Глава 1

Постановка задачи

1.1 Задача оценивания рисков на графах

Пусть $\mathbf{X} = \{x_1, \ldots, x_n\}$ - множество индивидов.
Пусть $T = \{1, 2, \ldots, t, \ldots, T\}$ - дискретные отрезки времени. Пусть в каждый момент t у каждого индивида x есть состояние $y_t(x)$: $y_t(x) = \mathbf{I}$ - инфицированный или $y_t(x) = \boldsymbol{S}$ - здоровый. Предлагается разработать модель, предсказывающую для каждого индивида x и момента t вероятность что индивид индивид инфицирован, $p_t(x) = \mathbf{P}(y_t(x) = \mathbf{I})$.
Обучающая выборка D состоит из потока данных, монотонных по времени. Данные бывают двух типов:

- $\langle t, (u, v)\rangle$ - контакт индивида u с индивидом v в момент времени t
- $\langle t, y_t(x)\rangle$ - смена состояния индивида в момент времени t

Критерием качества является максимизация логарифма правдоподобия модели:

$$Ł(w) = \frac{1}{D} \sum_{(t,x)\in D} [y_t(x) = \mathbf{I}] \log(p_t(x, w)) - [y_t(x) \neq \mathbf{I}] \log(1 - p_t(x, w)) \to \max_w \qquad (2.1)$$

В данной работе рассматриваются несколько последовательно усложняющихся моделей и сравнивается их качество.

Глава 2

Рассматриваемые модели

2.1 Частотная модель

В качестве основного уравнения протекания заражения берется модель SIS[цитата]. В ней вероятность инфицирования в момент t, раскладывается на $q_t(x)$ - вероятность инфицирования в интервал $[t-1,t]$, и $p_{t-1}(x)$ - вероятность инфицирования ранее:

$$p_t(x) = (1-\mu)p_{t-1}(x) + \beta(1-p_{t-1}(x))q_t(x) \tag{3.1}$$

Вероятность инфицирования в интервал $[t-1,t]$ можно оценить моделью логистической регрессии, основанной на количестве контактов в интервал:

$$q_t(x) = \sigma(w_1 k_t(x) - w_0) \tag{3.2}$$

$$k_t(x) = \sum_{\langle t:(x,v)\rangle} [t-1 \le t' \le t] \tag{3.3}$$

Данная модель не учитывает что у разных контактов x может быть как разная вероятность передачи инфекции x, так и разная вероятность инфицирования от x. Кроме того, модель не учитывает, что при получении информации о смене состояния x, вероятности всех недавно контактировавших с ним индивидов u так же должны измениться.

2.2 Учет вероятности передачи инфекции

Добавим в оценку количества контактов (формула 3.3) оценку вероятности передачи инфекции при данном контакте:

$$k_t(x) = \sum_{\langle t:(x,v)\rangle} [t-1 \le t' \le t] a_{t'}(x,v) \tag{3.4}$$

Вероятность передачи инфекции $a_{t'}(x,v)$ оцениваем так же с помощью модели логистической регрессии:

$$a_{t'}(x,v) = \sigma(-\alpha_0 + \sum_{j=1}^{m} \alpha_j f_j) \tag{3.5}$$

Здесь α_j - параметры модели, f_j - различные признаки контакта, например

- Длительность контакта
- Близость координат gps во время контакта
- Число совпадающих bluetooth маяков и wi-fi

В этой модели учитываются что различный характер контакта по-разному влияет на вероятность инфицирования в момент t, но не учли, что вероятность инфицирования зависит еще и от того, болен ли или здоров второй участник контакта.

