Predictive and Causal Inference of Cervical Cancer

CHAN Shun Hin (ISOM)
FONG Ki Wai (MATH)

May 30, 2021

Table of contents

Introduction

Predictions using classification models

Causal inference using Bayesian networks

Appendix: Mathematical details of Bayesian networks

Introduction

The problem

- A + test $\Rightarrow$ found (pre)cancerous cells; may need treatment
  - https://www.healthline.com/health/cervical-biopsy#results
- Can we predict biopsy result using demographic data, lifestyle, and medical history?
- Dataset: Cervical cancer dataset with patient details (2017)
  - from UCI ML Repository

The data set (2017)

From UCI ML Repository

- $N = 858$
- Originally 32 independent vars + 4 targets
- Deleted 3 targets out of concern
- Deleted 2 vars only w/ 0 values
- Deleted 2 vars w/ 91.7% missing values
- Impute the missing data by mode
- The target response: `Biopsy` (binary $\Rightarrow$ classificaton)
  - $+ : - = 55{:}803$
  - $+$ accounts for $< 7\%$ of all available cases

Variables

Attribute	Type
Age	Integer
Number of sexual partners	Integer
First sexual intercourse (age)	Integer
Num of pregnancies	Integer
Smokes	Boolean
Smokes (years)	Boolean
Smokes (packs/year)	Boolean
Hormonal Contraceptives	Boolean
Hormonal Contraceptives (years)	Integer
IUD	Boolean
IUD (years)	Integer
STDs	Boolean
STDs (number)	Integer
STDs:condylomatosis	Boolean
STDs:cervical condylomatosis	Boolean
STDs:vaginal condylomatosis	Boolean
STDs:vulvo-perineal condylomatosis	Boolean
STDs:syphilis	Boolean

Attribute	Type
STDs:pelvic inflammatory disease	Boolean
STDs:genital herpes	Boolean
STDs:molluscum contagiosum	Boolean
STDs:AIDS	Boolean
STDs:HIV	Boolean
STDs:Hepatitis B	Boolean
STDs:HPV	Boolean
STDs: Number of diagnosis	Integer
STDs: Time since first diagnosis	Integer
STDs: Time since last diagnosis	Integer
Dx:Cancer	Boolean
Dx:CIN	Boolean
Dx:HPV	Boolean
Dx	Boolean
Hinselmann: target variable	Boolean
Schiller: target variable	Boolean
Cytology: target variable	Boolean
Biopsy: target variable	Boolean

Tasks

(1) Prediction
- accurate prediction save biopsy cost (insurance companies like it)

(2) Further: Causal inference.
- More interesting
- Not only of interpreting Biopsy. e.g. IUD (contraceptive device) vs Dx.Cancer
- Figure out potential risk factors $\Longrightarrow$ prevention!

Do what?

- Yuki: Predictions using different classification models
- Lupe: Predictions and causal inference using Bayesian networks

Predictions using classification models

About the classification models

Before feeding these models with the training set ($N = 686$)

- Standardize the data according to the training set
- maybe over-/under-sampling a bit

Over-/under-sampling

We use

- SMOTE – over-sample the + cases
- Near Miss – under-sample the – cases

For the classification models, we tried

- SMOTE + cases to have the same count as – cases
- Near Miss the other way
- SMOTE + cases & Near Miss – cases to their geometric mean
- Nothing only for the random forest

Existing (classification) models

To compare the results with Al-Wesabi et al's (2018), we feed the over-/under-sampled training set into

- Gaussian Naive Bayes
- Decision Tree
- Logistic regression
- SVM
- kNN ($k = 5$)
- Random forest that Al-Wesabi et al didn't try

using Python with `scikit-learn` and `imbalanced-learn`

Metrics to measure and compare the performance

Given tp, fn, tn, fp are True $+$, False $-$, True $-$, False $+$

- Overall accuracy

$$\frac{tp + tn}{tp + fn + tn + fp} \tag{1}$$

- Weighted precision

$$\left[\frac{tp(tp + fn)}{tp + fp} + \frac{tn(tn + fp)}{tn + fn}\right] \div (tp + fn + tn + fp) \tag{2}$$

- $+$ recall

$$\frac{tp}{tp + fn} \tag{3}$$

Metrics to measure and compare the performance

Numbers

- Weighted specificity

$$\left[\frac{tp(tn+fp)}{tp+fn}+\frac{tn(tp+fn)}{tn+fp}\right]\div(tp+fn+tn+fp) \tag{4}$$

- Weighted F_1 & F_2 score ($\beta = 1, 2$)

$$\begin{aligned}&\frac{(1+\beta^2)tp}{(1+\beta^2)tp+\beta^2 fn+fp}\times\frac{tp+fn}{tp+fn+tn+fp}\\ +&\frac{(1+\beta^2)tn}{(1+\beta^2)tn+\beta^2 fp+fn}\times\frac{tn+fp}{tp+fn+tn+fp}\end{aligned} \tag{5}$$

