Постановки задач оптимизации в машинном обучении
Часть 1: обучаемые модели распространения риска на графе

Воронцов Константин Вячеславович

(Лаборатория машинного интеллекта МФТИ)

• Управление, информация, оптимизация • «Сириус», Сочи • 23–29 августа 2020

Содержание

Как оценить риск инфицирования по данным о контактах?

Выборка данных:

- $\langle t : (u, v) \rangle$ — *контакт* индивидов u и v в момент времени t
- $\langle t : y(x) \rangle$ — *состояние* y индивида x в момент времени t

Состояния:

- S (susceptible) — восприимчивый здоровый
- E (exposed) — латентный инфицированный
- I (infected) — инфицированный больной
- R (recovered) — выздоровевший невосприимчивый

Задача — построить модель индивидуального риска:
$p(y|t, x)$ — вероятность состояния y индивида x в момент t

Наша мотивация — объединить два типа моделей:

- модели риска, обучаемые по выборке данных
- модели распространения эпидемии по графу контактов

Оптимизационная задача обучения классификация

Обучающая выборка: $X^\ell = (x_i, y_i)_{i=1}^\ell$, $x_i \in \mathbb{R}^n$, $y_i \in \{-1, +1\}$
$f_1(x), \ldots, f_n(x)$ — признаковое описание объекта x

1. Фиксируется модель классификации, например, *линейная*:

$$a(x, w) = \operatorname{sign}\langle x, w\rangle = \operatorname{sign} \sum_{j=1}^{n} w_j f_j(x)$$

2. Функция потерь — пороговая или её верхняя оценка:

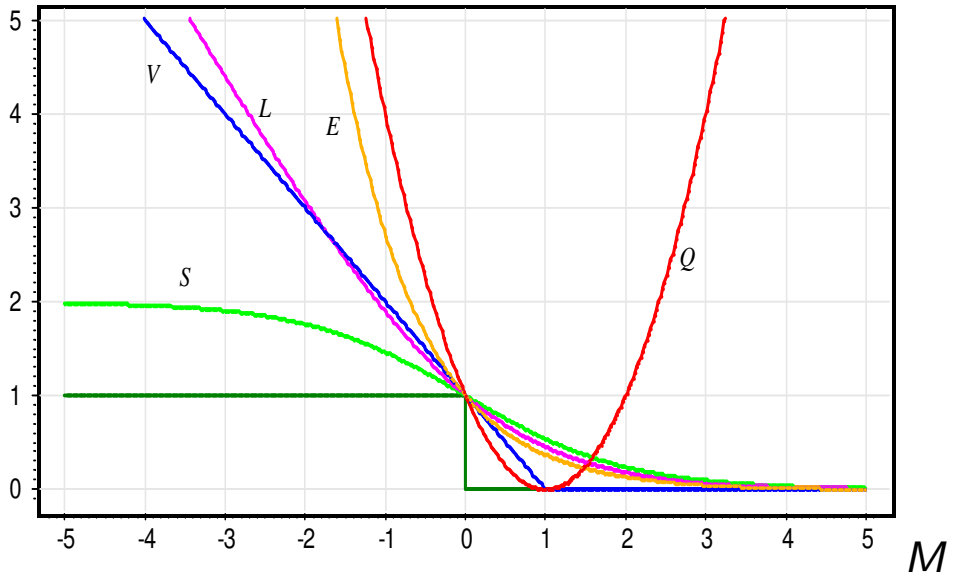
$$L_i(w) = \big[a(x_i, w) y_i < 0\big] = \big[\langle x_i, w\rangle y_i < 0\big] \leqslant \mathscr{L}\big(\langle x_i, w\rangle y_i\big)$$

3. Метод обучения — *минимизация эмпирического риска*:

$$\sum_{i=1}^{\ell} \big[\langle x_i, w\rangle y_i < 0\big] \leqslant \sum_{i=1}^{\ell} \mathscr{L}\big(\langle x_i, w\rangle y_i\big) \to \min_w$$