2.3 Рекурсивное оценивание рисков

Добавим в оценку количества контактов (формула 3.3) вероятность что второй участник контакта болен, $\tilde{p}_t(v)$:

$$k_t(x) = \sum_{\langle t:(x,v)\rangle} [t-1 \le t' \le t] a_{t'}(x,v)\tilde{p}_{t'}(v) \tag{3.6}$$

$$\tilde{p}_{t'}(v) = \begin{cases} 1, & \text{if } y_t(v) = \mathbf{I} \\ p_{t'}(v), & \text{if } y_t(v) \neq \mathbf{I} \end{cases} \tag{3.7}$$

Для всех предыдущих моделей вероятность инфицирования на шаге t, $p_t(x)$,зависела только от вероятности инфицирования на шаге $t-1$ (формула 3.1). Теперь добавляется зависимость на шаге t, $p_t(v)$. В качестве значения $p_t(v)$ можно либо брать значение с предыдущего шага $p_{t-1}(v)$, либо рекурсивно распространять градиент через суперпозицию функций, ограничившись определенной глубиной суперпозиции d.

2.4 Модель распространения рисков по сети

Предыдущие модели не учитывают, что при смене состояния индивида (появлении результата тестирования на наличие инфекции), например, $y_t(x) = \mathbf{I}$, должны рекурсивно измениться оценки рисков для всех индивидов, контактировавших с x для любого t' из отрезка времени $[t-d,t]$ («контактных» индивидов). После этого по цепочке должны изменится оценки рисков для индивидов, контактировавших с «контактными» индивидами, для моментов времени t'' из интервала $(t',t]$, и так далее.

Добавим в формулу оценивания вероятности инфицирования в момент t (формула 3.2) вероятность того, что x будет инфицирован в ближайшем будущем, на интервале $(t, t+d]$:

$$q_t(x) = \sigma(w_1 k_t(x) + w_2 b_t(x) - w_0) \tag{3.8}$$

$$b_t(x) = [t' \,:\, t < t' \leq t + d \quad \text{и} \quad y_{t'}(x) = \mathbf{I}] \tag{3.9}$$

При появлении информации о смене состояния индивида x на I, запускается *Алгоритм распространения рисков по сети*, при этом риск для x скачком увеличивается до 1, $\Delta p_t(x) = 1 - p_t(x)$, а затем увеличиваются оценки рисков для всех контактировавших с x, $p_t(v)$, и так по цепочке контактов.

2.5 Алгоритм распространения рисков по сети

Алгоритм состоит из двух функций: $BakwardUpdate(x, t)$ и $ForwardUpdate(x, t_0, t)$. При смене состояния индивида x запускается функция $BakwardUpdate(x, t)$. В ней для всех контактов x в момент t_0 из интервала $t - d, t$ пересчитывается вероятность $p_{t_0}(x)$ по формуле (3.1). Так как в этой формуле теперь $b_{t_0} = 1$, вероятность $p_{t_0}(x)$ увеличивается на $\Delta p_{t_0}(x)$. Функция $ForwardUpdate(x, t_0, t)$ обновляет оценку рисков для всех контактов x вперед с t_0 до t. Если приращение $\Delta p_{t'}(u)$ было достаточно велико, то запускается пересчет весов и для контактов $u(ForwardUpdate(u, t', t))$. Чтобы не допустить зацикливание, вводится множество просмотренных индивидов, и пересчет запускается только для непросмотренных индивидов. Псевдокод функций приведен ниже.

Algorithm 2.5.1 BakwardUpdate(x, t)

```
1: U = ∅
2: for t ∈ [t − d, t] : (x, v) do
3:     Пересчитать p'_t(x)
4: ForwardUpdate(x, td, t)
```

Algorithm 2.5.2 ForwardUpdate(x, t_0, t)

```
1: U := U ∪ {x}
2: for t ∈ [t − d, t] : (x, u ∉ U) do
3:     Пересчитать p_{t'}(u)
4: for t ∈ [t − d, t] : (x, u ∉ U) do
5:     if Δp'_t(u) > ϵ then
6:         ForwardUpdate(u, t', t)
```

Глава 3

Эксперименты

3.1 Данные

В качестве данных берем данные из приложения «ContactTracer». Данные включают в себя геопозицию пользователей в каждый момент времени, а так же сведения о wi-fi и bluetooth маяках. Из данных извдекаем поток, отсортированный по времени из двух типов объектов:

- $\langle t, (u, v)\rangle$ - контакт индивида u с индивидом v в момент времени t
- $\langle t, y_t(x)\rangle$ - смена состояния индивида в момент времени t