Results on GNB, DT, LR, SVM & 5NN

	GNB	Tree	LR	SVM	kNN
Accuracy	9.4	87.0	76.8	83.3	76.9
Recall	97.2	15.9	35.4	26.1	42.8
Specificity	91.2	20.7	38.1	30.1	45.1
Precision	89.7	89.0	89.4	89.4	90.1
F_1 score	6.9	87.9	82.0	86.0	82.2
F_2 score	5.6	87.3	78.5	84.2	78.6

Table: SMOTE-ed

Results on GNB, DT, LR, SVM & 5NN

	GNB	Tree	LR	SVM	kNN
Accuracy	63.5	34.5	44.5	53.8	60.1
Recall	63.9	82.0	76.2	70.2	63.7
Specificity	63.8	78.7	74.0	69.0	63.4
Precision	90.9	90.5	90.7	90.8	90.6
F_1 score	72.7	44.7	55.8	64.4	70.0
F_2 score	65.7	35.5	46.5	56.0	62.5

Table: Near Miss-ed

Results on GNB, DT, LR, SVM & 5NN

	GNB	Tree	LR	SVM	kNN
Accuracy	70.8	60.5	51.1	58.9	59.7
Recall	56.1	57.7	68.2	63.0	67.4
Specificity	57.1	57.8	67.0	62.7	66.9
Precision	90.7	90.0	90.4	90.5	91.0
F_1 score	78.2	70.3	62.2	69.0	69.7
F_2 score	72.8	63.0	53.4	61.3	62.0

Table: SMOTE-ed and Near Miss-ed

Results on GNB, DT, LR, SVM & 5NN

After repeating the workflow from p. 10 to p. 12 for 100 times,

- GNB is better and decision tree is worse c.f. Al-Wesabi
  - esp. in SMOTE followed by Near Miss
  - from different train–test split
    - us: 80:20, Al-Wesabi et al: 88:12
  - Near Miss instead of Tomek Links

With all the above metrics, which models work the best?

- SMOTE (+ Near Miss): 5NN
- Near Miss: GNB

Random forest

Tune hyper-parameters by grid searching with stratified 7-fold CV

- look at the $+$ recall of the SMOTE-ed and Near Miss-ed standardized training set
- Highest recall $\Rightarrow$ get the parameters $\Rightarrow$ plant the forest with
  - Neither
  - SMOTE-ed
  - Near Miss-ed
  - Both

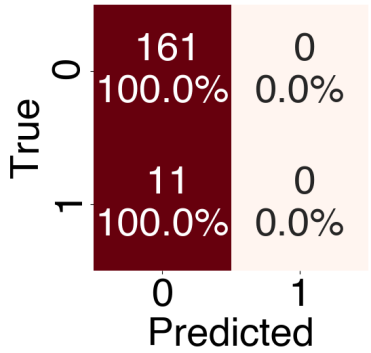
  standardized training set

$\Rightarrow$ compare the confusion matrices on the test set

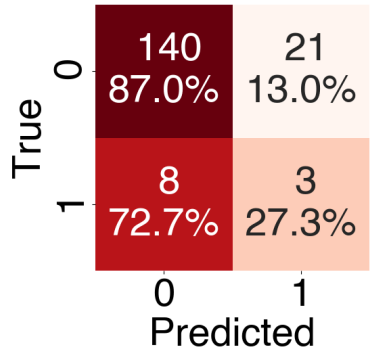
How are the fruit from the forests?

The confusion matrices with the same stdized training set w/(o) over-/under-sampling

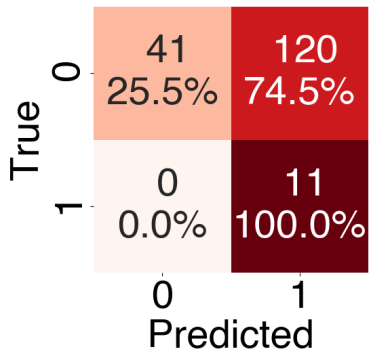
- `max_features` $= \lceil \log_2 28 \rceil = 5$
- `n_estimators` $= 10$
- `max_depth` $= 5$
- `min_samples_split` $= 2$



