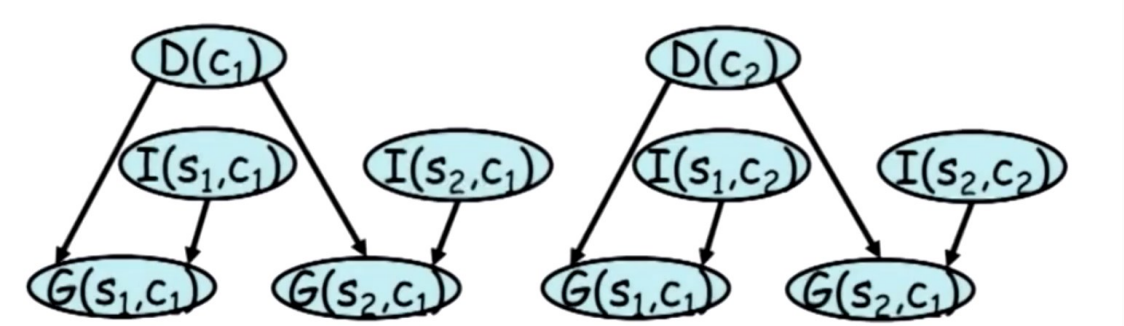
около t нет индекса = одинаковое для любого t





Nested Plates





Lecture plan

- Kindly reminder
- Template models brief
- Maximum Likelihood Estimation
- Likelihood, BIC, Bayesian scores
- Learning BN structure

Пример со смещенной монетой

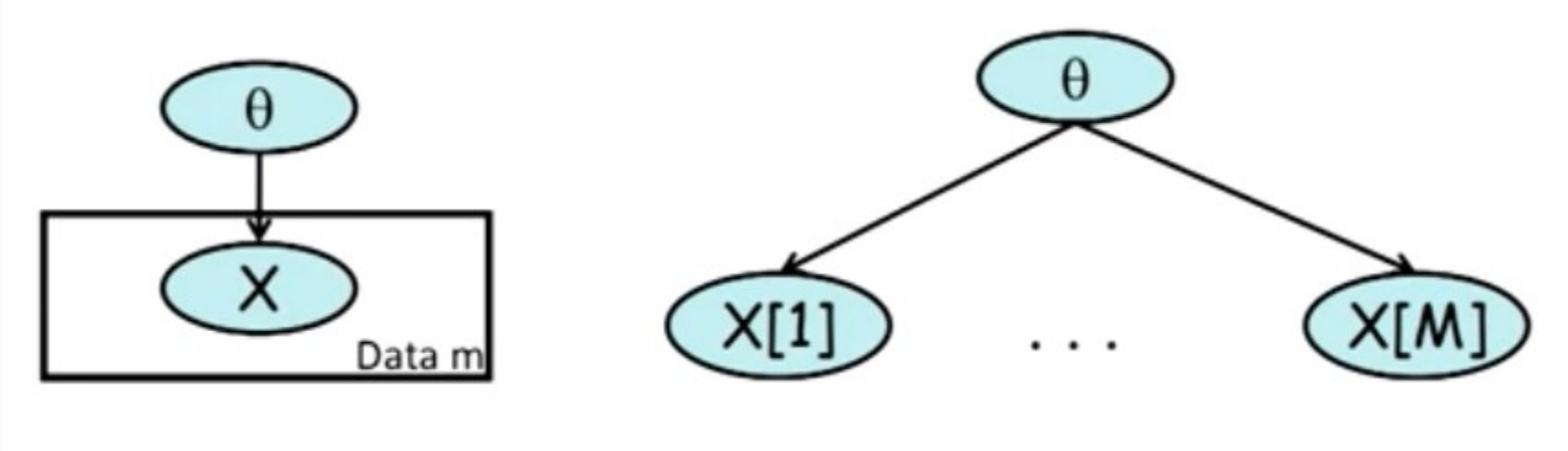
G это распределение Бернулли:

$$P(X = 1) = \theta, P(X = 0) = 1 - \theta$$

$D = \{x[1], \ldots, x[M]\}$ это IID (independent identically distributed) семплы из P:

- Броски независимы
- Броски из одного распределения

IID в виде PGM (Вероятностной графической модели)