Непрерывные верхние оценки пороговой функции потерь

Часто используемые непрерывные функции потерь $\mathscr{L}(M)$:



$[M < 0]$ — пороговая функция потерь

$V(M) = (1 - M)_+$ — кусочно-линейная (SVM)

$H(M) = (-M)_+$ — кусочно-линейная (Hebb's rule)

$L(M) = \log_2(1 + e^{-M})$ — логарифмическая (LR)

$Q(M) = (1 - M)^2$ — квадратичная (FLD)

$S(M) = 2(1 + e^{M})^{-1}$ — сигмоидная (ANN)

$E(M) = e^{-M}$ — экспоненциальная (AdaBoost)

Связь правдоподобия с аппроксимацией эмпирического риска

$X \times Y$ — в.п. с плотностью $p(x, y|w) = P(y|x, w)p(x)$.
$(x_i, y_i) \sim p(x, y|w)$ — простая (i.i.d.) выборка, $i = 1, \ldots, \ell$
$P(y|x, w)$ — вероятностная модель классификации

- *Максимизация правдоподобия (Maximum Likelihood, ML):*

$$\sum_{i=1}^{\ell} \log P(y_i|x_i, w) \to \max_w.$$

- *Минимизация аппроксимированного эмпирического риска:*

$$\sum_{i=1}^{\ell} \mathscr{L}\big(\langle x_i, w\rangle y_i\big) \to \min_w;$$

Эти задачи эквивалентны при $\mathscr{L}\big(\langle x_i, w\rangle y_i\big) = -\log P(y_i|x_i, w)$

$$\boxed{\text{модель } P(y|x, w)} \quad \rightleftarrows \quad \boxed{\text{модель } \langle x_i, w\rangle \text{ и } \mathscr{L}(M)}$$

Логистическая регрессия (двухклассовая)

Линейная модель классификации $a(x, w) = \text{sign}\langle x, w\rangle$

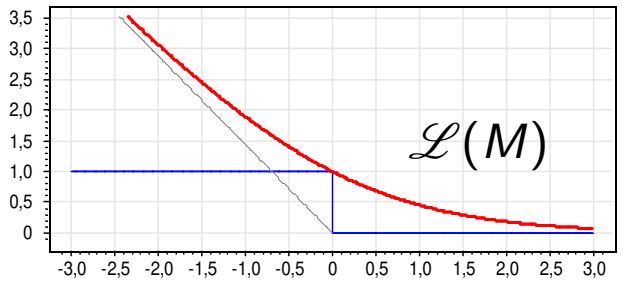
$M_i(w) = \langle x_i, w\rangle y_i$ — отступ (margin) для линейной модели

Максимизация правдоподобия данных с регуляризацией:

$$\sum_{i=1}^{\ell} \ln\bigl(1 + \exp(-\langle w, x_i\rangle y_i)\bigr) \;\to\; \min_w$$

Логарифмическая функция потерь:

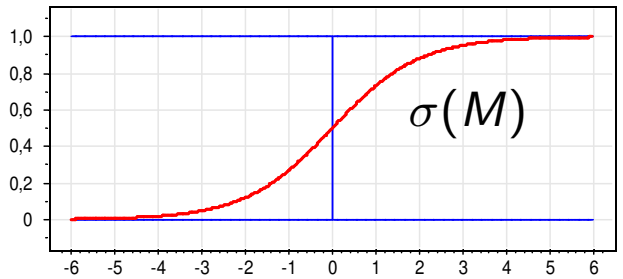
$$\mathscr{L}(M) = \ln\bigl(1 + e^{-M}\bigr)$$



Модель условной вероятности:

$$P(y|x, w) = \sigma(M) = \frac{1}{1+e^{-M}},$$

где $\sigma(M)$ — сигмоидная функция



Логистическая регрессия (многоклассовая)

Линейный классификатор при произвольном числе классов $|Y|$:

$$a(x, w) = \arg\max_{y \in Y} \langle w_y, x \rangle, \quad x, w_y \in \mathbb{R}^n$$

Вероятность того, что объект x относится к классу y:

$$P(y|x, w) = \frac{\exp\langle w_y, x \rangle}{\sum_{z \in Y} \exp\langle w_z, x \rangle} = \underset{y \in Y}{\text{SoftMax}} \langle w_y, x \rangle,$$

где SoftMax: $\mathbb{R}^Y \to \mathbb{R}^Y$ переводит произвольный вектор в нормированный вектор дискретного распределения.