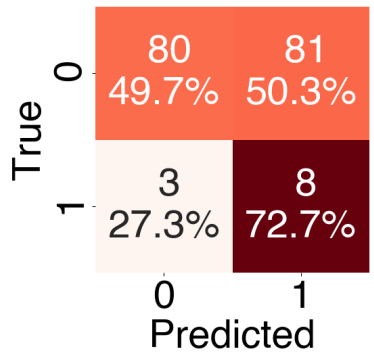
(a) Nothing



(b) SMOTE



(c) Near Miss



(d) Both

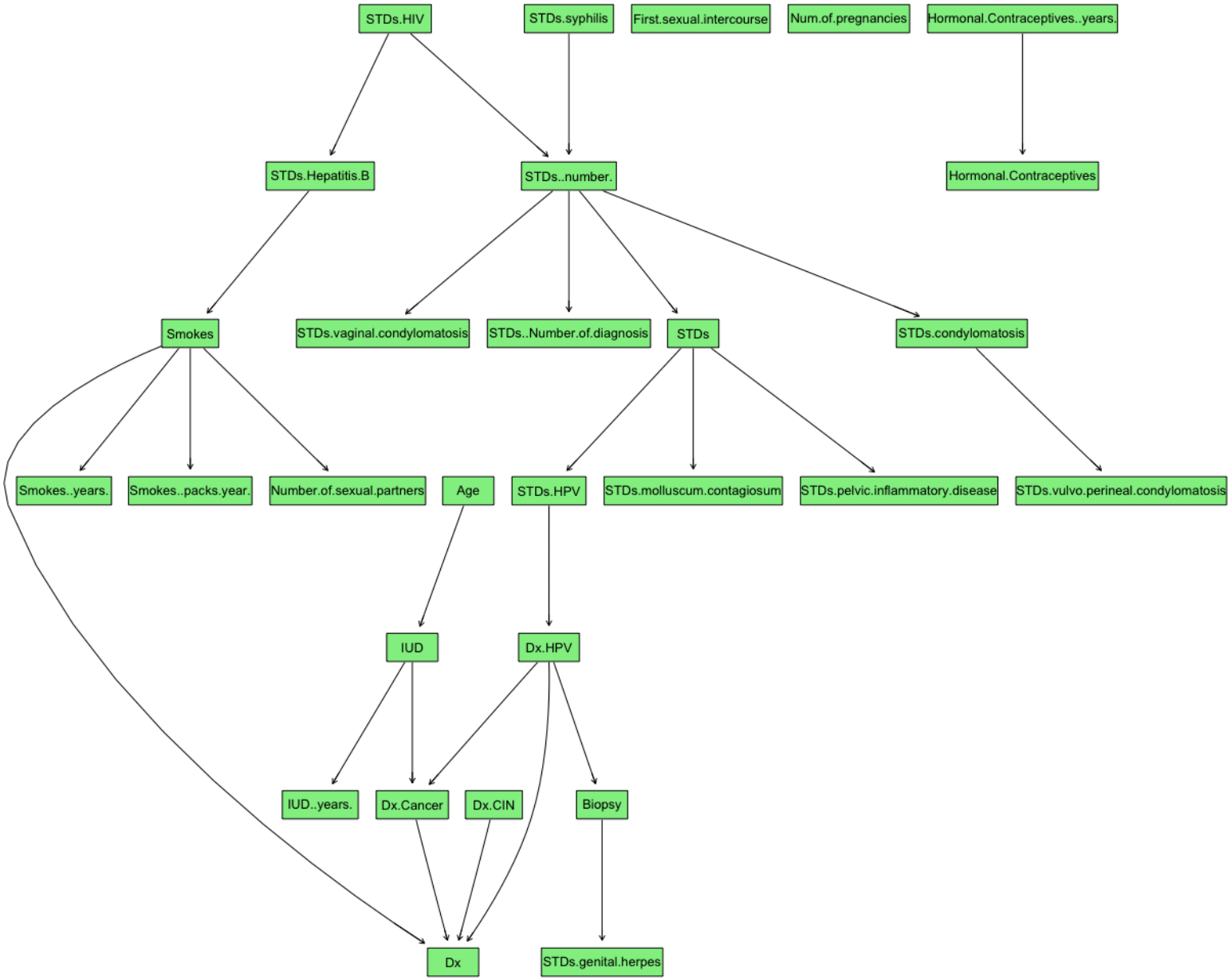
Use which to do the prediction?

Would suggest SMOTE on the standardized training set $\Rightarrow$ predict

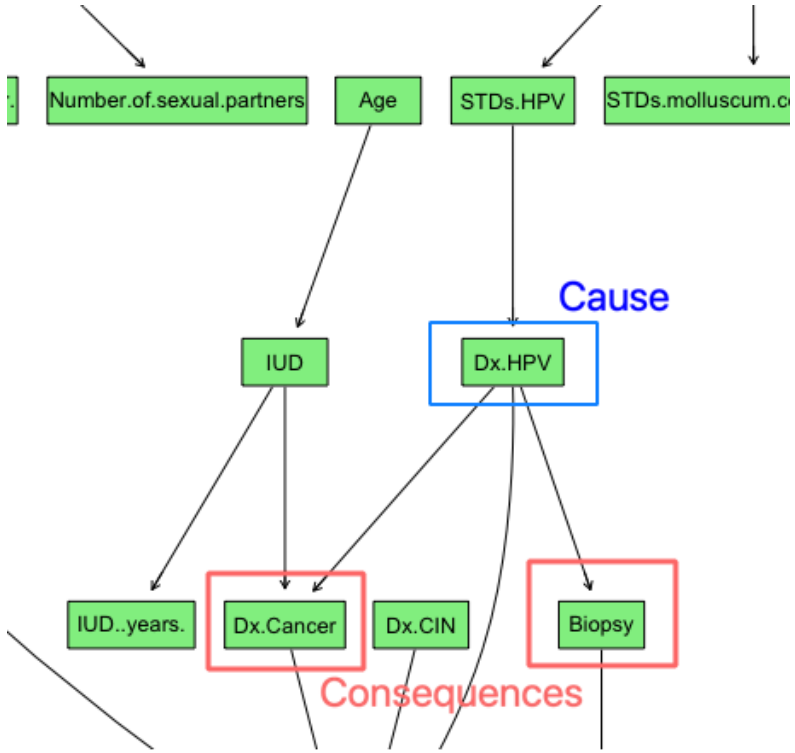
- at least identified some (not much though) $+$ cases
- the overall accuracy (83.1%) is okay
- not giving too many false alarms ($fp/(tn + fp) = 13.0\%$)

