Машинное обучение, ШАД

Занятие 4

Доверительные интервалы и статистические критерии

Материал по рассказан по доске

Два вспомогательных распределения

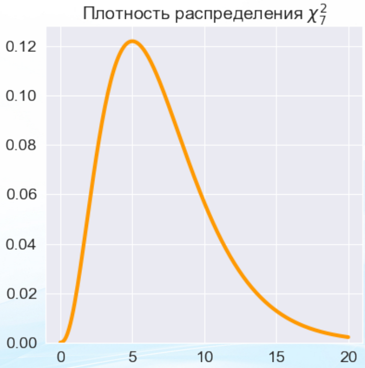
Распределение хи-квадрат

Обозначение: χ^2_k — хи-квадрат с k степенями свободы

- Параметр k — кол-во степеней свободы;
- Плотность

$$p(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2};$$

- $\chi^2_k = \Gamma(1/2, k/2)$;
- Если $\xi_1, ..., \xi_k$ — независимые $\mathcal{N}(0, 1)$, то $\xi_1^2 + ... + \xi_k^2 \sim \chi^2_k$;
- Если $\eta \sim \chi^2_k$, то $\mathsf{E}\eta = k, \mathsf{D}\eta = 2k$;
- $\chi^2_{k,p}$ — p-квантиль распределения χ^2_k.
- `scipy.stats.chi2(df=`k`)`



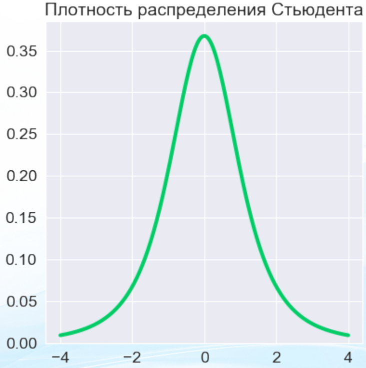
Распределение Стьюдента

Обозначение: T_k — распределение Стьюдента с k степенями свободы

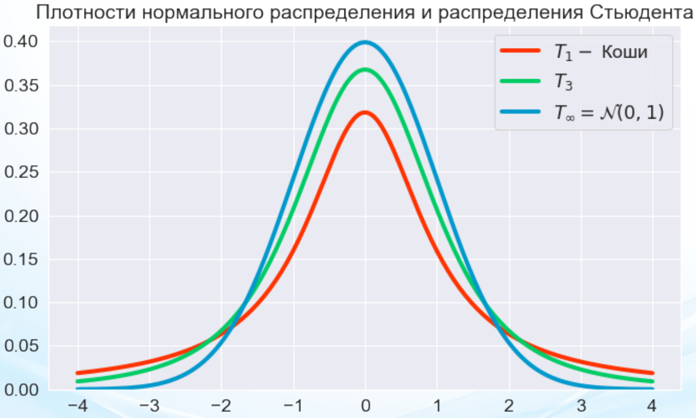
- Параметр k — кол-во степеней свободы;
- T_1 — распределение Коши
- $T_\infty = \mathcal{N}(0, 1)$
- Плотность

$$p(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{\pi k}\,\Gamma(k/2)} \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}};$$

- Если $\xi \sim \mathcal{N}(0, 1)$ и $\eta \sim \chi^2_k$ независимы, то $\zeta = \frac{\xi}{\sqrt{\eta/k}} \sim T_k$
- Если $\zeta \sim T_k$, то $\mathsf{E}\zeta = 0$ при $k > 1$
- Если $\zeta \sim T_k$, то $\mathsf{D}\zeta = \frac{k}{k-2}$ при $k > 2$
- $T_{k,p}$ — p-квантиль распределения T_k
- `scipy.stats.t(df=`k`)`



Сравнение распределений



Распределение Стьюдента похоже на нормальное, но обладает более тяжелыми хвостами, т.е. плотность медленнее убывает при $x \to \infty$.

Уильям Сили Госсет

Работал на пивоваренном заводе
Гиннеса в Дублине.

Чтобы предотвратить раскрытие конфиденциальной информации, Гиннесс запретил своим работникам публикацию любых материалов, независимо от содержавшейся в них информации.

Госсет выбрал себе псевдоним Student.

Стат. свойства линейной регрессии

Модель линейной регрессии (*повторение*)

Рассматриваем функциональную зависимость вида

$$y = y(x) = \theta_1 x_1 + \ldots + \theta_d x_d,$$

$x_1, \ldots, x_d$ — признаки,
$\theta = (\theta_1, \ldots, \theta_d)^T$ — вектор параметров.