Максимизация правдоподобия (log-loss) с регуляризацией:

$$-\sum_{i=1}^{\ell} \ln P(y_i|x_i, w) \;\to\; \min_w .$$

Структура исходных данных по распространению Covid-19

Дана выборка контактов и состояний индивидов:

- $\langle t : (u, v) \rangle$ — *контакт* индивидов u и v в момент времени t
- $\langle t : y(x) \rangle$ — переход индивида x в *состояние* y в момент t

Найти модель индивидуального риска:

- $p(y|t, x)$ — вероятность состояния y индивида x в момент t

Критерий:

- максимум правдоподобия наблюдаемых состояний

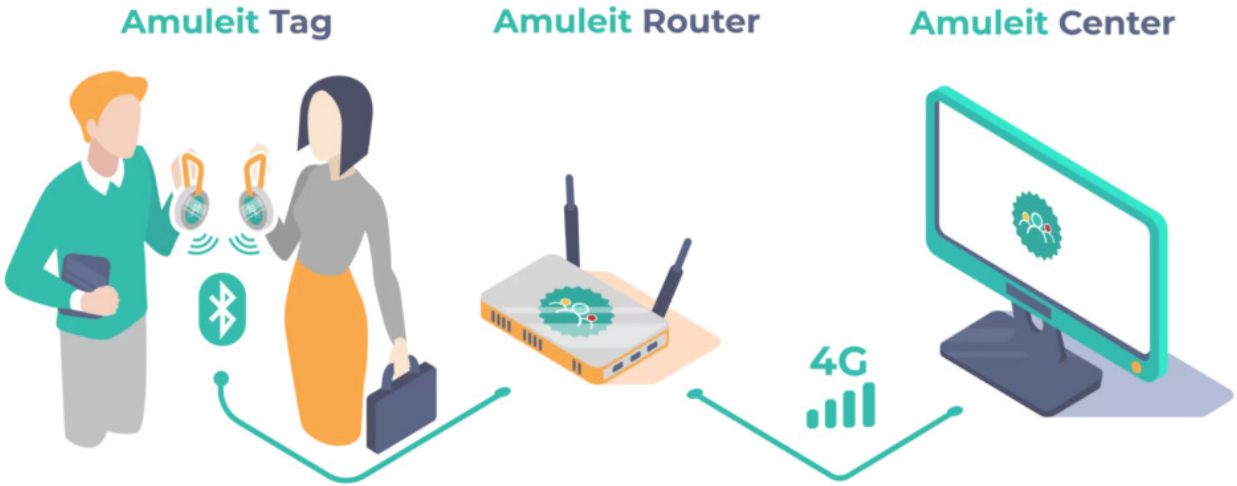
Обозначения и уточнения:

- $f(t, u, v)$ — информация о контакте, вектор признаков
- $y_t(x)$ — текущее состояние индивида x в момент t
- $y \in \{\mathtt{S}, \mathtt{I}\}$ — будем рассматривать только два состояния

Проект Amuleit: контроль эпидемии на предприятиях

Амулет — носимое устройство для сотрудников предприятий

Три основных элемента архитектуры системы Amuleit:



---

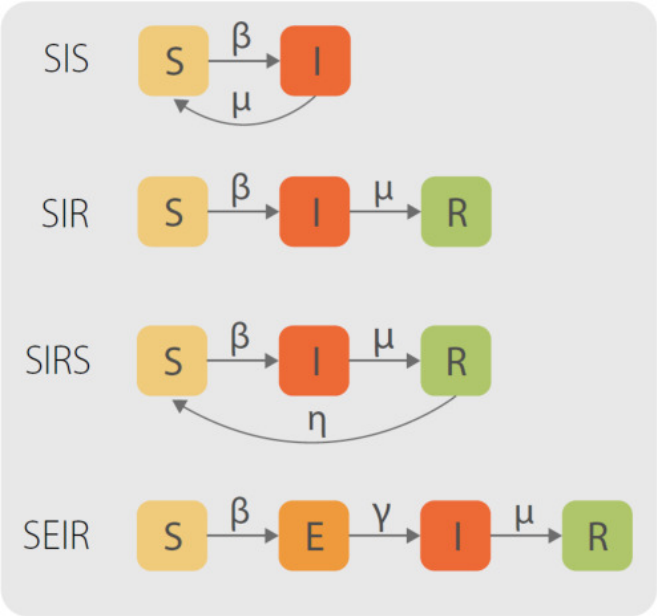
http://Amuleit.ru — сайт проекта

http://SoftTree.ru — сайт разработчика ООО Софттри (г.Пенза)

Популяционные (компартментные) модели SIS, SIR, SIRS, SEIR

$S(t)$ — число восприимчивых
$E(t)$ — число латентных носителей
$I(t)$ — число инфицированных
$R(t)$ — число выздоровевших
$S + E + I + R = \text{const}$ — вся популяция

SIR — постоянный иммунитет
SIRS — временный иммунитет
SEIR — латентное инфицирование



$$\begin{cases} S' = -\beta IS + \mu I \\ I' = \beta IS - \mu I \end{cases}$$

$$\begin{cases} S' = -\beta IS \\ I' = \beta IS - \mu I \\ R' = \mu I \end{cases}$$

$$\begin{cases} S' = -\beta IS \\ E' = \beta IS - \gamma E \\ I' = \gamma E - \mu I \\ R' = \mu I \end{cases}$$

R. Pastor-Satorras et al. Epidemic processes in complex networks. 2014

Модель SIS для оценивания индивидуального риска $p_t(x)$

$p_t(x) = P(\mathtt{I}|t,x)$ — вероятность, что x инфицирован в момент t
$q_t(x)$ — вероятность, что x получил инфекцию в момент t

$$\frac{\partial}{\partial t} P(\mathtt{I}|t,x) = -\mu P(\mathtt{I}|t,x) + \beta\big(1 - P(\mathtt{I}|t,x)\big)\, q_t(x)$$

Пусть время дискретно с шагом 1. Конечно-разностный аналог:

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

Максимизация правдоподобия для оценивания параметров:

$$\sum_{t,x} [y_t(x) = \mathtt{I}] \ln p_t(x) + [y_t(x) \neq \mathtt{I}] \ln\big(1 - p_t(x)\big) \to \max$$

Параметры: μ, β и параметры вероятностной модели $q_t(x)$

R. Pastor-Satorras et al. Epidemic processes in complex networks. 2014

Модель логистической регрессии для вероятности $q_t(x)$

Гипотеза: вероятность, что x получил инфекцию в $[t-1, t]$

$$q_t(x) = \sigma\big(w_1 k_t(x) - w_0\big)$$

монотонно возрастает по числу его контактов

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

Недостатки этой модели — она не учитывает, что

1. контакты имеют различную вероятность передачи инфекции
2. индивиды v, контактирующие с x, имеют различную вероятность $p_t(v)$ быть инфицированными
3. при изменении состояния $y_t(x)$ индивида x должны измениться вероятности $p_t(u)$ для индивидов u, проконтактировавших с x незадолго до момента t

Модель вероятности передачи инфекции

Вместо числа контактов будем оценивать сумму вероятностей передачи инфекции по всем контактам в интервале $[t-1, t]$:

$$k_t(x) = \sum_{\langle t' \colon (x,v)\rangle} \big[t-1 \leqslant t' \leqslant t\big]\, a_{t'}(x,v)$$

Вероятность передачи инфекции при контакте $\langle t \colon (x,v)\rangle$ можно оценить с помощью логистической регрессии:

$$a_t(x,v) = \sigma\Big(-\alpha_0 + \sum_{j=1}^{m} \alpha_j f_j(t,x,v)\Big),$$

где f_j — признаки потенциальной опасности контакта, например:
— расстояние между устройствами по уровню сигнала,
— продолжительность контакта,
— совпадение доступных wifi и bluetooth маяков;
коэффициенты α_j являются параметрами модели.