Causal inference using Bayesian networks

Bayesian networks: Introduction



Bayesian networks: Introduction



- `Dx.HPV` $\rightarrow$ `Dx.Cancer`: Probabilistic causal relationship
- Conditional independence: `Dx.Cancer` $\perp\!\!\!\perp$ `Biopsy` $\mid$ `Dx.HPV`
- Unsupervised learning usually

Bayesian networks: Definition

- $V = (X_1, \ldots, X_n)$: variable set/vertex set. x_i discrete, with supports $\{1, 2, \ldots, q_i\}$
- E: arc set; e.g. $X_1 \to X_2 \in E$
- $G = (V, E)$ defines a directed acyclic graph: directed graph without cycles

Causal inference: Overall idea

- Procedures
  - Given the dataset D, learn structure G
  - Assign posterior score $P(G|D)$ to G
  - MCMC approach: G is random. Sample $\mathcal{G} = \{G_1, G_2, \ldots, G_N\}$ based on $P(G|D)$; Sample from posterior modes
  - Estimate the edge feature, $P(X_i \to X_j|D)$ using $\mathcal{G}$ (Model averaging)
  - Estimate $P(X = x|Y = y)$ for interesting relationship, for some X and Y using $G^* = \arg\max_{G \in \mathcal{G}}\{P(G|D)\}$ (Maximum a posteriori, MAP)

Results: Prediction

- Can be used to predict; not our main goal
- Calculate $P(\texttt{Biopsy} = 1|X)$ by simulation