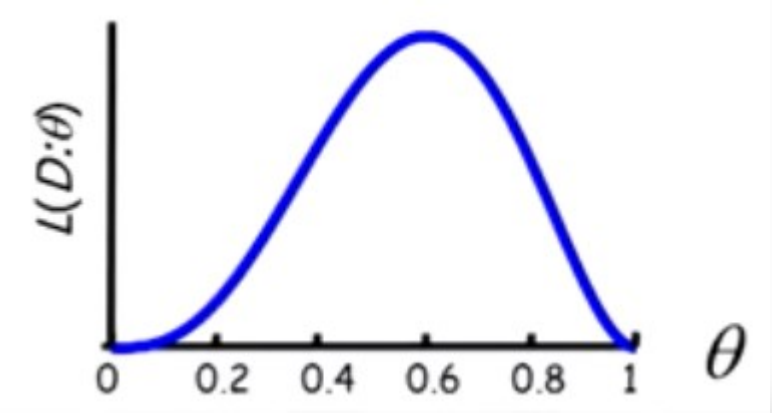
Оценка максимального правдоподобия (MLE)

Цель: найти $\theta \in [0,1]$ которая хорошо предсказывает D

Качество предсказания: правдоподобие D при известном θ

$$L(\theta : D) = P(D|\theta) = \prod_{m=1}^{M} P(x[m]|\theta)$$

$$L(\theta : \langle H, T, T, H, H \rangle)$$



Maximum Likelihood Estimator

- Наблюдения: M_H орлов и M_T решек
- Ищем θ максимизируя правдоподобие
$$\mathrm{L}(\theta: M_H, M_T) = \theta^{M_H}(1-\theta)^{M_T}$$
- Сводим к максимизации log-likelihood
$$l(\theta: M_H, M_T) = M_H \log\theta + M_T \log(1-\theta)$$
- Дифференцируем log-likelihood и решаем для заданного θ:
$$\hat{\theta} = \frac{M_H}{M_H + M_T}$$

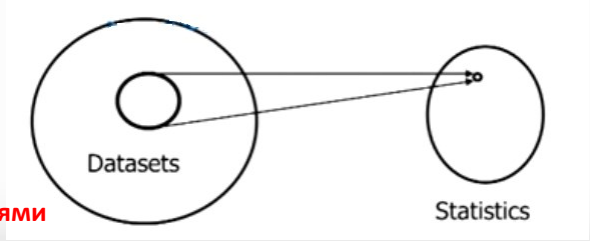
Sufficient Statistics

- В примере с монетой для вычислений нужны только M_H и M_T
- Функция s(D) называется sufficient statistic из экземпляров в вектор R^k если для любых двух наборов данных D и D` и любой $\theta \in \Theta$ выполняется условие:

S(D) = S(D`)

Если $\sum_{x[i] \in D} s(x[i]) = \sum_{x[i] \in D`} s(x[i])$ то $L(\theta : D) = L(\theta : D`)$

Набор данных с несколькими последовательностями бросков



Sufficient Statistic для Мультинома

- Для набора данных D на переменной *X* с *k* возможными значениями, sufficient statistics это просто счетчики $\langle M_1, \ldots, M_k \rangle$, где M_i число раз, когда $\mathrm{X}[m] = x^i$

- Sufficient statistic $s(x)$ это *k*-мерный tuple где

$$s(x^i) = (0, \ldots, 0, \underset{i}{1}, 0, \ldots, 0), \sum s(x[m]) = \langle M_1, \ldots, M_k \rangle$$