Предполагается, что данные удовлетворяют соотношению

$$Y_i = y(x_i) + \varepsilon_i = \theta_1 x_{i1} + \ldots + \theta_d x_{id} + \varepsilon_i, \quad i = 1, \ldots, n,$$

$x_i = (x_{i1}, \ldots, x_{id})$ — признаковые описания объекта i
обычно неслучайные,
ε_i — случайная ошибка измерений.

Модель линейной регрессии *(повторение)*

Введем обозначения

$$Y = \begin{pmatrix} Y_1 \\ \dots \\ Y_n \end{pmatrix}, \qquad X = \begin{pmatrix} x_{11} & \dots & x_{1d} \\ \dots & & \\ x_{n1} & \dots & x_{nd} \end{pmatrix}, \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_n \end{pmatrix}.$$

Матричная форма записи данных

$$Y = X\theta + \varepsilon.$$

$X \in \mathbb{R}^{n \times d}$ — регрессоры или матрица признаков,

$Y \in \mathbb{R}^{n}$ — отклик.

Матричный вид зависимости: $y(x) = x^T\theta$.

Метод наименьших квадратов *(повторение)*

Задача: $\|Y - X\theta\|_2^2 \to \min\limits_{\theta \in \mathbb{R}^d}$.

Решение: $\widehat{\theta} = \left(X^T X\right)^{-1} X^T Y$ если $rank\ X = d$.

Предсказания: $\widehat{y}(x) = x^T \widehat{\theta}$.

Предположения и следствия:

1. $\mathsf{E}\varepsilon = 0$ — несмещенность:
   - $\widehat{\theta}$ — несмещенная оценка θ,
   - $\widehat{y}(x)$ — несмещенная оценка $y(x)$.
2. $\mathsf{E}\varepsilon = 0$ и $\mathsf{D}\varepsilon = \sigma^2 I_n$ — несмещенность и гомоскедастичность:
   - $\mathsf{D}\widehat{\theta} = \sigma^2 \left(X^T X\right)^{-1}$, $\mathsf{D}\widehat{y}(x) = \sigma^2 x^T \left(X^T X\right)^{-1} x$;
   - $\widehat{\sigma}^2 = \frac{1}{n-d} \left\| Y - X\widehat{\theta} \right\|_2^2$ — несмещенная оценка σ^2.
3. $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ — гауссовская линейная модель:
   - МНК совпадает с ОМП для θ.

Доверительные интервалы при $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$

Величина	**Интервал**
σ	$\left(\sqrt{\lVert Y - X\widehat{\theta} \rVert^2 \Big/ \chi^2_{n-d,\frac{1+\alpha}{2}}}, \quad \sqrt{\lVert Y - X\widehat{\theta} \rVert^2 \Big/ \chi^2_{n-d,\frac{1-\alpha}{2}}} \right)$
Дов. интервал для размера шума в отклике	
θ_j	$\left(\widehat{\theta}_j \pm T_{n-d,\frac{1+\alpha}{2}} \cdot \widehat{\sigma} \sqrt{(X^T X)^{-1}_{jj}} \right)$
Дов. интервал для коэффициента перед j-м признаком	
$x_0^T \theta$	$\left(x_0^T \widehat{\theta} \pm T_{n-d,\frac{1+\alpha}{2}} \cdot \widehat{\sigma} \sqrt{x_0^T (X^T X)^{-1} x_0} \right)$
Дов. интервал для **среднего** отклика на объекте x_0	
$x_0^T \theta + \varepsilon$	$\left(x_0^T \widehat{\theta} \pm T_{n-d,\frac{1+\alpha}{2}} \cdot \widehat{\sigma} \sqrt{1 + x_0^T (X^T X)^{-1} x_0} \right)$
Предск. интервал для **наблюдаемого** отклика на объекте x_0	

Число α — уровень доверия, обычно $\alpha = 0.95$

Значим ли признак x_j?

Гипотеза о **не**значимости коэффициента θ_j

$\mathrm{H}_0\colon \theta_j = 0 \;\; vs. \;\; \mathrm{H}_1\colon \theta_j \; \{<, \neq, >\} \; 0$

Критерий Стьюдента (t-test)

$$T_j(X, Y) = \frac{\widehat{\theta}_j}{\widehat{\sigma}\sqrt{(X^T X)^{-1}_{jj}}} \overset{\mathrm{H}_0}{\sim} T_{n-d},$$

где $T_j(X, Y)$ — t-статистика критерия.