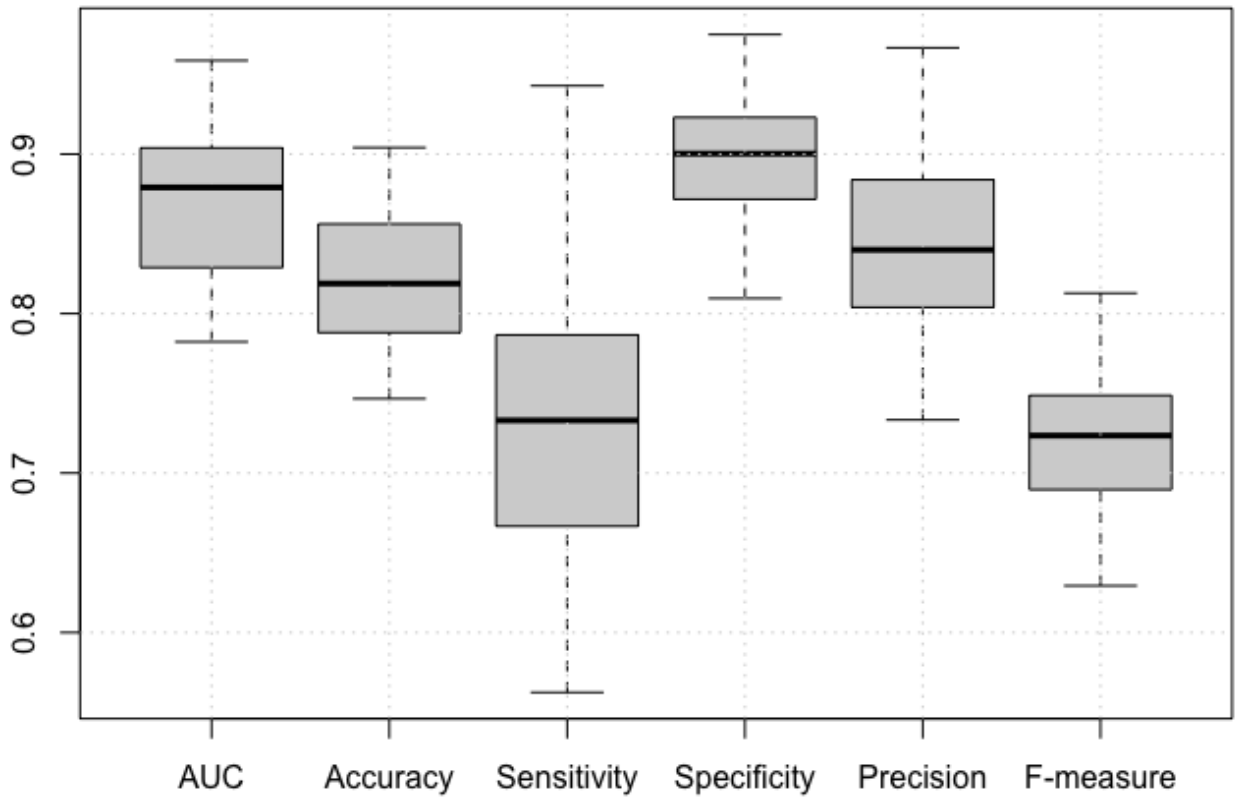
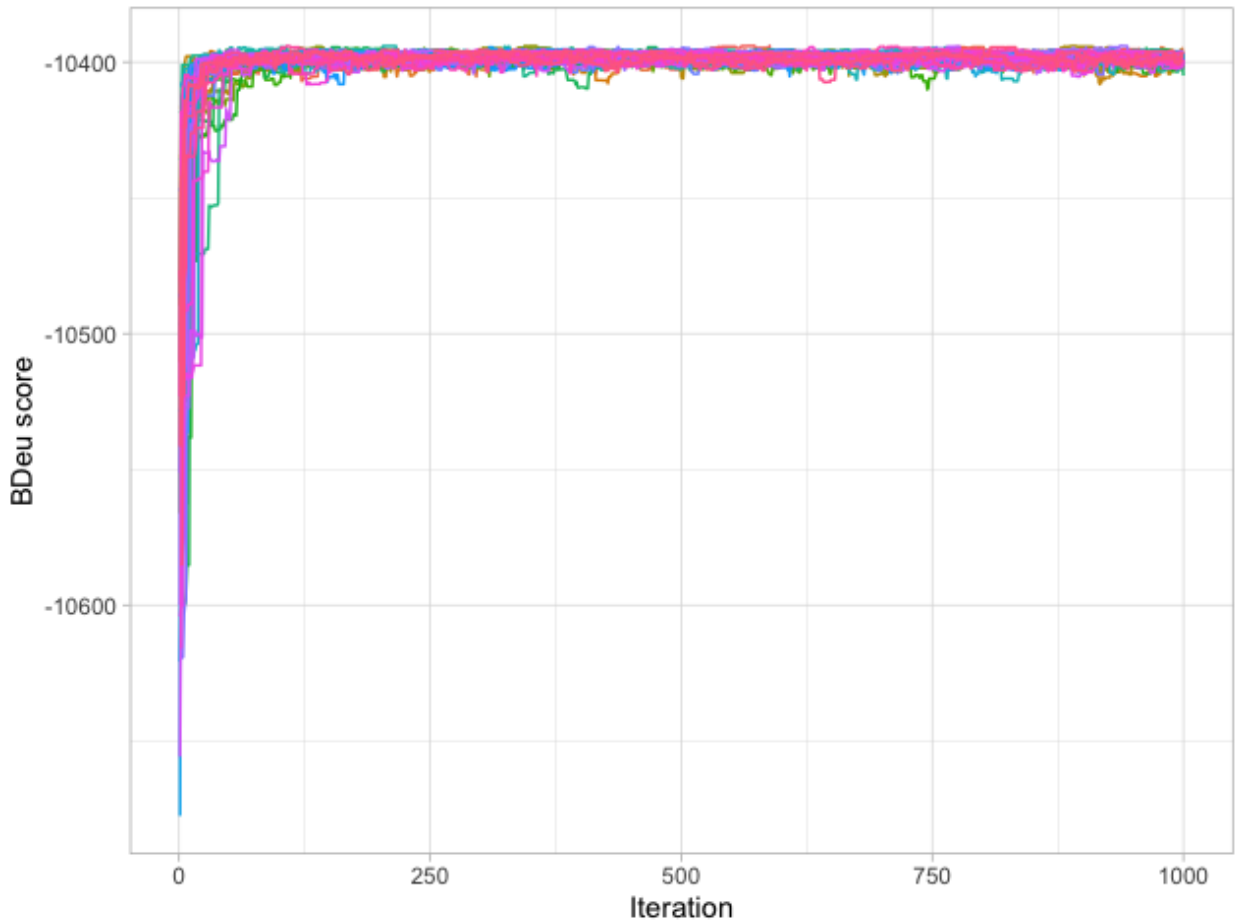
