ИТМО

Фреймворк для вероятностного моделирования на основе байесовских сетей BAMT

Деева Ирина, с.н.с. ЛабКИИ, университет ИТМО

Обо мне

Деева Ирина https://github.com/Anaxagor

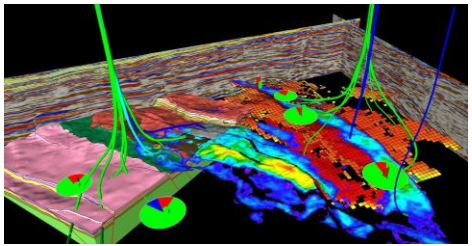
- Кандидат физико-математических наук
- Старший научный сотрудник ЛабКИИ
- Руководитель группы Probabilistic AI

Agenda

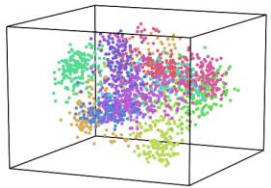
ИТМО

- Как моделировать всё и сразу одной моделью?
- Что такое байесовские сети и как с ними работать?
- Какие есть инструменты для работы с БС;
- Что такое BAMT и в чём его преимущества?
- Как вы можете использовать BAMT уже сейчас – практические примеры;
- Что Open Source грядущий нам готовит?

О дивный, ~~новый~~ многомерный мир



Предметная область



Получение данных с N признаками

Задачи:

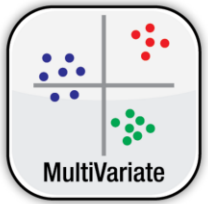
1) Предсказание
2) Анализ признаков
3) Анализ выбросов
4) Моделирование случайных величин

Как решать все эти задачи на многомерных данных?

Подходы к многомерному анализу данных

Предсказание:

1) Модели регрессии
2) Модели классификации



Анализ выбросов:

1) Квантильный анализ
2) Ящики с усами
3) Диаграммы рассеяния

Анализ признаков:

1) Корреляции
2) PCA
3) Факторный анализ
4) Дискриминантный анализ

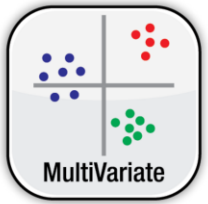
Что не так?

- Нет универсальности, для каждой цели свои модель и подход
- Анализ только частных взаимодействий
- Слабая интерпретируемость
- Плохо работают со смешанными данными и нелинейными зависимостями

Подходы к многомерному анализу данных

Предсказание:

1) Модели регрессии
2) Модели классификации

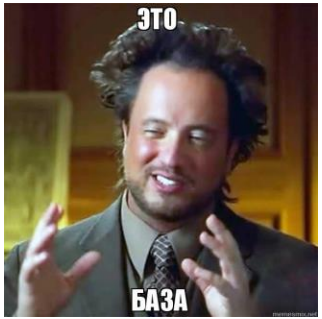


Анализ выбросов:

1) Квантильный анализ
2) Ящики с усами
3) Диаграммы рассеяния

Анализ признаков:

1) Корреляции
2) PCA
3) Факторный анализ
4) Дискриминантный анализ



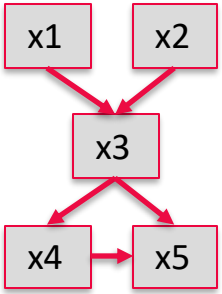
Ответ:

Многомерное распределение!

Байесовская сеть (БС)

Вероятностные графические модели – фреймворк для работы с вероятностными задачами.

Байесовская сеть – направленный ацикличный граф, в узлах которого располагаются распределения признаков, а рёбра обозначают условные зависимости между признаками.



$$x_1 \perp x_4 | x_3$$

$$p(\boldsymbol{X}) = \prod_i p(x_i | x_1, \dots, x_{i-1})$$

$$p(x_1, \dots, x_5) = p(x_1)p(x_2|x_1) \times \cdots$$
$$\dots \times p(x_3|x_1, x_2)p(x_4|x_3)\, p(x_5|x_3, x_4)$$