Для $\mathrm{H}_1\colon \theta_j \neq 0$ критерий имеет вид $\left\{|T_j(X, y)| > T_{n-d, \frac{1-\alpha}{2}}\right\}$,

где число α — уровень значимости, обычно $\alpha = 0.05$.

Если H_0 не отвергается, то можно считать,
что θ_j отклоняется от нуля статистически незначимо.

Пример: *AB-тест*

Пользователи делятся случайно на две независимые группы:

1. *Контрольная группа A* — использует **старую** ML-модель;
   $Y_1 = (Y_{11}, ..., Y_{n1}), \mathsf{E} Y_{i1} = a$ — результаты.
2. *Исследуемая группа B* — использует **новую** ML-модель;
   $Y_2 = (Y_{12}, ..., Y_{m2}), \mathsf{E} Y_{i2} = b$ — результаты.

Что может быть результатом?

- Сумма покупки.
- Клик по рекламе.
- Длина сессии.
- Регистрация пользователя на сервисе.
- и т.д.

Гипотезы:

$\mathsf{H}_0 : a = b$ — отсутствие эффекта,
$\mathsf{H}_1 : a \neq b$ — эффект присутствует.

Сведем линейной регрессии

Введем обозначения

$$Y = \begin{pmatrix} Y_{11} \\ \dots \\ Y_{n1} \\ Y_{12} \\ \dots \\ Y_{m2} \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 0 \\ \dots & \\ 1 & 0 \\ 1 & 1 \\ \dots & \\ 1 & 1 \end{pmatrix}, \quad \theta = \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_{n+m} \end{pmatrix}.$$

Линейная регрессия предполагает зависимость $Y = X\theta + \varepsilon$.

Тогда $Y_{i1} = \theta_0 + \varepsilon_i$ и $Y_{i2} = \theta_0 + \theta_1 + \varepsilon_{n+i}$, следовательно

- θ_0 — среднее в группе A,
- θ_1 — **эффект от эксперимента**.

Вывод: для проверки AB-теста нужно построить интервал для θ_1 и проверить гипотезу $\mathsf{H}_0 \colon \theta_1 = 0$ критерием Стьюдента.

Учет признаков: аналог CUPED в AB-тестах

Если есть дополнительные признаки, например, сумма прошлых покупок, возраст пользователя, его пол, можно учесть эти признаки

$$Y = \begin{pmatrix} Y_{11} \\ \dots \\ Y_{n1} \\ Y_{12} \\ \dots \\ Y_{m2} \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 0 & \overrightarrow{x_{11}} \\ \dots & & \\ 1 & 0 & \overrightarrow{x_{n1}} \\ 1 & 1 & \overrightarrow{x_{12}} \\ \dots & & \\ 1 & 1 & \overrightarrow{x_{m2}} \end{pmatrix}, \quad \theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \overrightarrow{\theta_2} \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_{n+m} \end{pmatrix}.$$

Линейная регрессия предполагает зависимость $Y = X\theta + \varepsilon$.

Тогда $Y_{i1} = \theta_0 + \overrightarrow{x_{i1}}^T \overrightarrow{\theta_2} + \varepsilon_i$ и $Y_{i2} = \theta_0 + \theta_1 + \overrightarrow{x_{i1}}^T \overrightarrow{\theta_2} + \varepsilon_{n+i}$, след.

- θ_0 — среднее в группе A,
- θ_1 — **эффект от эксперимента** при *фиксированных* остальных факторах.

Вывод: для проверки AB-теста нужно построить интервал для θ_1 и проверить гипотезу $\mathsf{H}_0\colon \theta_1 = 0$ критерием Стьюдента.

Гетероскедастичность
и анализ остатков

Остатки

$\mathrm{D}\varepsilon = \sigma^2 I_n$ — гомоскедастичность. Обратное — гетероскедастичность.

В качестве оценки ошибки ε_i рассмотрим остатки $e_i = Y_i - \widehat{Y}_i$

Проблема: $\mathrm{D}e_i \neq \sigma^2$ при гомоскедастичности. Покажем это.