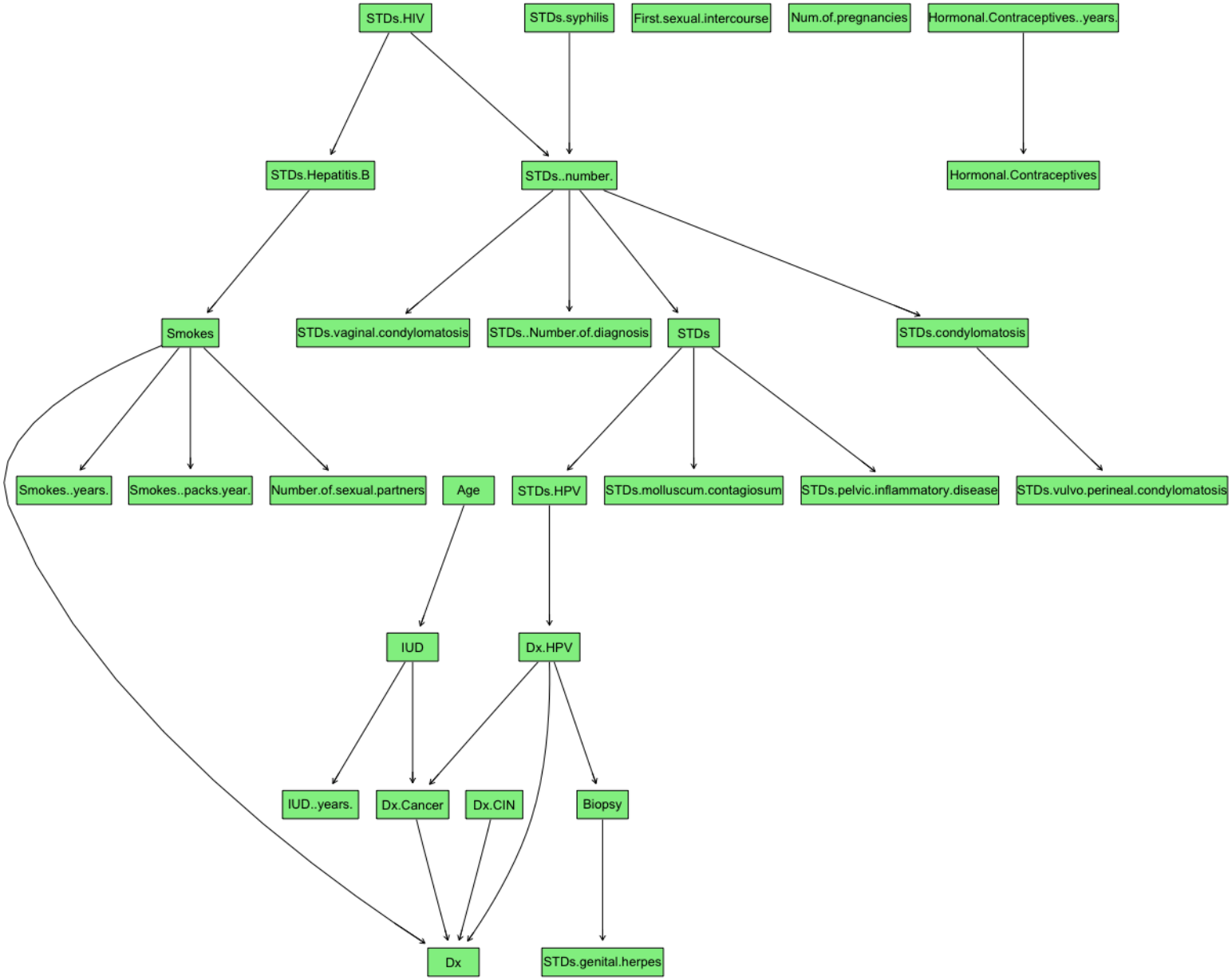