Обучение БС

Обучение байесовских сетей

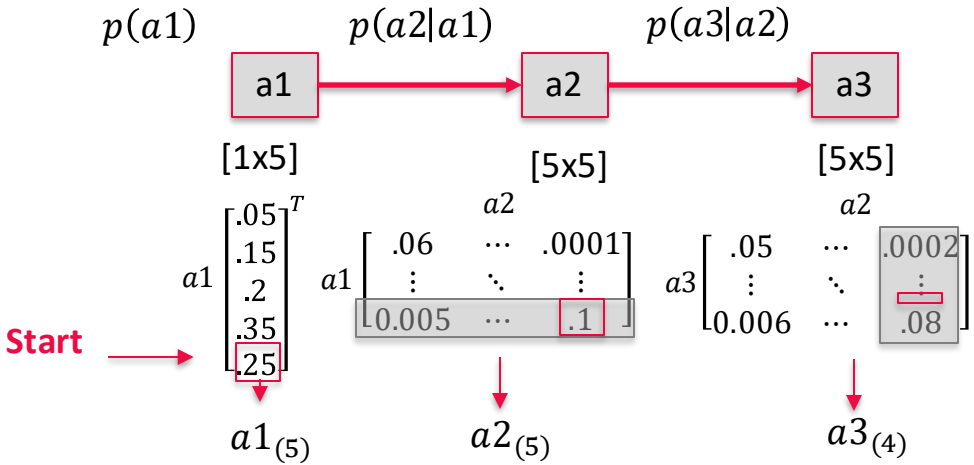
Обучение структуры – нахождение структуры БС из данных

Обучение параметров – нахождение параметров распределений в узлах БС после нахождения структуры

$$\underbrace{\mathrm{P}(\mathcal{M} \mid \mathcal{D})}_{\text{learning}} = \underbrace{\mathrm{P}(\mathcal{G} \mid \mathcal{D})}_{\text{structure learning}} \cdot \underbrace{\mathrm{P}(\Theta \mid \mathcal{G}, \mathcal{D})}_{\text{parameter learning}}$$

Вероятностный вывод

Предковая выборка (ancestral sampling) – выборка из ориентированной модели на основе топологической сортировки вершин графа, обеспечивающей, что выборка из предковых узлов будет проводиться раньше, чем из узлов-потомков.



Ваше случайное значение готово!

$$\boldsymbol{a}_j = [a1_{(5)}, a2_{(5)}, a3_{(4)}]_j, \qquad j = 1 \dots N$$

Используемая модель МСВ:

$$p(\boldsymbol{a}) = p(a1)p(a2|a1)p(a3|a2)$$

Существующие библиотеки

ИТМО

Библиотека	Язык программирования	Обучение на смешанных данных	Продвинутые алгоритмы обучения	Работа с нелинейными зависимостями	Алгоритмы для больших БС	Доп. инструменты для использования
BAMT	Python	✓	✓	✓	✓	✓
bnlearn	R	✓	✓	✗	✗	✓
BiDAG	R	✗	✗	✗	✓	✓
pomegranate	Python	✗	✗	✗	✗	✗
DEAL	R	✓	✓	✗	✓	✗
pgmpy	Python	✓	✗	✗	✗	✓
BayesSuite	Python	✗	✓	✗	✓	✓
Tetrad	Java	✓	✓	✗	✓	✗

Что такое BAMT?

ИТМО

BAMT – open-source фреймворк для вероятностного моделирования на основе байесовских сетей.

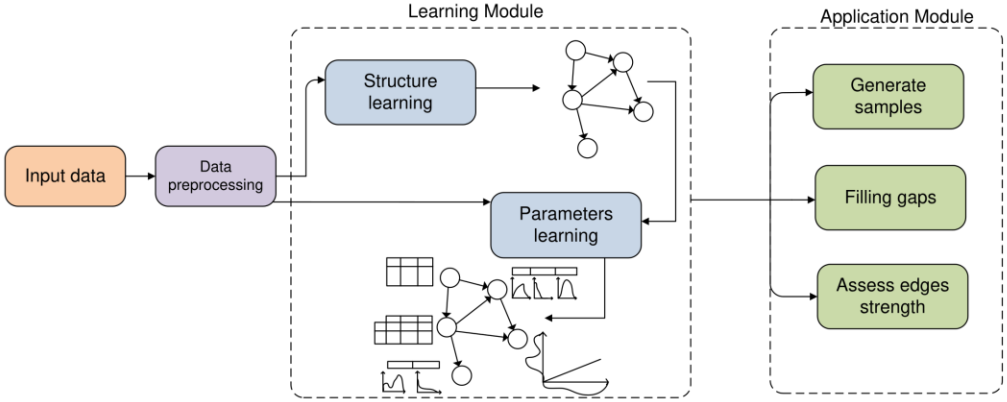
Основные идеи:

- **Построение структур сетей на основе данных** различными алгоритмами, построение композитных БС с моделями МО;
- **Обучение параметров распределений в узлах сети**;
- Поддержка **различных типов данных** (дискретные, непрерывные);
- Генерация **синтетических данных**;
- Поиски нетривиальных зависимостей в данных (нелинейных);
- **Модульность**, расширяемость, интегрируемость с ML-инструментами;
- Сочетание **легковесного API** для конечного пользователя и расширенного конфигурирования для исследовательских задач.

★ 62 звезды

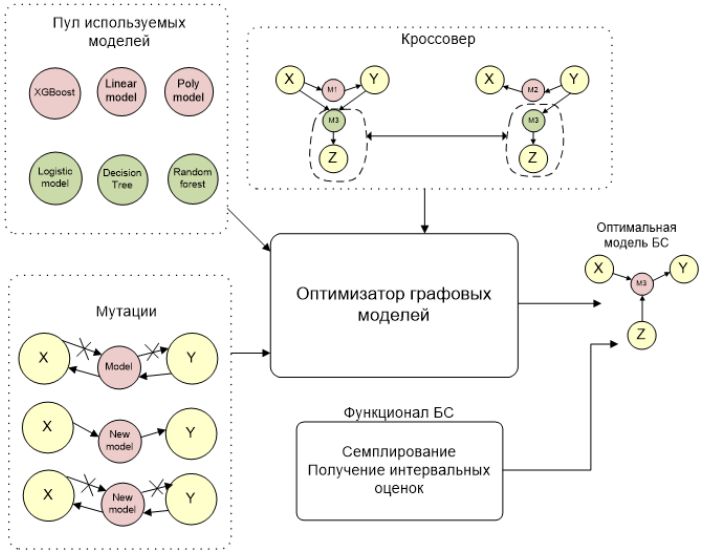
493 просмотра в неделю

Что позволяет делать BAMT

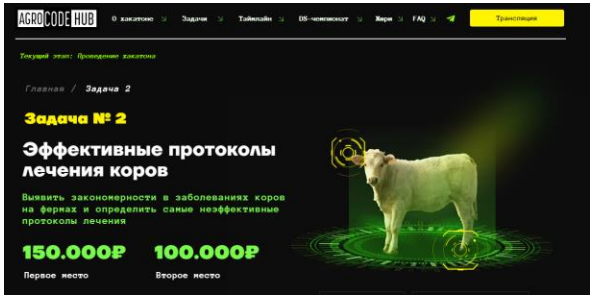


Обучение + применение

Продвинутые алгоритмы обучения



Анализ зависимостей



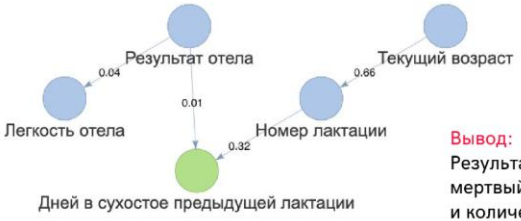
Задача: разработать инструмент для выявления закономерностей в заболеваниях коров.

