Распространение рисков на графах

Постановка задачи

Существуют данные о пользователях двух типов:

- t, (u, v) - контакт между пользователями u, v в момент времени t
- t, $y(x)$ - состояние пользователя x в момент времени t

Состояния пользователей могут I(infected), S(susceptible) или еще R(recovered), E(exposed) и другие.
Необходимо по последнему известному состоянию $y_t(x)$ построить вероятностную модель $p_t(y, x)$ - вероятность состояния y для пользователя x в момент t.

Далее считаем что состояний всего два - Infected и Susceptible.

Методы решения

Простая частотная модель

Вводится число контактов у пользователя x в момент t
за время δ - $k_t(x, \delta)$.
Вероятность заражения считается как $p_t(x) = \sigma(k_t(x, \delta))$, где в качестве $\sigma()$ берется например эмпирическая функция распределения $k_t(x, \delta)$.

Логистическая регрессия

Вводится число контактов с инфицированными $k_t^I(x, \delta)$. Пусть $p_t(x, w) = \sigma(\log p_{t-\delta} - w_0 + w_1 \cdot k_t(x, \delta) + w_2 \cdot k_t^I(x, \delta))$.
Здесь $\sigma(z) = \frac{1}{1+e^{-z}}$ - сигмоидная функция. Кроме того, по постановке задачи все коэффициенты должны быть неотрицательными, поэтому это логистическая регрессия с неотрицательными коэффициентами.
Предлагается обучать на парах (x, t) с функцией ошибки
$L(w) = \frac{1}{|D|} \sum_{(x,t)\in D} [y_t = I] \log p_t(x, w) + [y_t \neq I] \log(1 - p_t(x, w)) \to \max_w$

Данные и эксперименты

Данные приходят из приложения ContactTraicer. Есть данные о местоположении пользователя - таблица (userid, longtitude, latitude, timestamp) и

данные о плохом самочувствии пользователя в формате (userid, timestamp).
Для решения задачи проведена работа с данными реализован алгоритм преобразования данных в таблицу контактов (userid, userid, timestamp) и подготовки обучающей выборки.
Обучающая выборка представляет собой временную сетку (userid, timestamp, features) с шагом δ.
На данный момент данных мало, поэтому качество моделей оставляет желать лучшего.

Проведены эксперименты:

- С построением частотной модели. На отложенной выборке precision 0.158, recall 0.167.
- С построением логистической регрессии. Для этого реализован класс логистической регрессии с неотрицательными коэффициентами на pytorch.
  Так как модель зависит от $\log p_{t-\delta}$, реализован процесс обучения с динамическим пересчетом вероятностей после ряда итераций градиентного спуска. На той же отложенной выборке получила качество precision 0.19, recall 0.4.