Figure: Boxplot of the AUCs of the ROC curves of 20 replications with SMOTE oversampled dataset.

Results: MCMC convergence of 50 replications



Results: MAP graph



Results: MAP graph

- Correct patterns almost
- IUD:Intrauterine device
- Age $\rightarrow$ IUD? More averaging is better

Results: Edge features

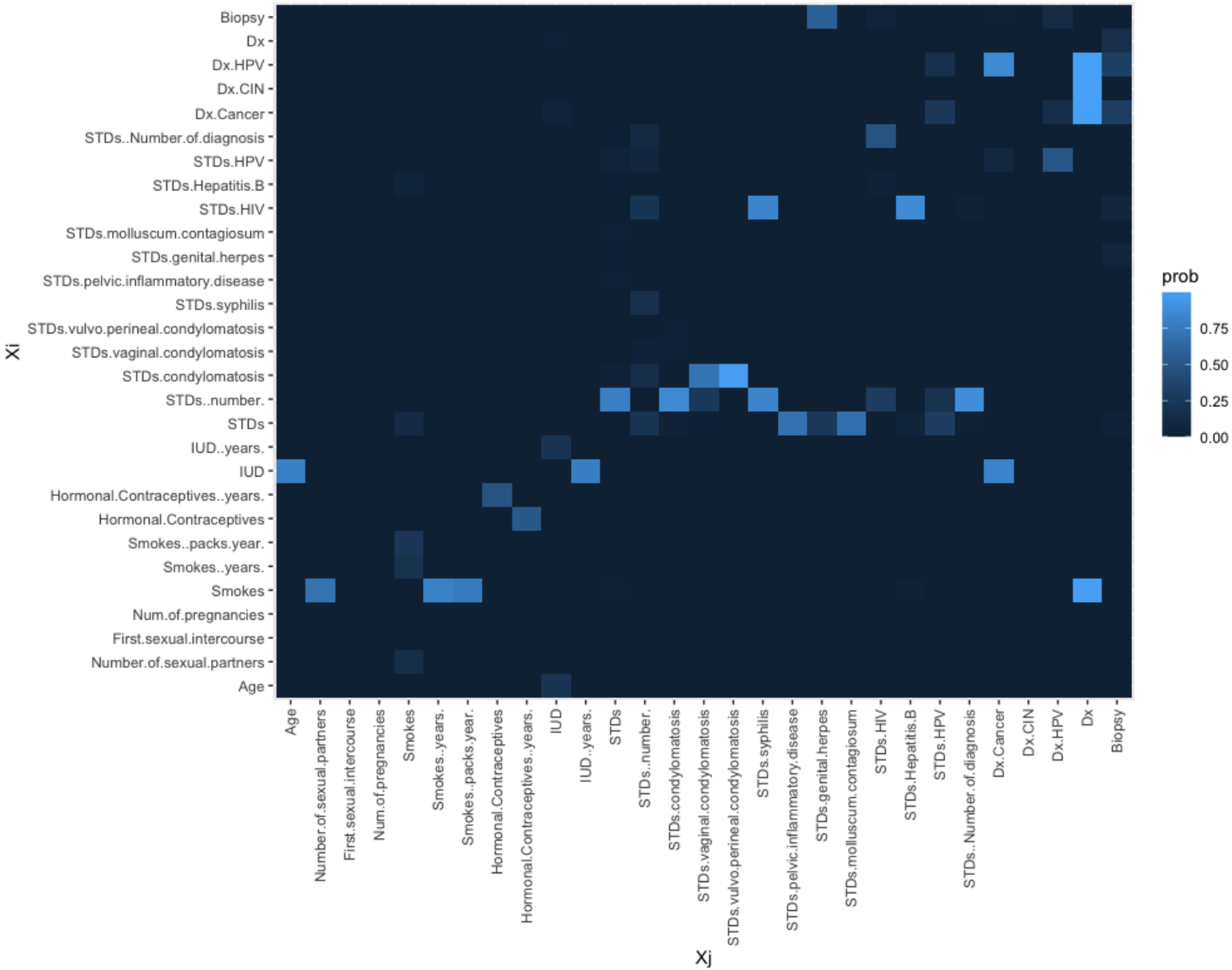
- Edge features:

$$\hat{P}(X_i \to X_j | D) = \frac{1}{M - B} \sum_{m=B+1}^{M} 1_{X_i \to X_j}(G_m), \qquad (6)$$

  B: Burn-in size of MCMC

- "Strength" of the causal relationship

Results: Edge features



Results: Edge features

- $\hat{P}(\text{Dx.HPV} \rightarrow \text{Dx.Cancer}|D) = 0.8601$
- $\hat{P}(\text{IUD} \rightarrow \text{Dx.Cancer}|D) = 0.82518$
- $\hat{P}(\text{Dx.HPV} \rightarrow \text{Biopsy}|D) = 0.3126$
- $\hat{P}(\text{Smokes} \rightarrow \text{Dx.HPV}|D) = \hat{P}(\text{IUD} \rightarrow \text{Dx.HPV}|D) = 0$

Results: Probability Estimation

- $\hat{P}(\text{Dx.HPV} \rightarrow \text{Dx.Cancer}|D) = 0.8601$
  - $P(\text{Dx.Cancer}|\text{Dx.HPV} = 0) = 0.00246$
  - $P(\text{Dx.Cancer}|\text{Dx.HPV} = 1) = 0.86858$
  - Odds ratio $= 2680$
  - Relative risk $= 353.1$
- Signficiant risk increases in cancer with HPV infection

Results: Probability Estimation

- Use the MAP network to estimate (easier to interpret)
- $\hat{P}(\text{IUD} \rightarrow \text{Dx.Cancer}|D) = 0.82518$
  - $P(\text{Biopsy} = 1|\text{Dx.HPV} = 0) = 0.05846$
  - $P(\text{Biopsy} = 1|\text{Dx.HPV} = 1) = 0.33486$
  - Odds ratio $= 8.1$
  - Relative risk $= 5.728$
- The use of IUD increases risk of cervical cancer

Results: Probability Estimation

- $\hat{P}(\text{Dx.HPV} \rightarrow \text{Biopsy}|D) = 0.3126$
  - $P(\text{Biopsy} = 1|\text{Dx.HPV} = 0) = 0.05846$
  - $P(\text{Biopsy} = 1|\text{Dx.HPV} = 1) = 0.33486$
  - Odds ratio $= 85.6$
  - Relative risk $= 5.72$
- Well-known