3.2 Процесс обучения

Модели обучаем в парадигме онлайн-обучения[цитата]: на каждом шаге мы

- Получаем элемент из потока
- Делаем предсказание вероятности инфицирования $p_t(x)$
- Получаем состояние индивида $y_t(x)$
- Считаем ошибку на одном элементе
- Делаем градиентный шаг

Рассмотрим формулу (3.1): $p_t(x) = (1-\mu)p_{t-1}(x) + \beta(1-p_{t-1}(x))q_t(x)$. В ней в левой части присутствует вероятность, а в правой - вероятности складываются с множителями-параметрами модели $(1-\mu)$ и β. Поэтому мы накладываем на μ и β ограничения - $0 \leq \mu, \beta \leq 1$. В качестве оптимизирующего алгоритма используется проективный градиентный спуск.
Мы получаем данные потоком по одному элементу. Так как функция ошибки (формула 2.1) аддитивная относительно элементов выборки, то на каждом шаге мы считаем градиент скользящим средним.

3.3 Сравнение моделей

Сравнение моделей происходит так:

- Модели обучаются на заданном потоке данных
- Берется история предсказаний
- По истории предсказаний считается precision-recall кривая

Мы сравниваем модели со всеми фиксированными гиперпараметрами, кроме параметра скользящего среднего и шага градиента - эти параметры мы оптимизируем по AUC. Мы делаем это не на отложенной выборке, так как контактов очень мало. На графике сравнения моделей (Рис. 1) мы видим что добавление признаков улучшает качество модели, но незначительно.

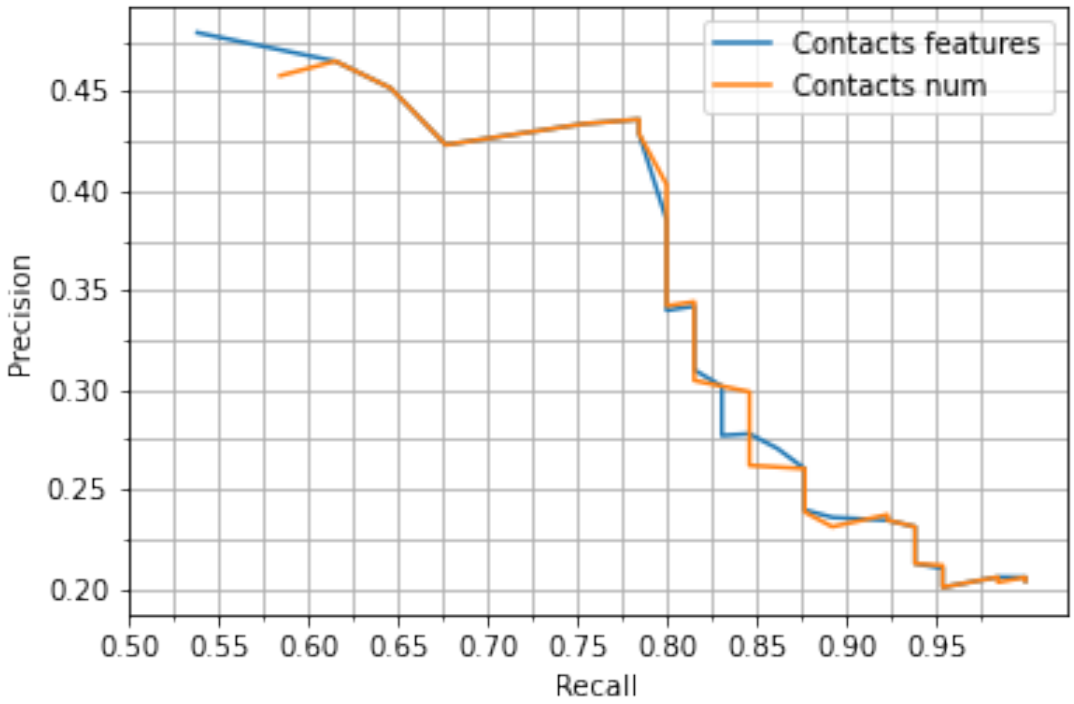

Рис. 1. Сравнение частотной модели и модели с признаками контактов