Модель с рекурсивным оцениванием риска

Добавим оценки вероятностей $\tilde{p}_{t'}(v)$, что v инфицирован:

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

$$\tilde{p}_t(v) = \begin{cases} 1, & y_t(v) = \mathtt{I}; \\ p_t(v), & y_t(v) \neq \mathtt{I}. \end{cases}$$

Модель становится рекуррентной. Варианты обучения:

- брать текущие значения $p_t(v)$ с предыдущей итерации
- распространять градиент через суперпозицию функций, как в рекуррентных нейронных сетях
- промежуточный вариант: ограничивать распространение градиента небольшой глубиной рекурсии

Модель распространения риска по сети

Изменение состояния $y_t(x)\colon \mathtt{S} \to \mathtt{I}$ увеличивает оценки риска $p_{t'}(x)$ и $p_{t'}(u)$ для всех u, контактировавших с x, и цепочек контактов $x \to u \to v \to \ldots$ в недавнем прошлом $t' \in [t-d, t]$.

Индикатор, что x будет инфицирован в интервале $(t, t+d]$:

$$b_t(x) = \left[\exists t'\colon\ t < t' \leqslant t+d \ \text{ и } \ y_{t'}(x) = \mathtt{I}\right]$$

Добавим его в логистическую модель $q_t(x)$:

$$q_t(x) = \sigma\big(w_1 k_t(x) + w_2 b_t(x) - w_0\big)$$

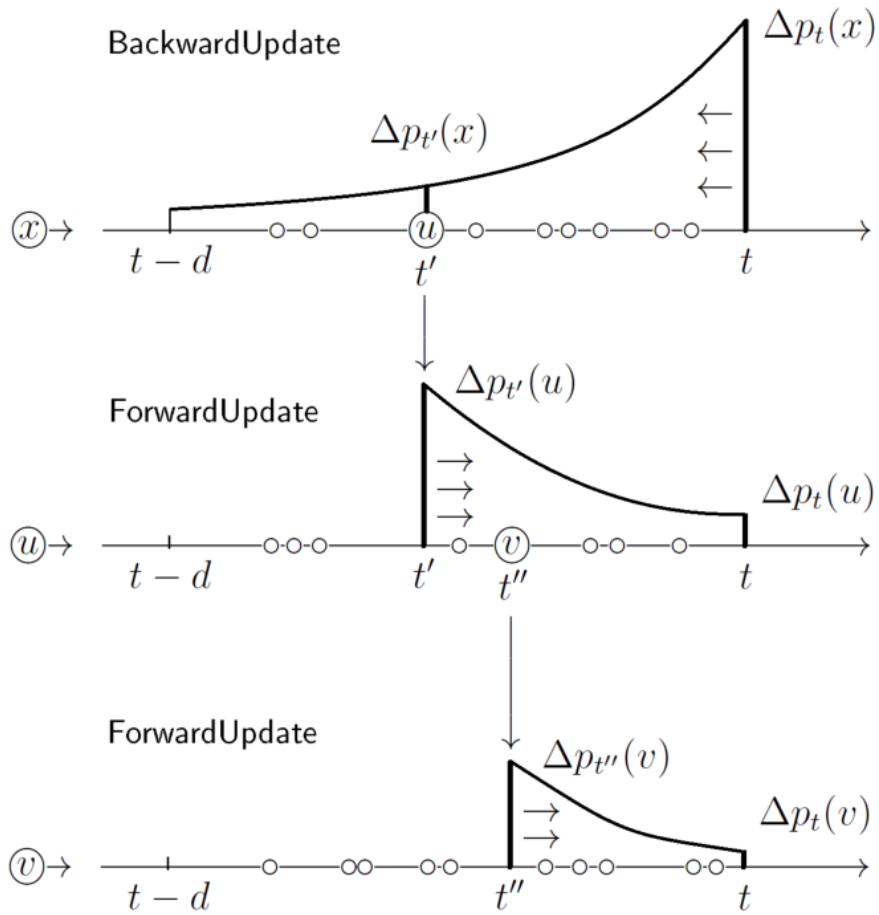
Алгоритм распространения риска по сети контактов запускается при переключении $y_t(x)\colon \mathtt{S} \to \mathtt{I}$, при этом риск скачком увеличивается до единицы, $\Delta p_t(x) = 1 - p_t(x)$, увеличиваются оценки $p_t(u)$, $p_t(v)$ и далее по цепочке контактов.