Results: Probability Estimation

- $\hat{P}(\text{Smokes} \rightarrow \text{Dx.HPV}|D) = \hat{P}(\text{IUD} \rightarrow \text{Dx.HPV}|D) = 0$
  - $P(Dx.HPV = 1|\text{Smokes}= 1) = 0.02112072$
  - $P(Dx.HPV = 1|\text{Smokes}= 0) = 0.02107726$
  - Odds ratio $= 1.00210$
  - Relative risk $= 1.00206$
- Smoking or not does not affect HPV diagnosis

Summary

Causal inference using Bayesian network

Apply to any other datasets

One model, multiple predictions

Good interpretation:

- MAP: A clear network for the relationship of variables
- Stregnth of causal relationship
- Predict risks

Appendix: Mathematical details of Bayesian networks

Score-based structure learning: Posterior likelihood

- Conditional independence: Joint density function decomposition

$$P(X_1, \ldots, X_2) = \prod_{i=1}^{n} P(X_i | pa(X_i)), \tag{7}$$

$pa(X_i)$: parent set of X_i

- The likelihood function is

$$P(D|\Theta) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta_{ijk}, \tag{8}$$

N_{ijk}: Number of case in D that $\{X_i = j | pa(X_i) = k\}$; r_i: Number of possible parent configuration.

Score-based structure learning: Posterior likelihood

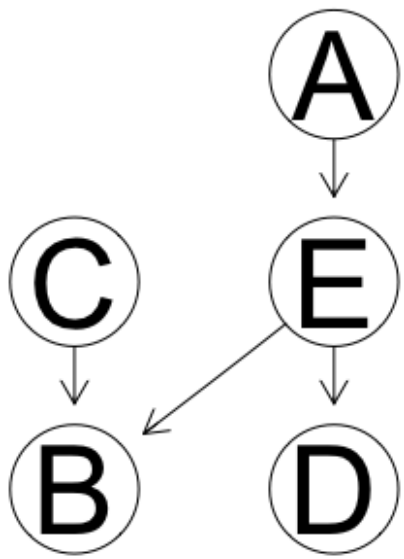
- Assume that $p(\Theta|G) \propto \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta_{ijk}^{\alpha_{ijk}}$, i.e., product Dirichlet distribution with parameters α
- Assume $P(G) \propto 1$
- Apply Bayes' theorem, the posterior is

$$\log P(G|D) \propto \int P(D|\theta, G) P(\theta|G) d\Theta P(G)$$

$$\propto \sum_{i=1}^{n} \left[\log \frac{\Gamma(\alpha_{i.k})}{\Gamma(\alpha_{i.k} + N_{i.k})} + \sum_{j=1}^{q_i} \prod_{j=1}^{r_i} \frac{\Gamma(N_{ijk} + \alpha_{ijk})}{\Gamma(\alpha_{ijk})} \right] \tag{9}$$

Score-based structure learning: Order MCMC

- Sample on the partial ordering space.
- Define $\prec = (\prec_{(1)}, \ldots, \prec_{(n)})$ be the ordering of nodes, which are the permutation of nodes in V
- $\prec_{(i)}$ has higher order than $\prec_{(j)}$ iff the nodeAs corresponding to $\prec_{(j)}$ can only be non-ancestor of $\prec_{(i)}$
- e.g. $\prec = (A, E, C, B, D)$



Score-based structure learning: Order MCMC

- Initialize an order $\prec_0 = (\prec_{(1)}, \ldots, \prec_{(n)})$
- Interchange two orders. New order: $\prec'$
- Accept $\prec'$ (i.e. set $\prec_1 = \prec'$, else $\prec_1 = \prec_0$) with probability

$$\min\left\{1, \frac{P(\prec' | D)}{P(\prec | D)}\right\}, \tag{10}$$

where

$$P(\prec | D) = \prod_{i=1}^{n} \sum_{\Pi \in \Pi_i^{\prec}} P(X_i | \Pi_i) \tag{11}$$

- Then, sample a graph consistent to order $\prec_1$, name it G_1
- Repeat, until the chain converges

Probability Estimation

- Target: Estimate $P(X_i = x|E = e)$; E: evidences

Algorithm 1: Likelihood-Weighted (LW) sampling

Result: Write here the result
Input: A graph G, an event $Y = y$ and an evidence with k vairables, $E = e \equiv (e_1, \ldots, e_k)$;
Initialize $w \leftarrow 1$;
for $i = 1, 2, \ldots, n$ **do**
  if $X_i \in E$ **then**
    Set $x_i \leftarrow e_i$;
    Update $w \leftarrow w \times P(e_i|\Pi_i)$, where Π_i is the parent set of X_i in the graph G ;
  else
    Sample x_i from $P(X_i|\Pi_i)$;
  end
end
return $(x_1, \ldots, x_n), w$

Probability Estimation

- Conduct Algorithm 1 for L times.

$$\hat{P}(Y = y|E = e) = \frac{\sum_{\ell=1}^{L} w_\ell I\{y = y_\ell\}}{\sum_{\ell=1}^{L} w_\ell}, \tag{12}$$

Thank you! Q&A