$$L(\theta : D) = \prod_{i=1}^{k} \theta_i^{M_i}$$

Maximum Likelihood Estimation

Принцип MLE в общем случае: выбираем θ для максимизации $L(D:\Theta)$

MLE для байесовских сетей

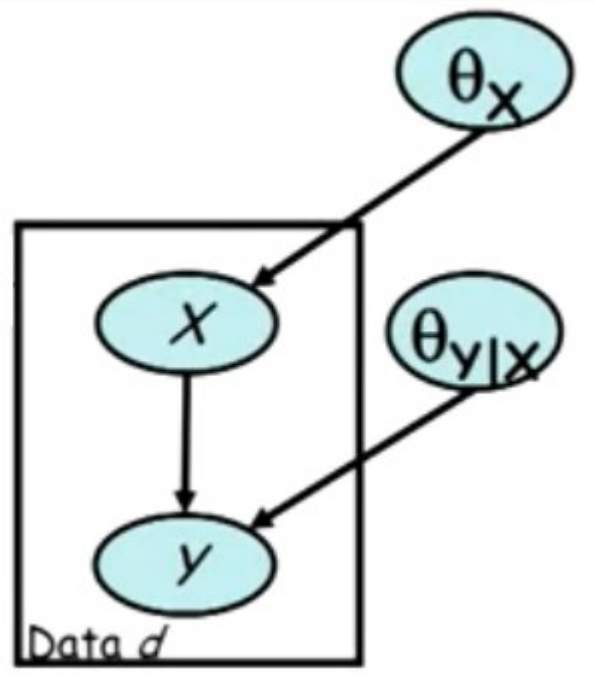
- Параметры: $\{\theta_x : x \in val(X)\}, \{\theta_{y|x} : x \in val(X), y \in val(Y)\}$

$$\mathrm{L}(\Theta : D) = \prod_{m=1}^{M} P(x[m], y[m] : \theta)$$

$$= \prod_{m=1}^{M} P(x[m] : \theta) P(y[m] | x[m] : \theta)$$

$$= \left(\prod_{m=1}^{M} P(x[m] : \theta) \right) \left(\prod_{m=1}^{M} P(y[m] | x[m] : \theta) \right)$$

$$= \left(\prod_{m=1}^{M} P(x[m] : \theta_X) \right) \left(\prod_{m=1}^{M} P(y[m] | x[m] : \theta_{Y|X}) \right)$$



Local likelihood

MLE для байесовских сетей

- Likelihood для Байесовской сети более общего вида

$$\mathrm{L}(\Theta : D) = \prod_{m=1}^{M} P(x[m] : \Theta)$$

$$= \prod_{m} \prod P(x_i[m] | U_i[m] : \Theta_i)$$

parents of x_i

chain rule

$$= \prod_{m} \prod_{i}^{m} P(x_i[m] | U_i[m] : \Theta_i) = \prod_{i} L_i(D : \Theta_{\mathrm{i}})$$

Local likelihood

MLE для табличных CPDs

$$\prod_{m=1}^{M} P(x[m]|u[m]:\theta) = \prod_{m=1}^{M} P\left(x[m]|u[m]:\theta_{X|U}\right)$$

$$= \prod_{x,u} \left(\prod_{m:x[m]=x,\, u[m]=u} P\left(x[m]|u[m]:\theta_{X|U}\right) \right)$$

$$= \prod_{x,u} \left(\prod_{m:\, x|m|=x,\, u|m|=u} \theta_{x|u} \right)$$

$$= \prod_{x,u} \theta_{x|u}^{M_{[x,u]}}$$

$$\theta_{x|u} = \frac{M[x,u]}{\sum_{x`} M[x`,u]} = \frac{M[x,u]}{M[u]}$$

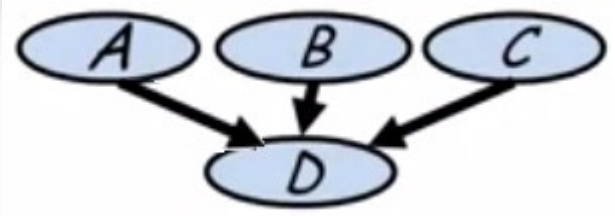
Итого

- Для Байесовской Сети с disjoint параметрами в CPD, правдоподобие декомпозируется как произведение локальных функций правдоподобия, по одной на каждую переменную
- Для табличных CPD, локальные правдоподобия можно дальше декомпозировать как произведение правдоподобий для мультиномов, одно для каждой комбинации родителей