Условия: 30 команд, 40 часов времени

- Частные корреляции
- Только линейные зависимости
- Не учитываются взаимодействия признаков

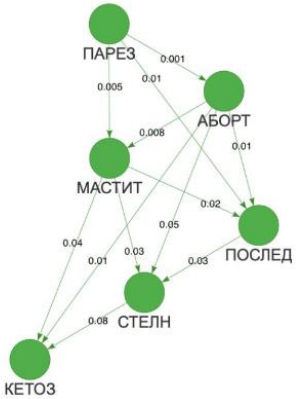
- Причинно-следственные связи
- Учёт взаимодействия признаков
- Работа с нелинейными зависимостями

Решение

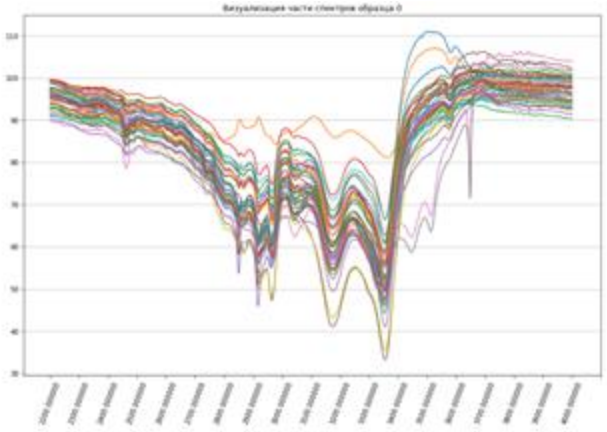


Байесовская сеть связей между событиями, получен при помощи библиотеки BAMT

Вывод:
Результат отела (живой/мертвый) влияет на легкость и количества дней сухостоя, как и возраст и номер лактации



Отбор признаков - инфохимия

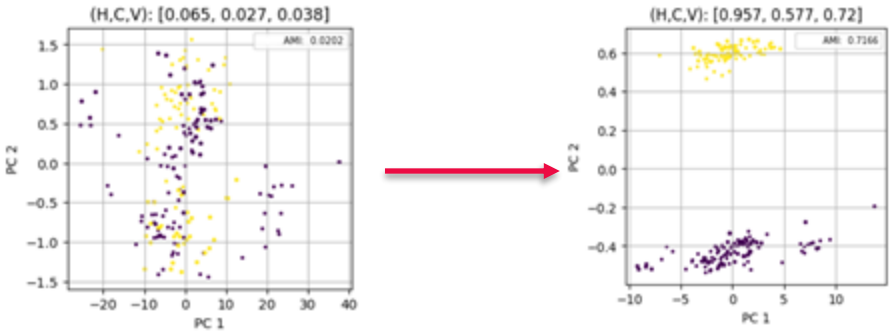


Исходные данные – спектры вещества.

Задача – определить наиболее информативные частоты для моделирования типа вещества.

Марковское окружение – узлы, несущие наибольшую информацию о целевой переменной

Представим спектры – как узлы БС



Отбор признаков - финансы

Входные данные:

Данные клиентов банка

14 дискретных параметров.

3 непрерывных параметра.

Выходные данные:

Предсказать факт открыт депозита (yes/no).

Отобранные признаки:

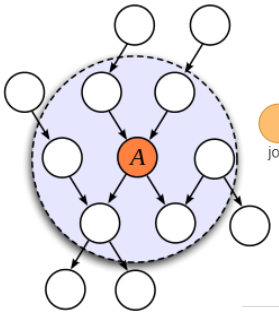
Contact – как контактировали с клиентом (по телефону, очно),

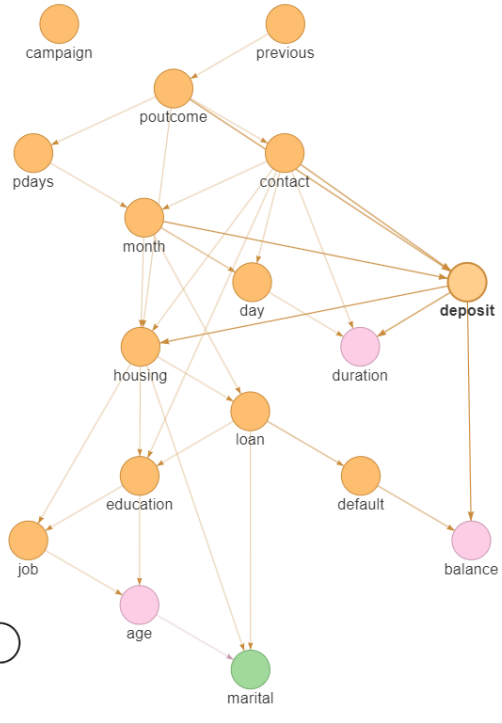
Poutcome – результат маркетинговой компании в отношении клиента,