Алгоритм распространения рисков по сети контактов

Графики зависимости приращения риска Δp_t от времени t

функция **BackwardUpdate**(x, t);
$U := \varnothing$;
для всех $\langle t' \in [t-d, t] : (x, u) \rangle$
  пересчитать риск $p_{t'}(x)$;
ForwardUpdate$(x, t-d, t)$;

функция **ForwardUpdate**(x, t_0, t);
$U := U \cup \{x\}$;
для всех $\langle t' \in [t_0, t] : (x, u \notin U) \rangle$
  пересчитать риск $p_{t'}(u)$;
для всех $\langle t' \in [t_0, t] : (x, u \notin U) \rangle$
  если $\Delta p_{t'}(u) > \varepsilon$ **то**
    ForwardUpdate(u, t', t);



Модель индивидуального риска (собираем всё воедино)

Максимизация правдоподобия для оценивания параметров:

$$\sum_{t,x} [y_t(x) = \mathtt{I}] \ln p_t(x) + [y_t(x) \neq \mathtt{I}] \ln\bigl(1 - p_t(x)\bigr) \to \max$$

Модель SIS для вероятности, что x инфицирован в момент t:

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

Модель вероятности, что x получил инфекцию в $[t-1, t]$:

$$q_t(x) = \sigma\bigl(w_1 k_t(x) + w_2 b_t(x) - w_0\bigr)$$
$$k_t(x) = \sum_{\langle t' : (x,v)\rangle} \bigl[t - 1 \leqslant t' \leqslant t\bigr]\, a_{t'}(x, v)\, \tilde{p}_{t'}(v)$$
$$b_t(x) = \bigl[\exists t' : t < t' \leqslant t + d \text{ и } y_{t'}(x) = \mathtt{I}\bigr]$$

Модель вероятности передачи инфекции при контакте (x, v) в t:

$$a_t(x, v) = \sigma\bigl(-\alpha_0 + \sum_{j=1}^{m} \alpha_j f_j(t, x, v)\bigr)$$

Замечания об Алгоритме распространения рисков

Особенности метода стохастического градиента:

- индивида x выбираем случайно
- по времени t проходим последовательно, формируя сбалансированную по классам $\{\mathtt{S}, \mathtt{I}\}$ выборку
- обновляем $p_t(u)$ при фиксированных параметрах
- только полностью обработав данные по индивиду x, делаем накопленный градиентный шаг по параметрам

Возможные модификации:

- Ограничение неотрицательности коэффициентов w_k, α_j
- Учёт большего числа состояний $\mathtt{S}$, $\mathtt{E}$, $\mathtt{I}$, $\mathtt{R}$ и др.
- Кластеризация индивидов по когортам

Резюме

Мы объединили три подхода к моделированию:

- эпидемиологические модели на уровне индивидов
- машинное обучение для оценивания параметров
- алгоритмы распространения информации по графу

Новизна таких моделей в математической эпидемиологии:

- большие данные о контактах появились только недавно, в связи с пандемией Covid-19
- распространение информации в прошлое по графу контактов
- использование методов машинного обучения