Lecture plan

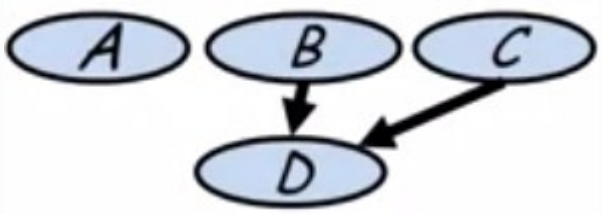
- Kindly reminder
- Template models brief
- Maximum Likelihood Estimation
- Likelihood, BIC, Bayesian scores
- Learning BN structure

Важность точного построения



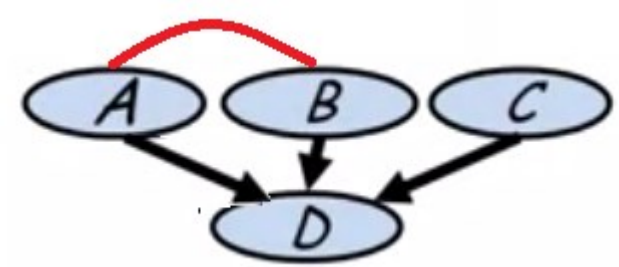
Оригинальный граф

Пропустили ребро



- Некорректные независимости
- Нельзя выучить правильное распределение P*
- Лучше обобщающая способность

Лишнее ребро



- Ложные зависимости
- Можем выучить правильное распределение P*
- Увеличенное число параметров
- Хуже обобщающая способность

Что нужно для построения?

- Метрика
- Оптимизатор

Пример

- Найти (G, θ) которая максимизирует правдоподобие

$$score_L(G:D) = l\left(\left(\hat{\theta}, G\right):D\right)$$

Пример

G_0 

$$score_L(G_0 : D) = \sum_m \left(\log \hat{\theta}_{x[m]} + \log \hat{\theta}_{y[m]}\right)$$

G_1 

$$score_L(G_1 : D) = \sum_m \left(\log \hat{\theta}_{x[m]} + \log \hat{\theta}_{y[m]|x[m]}\right)$$

$$score_L(G_1 : D) - score_L(G_0 : D) = \sum_m \left(\log \hat{\theta}_{y[m]|x[m]} - \log \hat{\theta}_{y[m]}\right)$$

$$= \sum_{x,y} M[x,y] \log \hat{\theta}_{y|x} - \sum_y M[y] \log \hat{\theta}_y$$

$$= M \sum_{x,y} \hat{P}(x,y) \log \hat{P}(y|x) - M \sum_y P(y) \log \hat{P}(y)$$

$$= M \left(\sum_{x,y} \hat{P}(x,y) \log \hat{P}(y|x) - \sum_{x,y} \hat{P}(x,y) \log \hat{P}(y) \right)$$

$$= M \left(\sum_{x,y} \hat{P}(x,y) \log \frac{\hat{P}(y,x)}{\hat{P}(x)\hat{P}(y)} \right) = M \cdot I_{\hat{P}}(X;Y)$$

Избегаем переобучения

- Ограничиваем пространство
  - Число родителей или число параметров
- Дополнительные штрафы на сложность
  - Явные
  - Bayesian score усредняется по всем возможным значениям параметров

BIC score

- Штрафуем сложность

$$score_{BIC}(G:D) = score_L(G:D) - \frac{\log M}{2} Dim[G]$$

Dim[G] = 2 в степени числа ребер

Bayesian score

Marginal likelihood Prior over structures

$$P(G:D) = \frac{P(D|G)P(G)}{P(D)}$$

Marginal probability of Data

$$score_B(G:D) = \log P(D|G) + \log P(G)$$

Marginal Likelihood

$$P(\mathcal{D} \mid \mathcal{G}) = \prod_i \prod_{\boldsymbol{u}_i \in Val(\mathbf{Pa}_{X_i}^{\mathcal{G}})} \frac{\Gamma(\alpha_{X_i|\boldsymbol{u}_i})}{\Gamma(\alpha_{X_i|\boldsymbol{u}_i} + M[\boldsymbol{u}_i])} \prod_{x_i^j \in Val(X_i)} \left[\frac{\Gamma(\alpha_{x_i^j|\boldsymbol{u}_i} + M[x_i^j, \boldsymbol{u}_i])}{\Gamma(\alpha_{x_i^j|\boldsymbol{u}_i})} \right]$$