Month – время последнего контакта

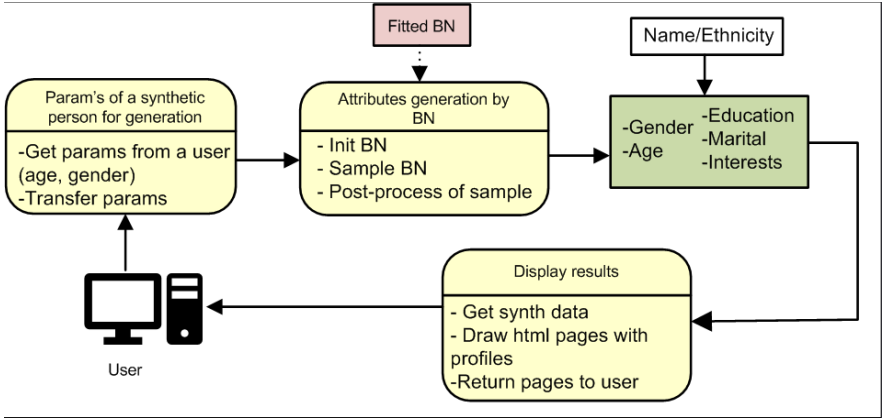
Время обучения **на полных признаках** ~ 30 сек. ROC-AUC=0.92

Время обучения **на отобранных признаках** ~ 15 сек. ROC-AUC=0.9

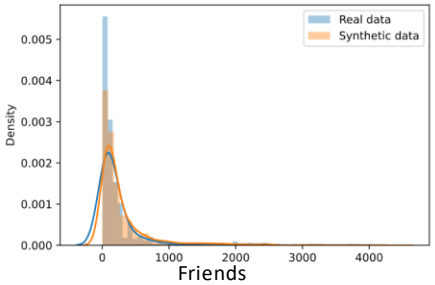


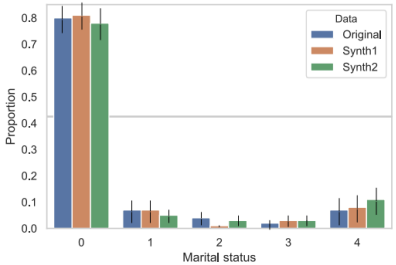


Генерация синтетических данных



Pipeline генератора профилей





Profile picture

Name : Lyudmila

Age: 47

Gender: female

High education: Yes

Family status: married

User interests:

life, child, children, human, love, relationships, peace

coffee, oil, add, water, tea, taste, eggs, salt, dough, meat, spoons, sugar, milk

life, heart, love, mom, girl, near, hands, home

Profile picture

Name : Konstantin

Age: 32

Gender: male

High education: No

Family status: not specified

User interests:

invitation, friends, public, subscribe, people, cool

series, games, match, tournament, teams, season, place

competition, participation, participants, link, necessary, repost

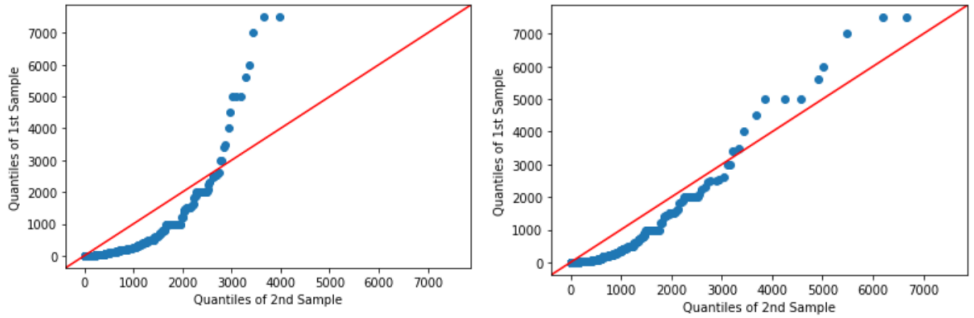
Примеры профилей

Композитные БС

При **композитном подходе** модели МО подбираются прямо **во время обучения байесовских сетей** и являются частью модели байесовской сети.
В этом случаи узлы БС делятся на два типа:

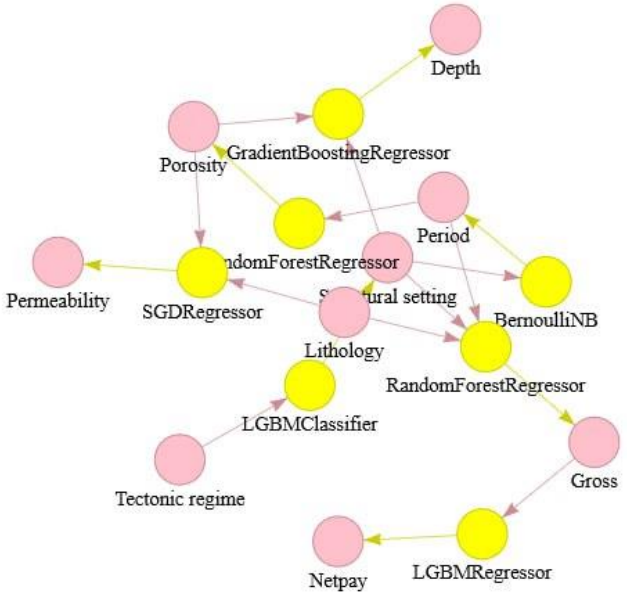
- узлы-переменные
- узлы-модели

Результат вероятностного вывода



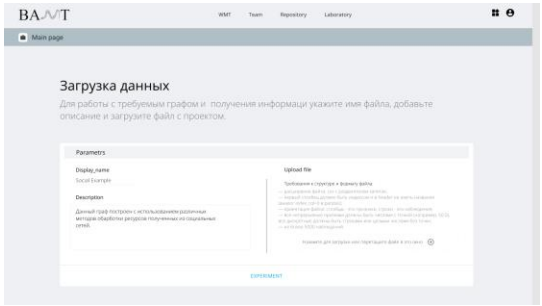
Классическая БС с линейной регрессией

Композитная БС с подобранной RFRegressor регрессией

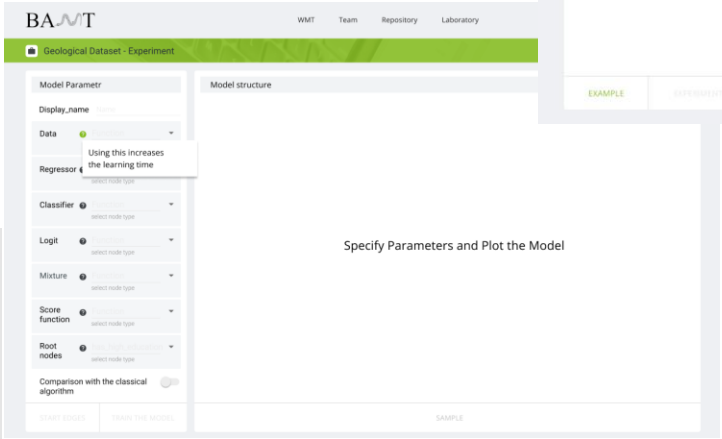


Web-BAMT

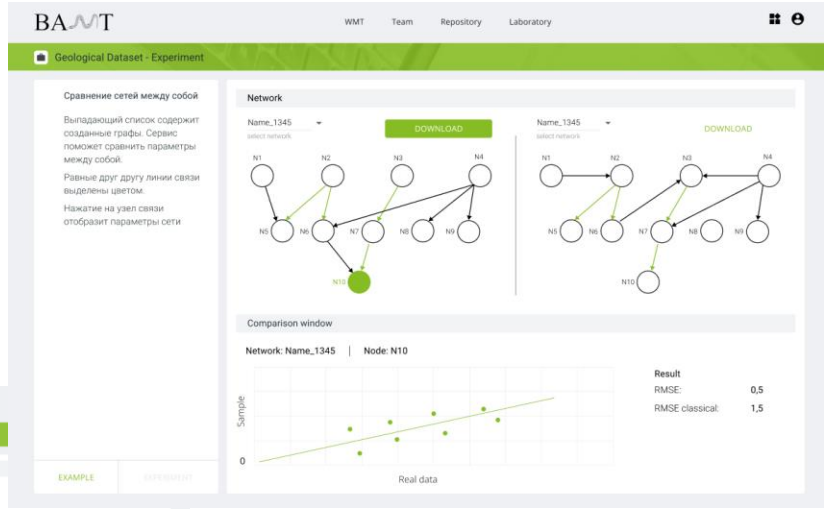
https://github.com/aimclub/Web-BAMT



1. Загрузить данные



2. Настроить и обучить



3. Проанализировать, выгрузить результат

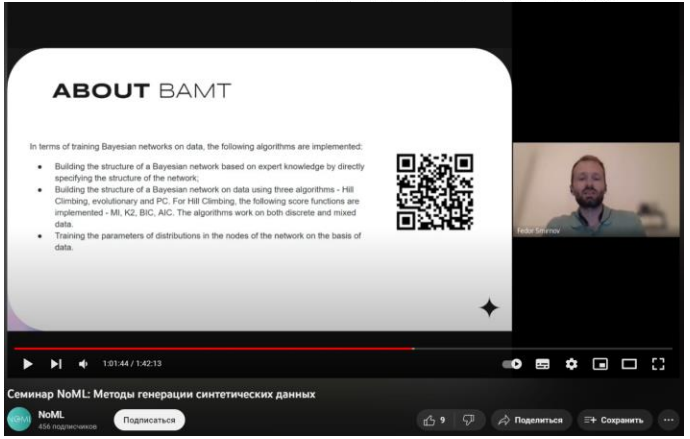
Опыт Open Source

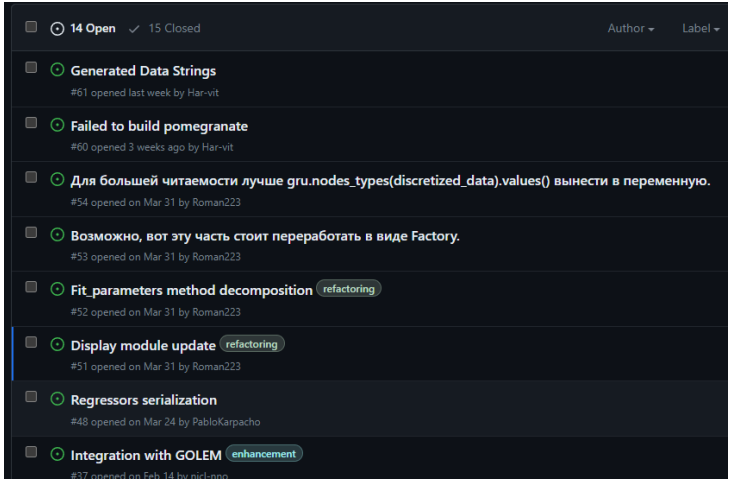
Плюшки от Open Source:

- О нас узнают и наш продукт активнее используют;
- Сообщество помогает улучшать продукт, искать ошибки;
- Сообщество подкидывает идеи для развития функционала;
- Нас чаще цитируют.

Трудности:

- Код необходимо поддерживать и это забирает время;
- Использование порождает вопросы, что также времязатратно.





А что дальше?

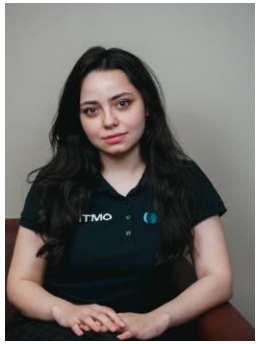
ИТМО

- Развитие функциональности BAMT в сторону повышения качества обучения БС;
- Развитие прикладных инструментов использования BAMT для анализа данных, моделей ML и т. д.;
- Повышение качества кода, развитие архитектуры;
- Популяризация.

Команда разработчиков

ИТМО







Кто следующий?

Если интересно поучаствовать – пишите!

https://github.com/aimclub/BAMT

Спасибо
за внимание!

ITMO more than a UNIVERSITY