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt \qquad \Gamma(x) = x \cdot \Gamma(x-1)$$

Lecture plan

- Kindly reminder
- Template models brief
- Maximum Likelihood Estimation
- Likelihood, BIC, Bayesian scores
- Learning BN structure

Строим деревья/леса

- Леса
  - Максимум один родитель для каждой переменной
- Почему деревья?
  - Математика
  - Эффективная оптимизация
  - Разреженная параметризация

Обучение

- p(i)=родитель X_i или 0 если у X_i нет родителей

$$\text{score}(\mathcal{G} : \mathcal{D}) = \sum_i \text{score}(X_i \mid \mathbf{Pa}_{X_i}^{\mathcal{G}} : \mathcal{D})$$

$$= \sum_{i:p(i)>0} \text{score}(X_i \mid X_{p(i)} : \mathcal{D}) + \sum_{i:p(i)=0} \text{score}(X_i : \mathcal{D})$$

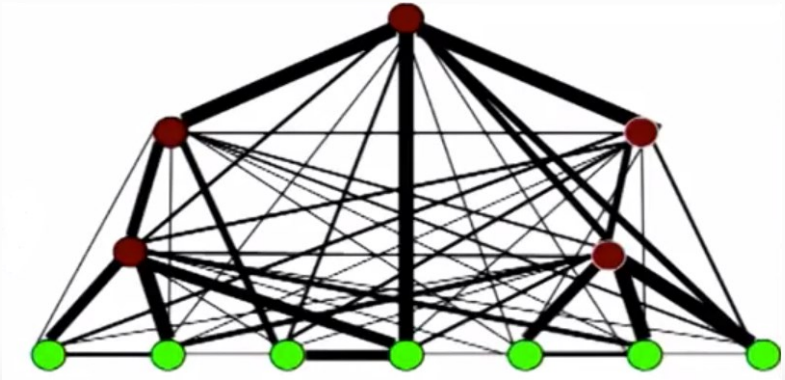
$$= \sum_{i:p(i)>0} \left(\text{score}(X_i \mid X_{p(i)} : \mathcal{D}) - \text{score}(X_i : \mathcal{D})\right) + \sum_{i=1}^{n} \text{score}(X_i : \mathcal{D})$$

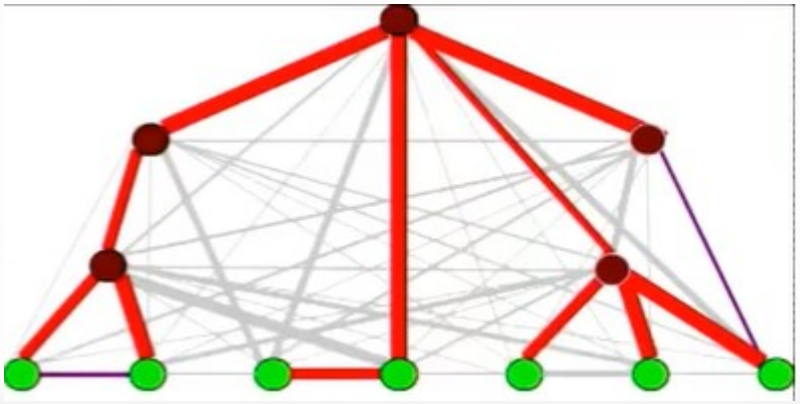
Улучшение по сравнению с пустой сетью

Score of empty network

Обучение

- Определить неориентированный граф с вершинами {1,…,n}
- Назначаем w = max(score, 0)
- Строим остовное дерево Крускалом или еще чем-то
- Удаляем нулевые ребра





Что-то кроме деревьев?

- В общем случае всё не так очевидно
  - Пример: Если позволить наличие двух родителей, то жадный алгоритм уже не гарантирует оптимальное множество
- Теорема
  - Поиск сети со структурой имеющий максимальный score для ситуации когда для каждой переменной разрешено не более *k* родителей является NP-трудной для *k>1*

Эвристический поиск

- Операторы:
  - Пошаговые: добавление, удаление, инверсия ребер
  - Глобальные шаги
- Техники:
  - Greedy hill-climbing
  - Best first search
  - Simulated Annealing
  - …