Вектор остатков $e = Y - \widehat{Y} = (I_n - H)Y, \quad$ где $H = X\left(X^T X\right)^{-1} X^T$;

$$\mathrm{D}e = (I_n - H)\mathrm{D}Y(I_n - H)^T = \sigma^2(I_n - H)(I_n - H)^T = \sigma^2(I_n - H),$$

$$\text{где } HH^T = X\left(X^T X\right)^{-1} X^T X \left(X^T X\right)^{-1} X^T = X\left(X^T X\right)^{-1} X^T = H.$$

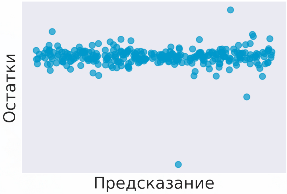
Проверять на однородность дисп. нужно **поправленные остатки**:

$$\widehat{e}_i = \frac{e_i}{\sqrt{\widehat{\mathrm{D}e_i}}} = \frac{e_i}{\sqrt{\frac{\|Y - X\widehat{\theta}\|^2}{n-d}(1 - H_{ii})}} \text{ — стюдентизированные остатки}$$

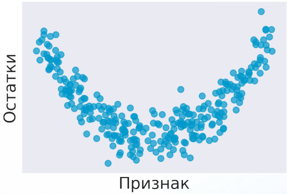
Визуальный анализ

Строятся графики зависимости $\widehat{e}_i$ от y, x, i

Все хорошо

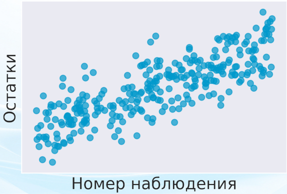


Есть выбросы



Нужно добавить x^2

Гетероскедастичность

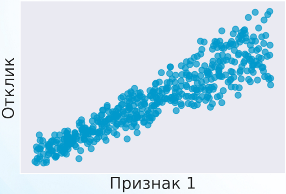


Тренд

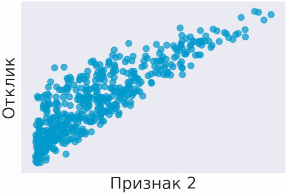
Неправильная модель

Визуальный анализ

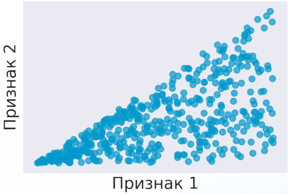
Что будет если строить графики зависимостей таргета от признаков:



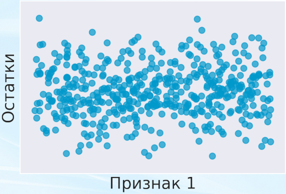
Гетероскедастичность?



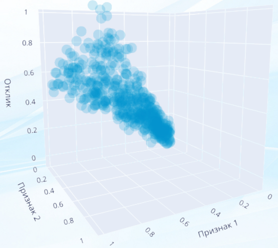
Гетероскедастичность?



Признаки зависимы



Нет, все хорошо!



Что делать при гетероскедастичности?

- Если нужна только оценка θ — ничего;
- Если есть предположения о природе гетероскедастичности, взвесить наблюдения:
$$Y_i/\widehat{\sigma}_i = (x_i/\widehat{\sigma}_i)^T \theta + \varepsilon_i,$$
где $\widehat{\sigma}_i$ — предполагаемая дисперсия при i-м измерении;
- Преобразование признаков и отклика, напр., Бокса-Кокса:
$$Z_i = \begin{cases} \ln Y_i, & \lambda = 0 \\ (Y_i^\lambda - 1)/\lambda, & \lambda \neq 0 \end{cases}$$
Величина λ подбирается по графику зависимости MSE от λ
- Использовать специальные оценки дисперсии, устойчивые к гетероскедастичности.

Устойчивые оценки дисперсии

Пусть $\mathsf{E}\varepsilon = 0$ и $\mathsf{D}\varepsilon = V$.

Тогда $\Sigma = \mathsf{D}\widehat{\theta} = \left(X^T X\right)^{-1} X^T V X \left(X^T X\right)^{-1}$.

1. $V = \sigma^2 I_n$ — гомоскедастичность:
   $\Sigma = \sigma^2 \left(X^T X\right)^{-1}$ — дисперсия оценки коэффициентов;
   $\widehat{\Sigma} = \widehat{\sigma}^2 \left(X^T X\right)^{-1}$ — оценка дисперсии оценки коэффициентов;
2. $V = \operatorname{diag}\left(\sigma_1^2, \dots, \sigma_n^2\right)$ — отсутствие автокорреляций:
   $\Sigma = \left(X^T X\right)^{-1} X^T \cdot \operatorname{diag}\left(\sigma_1^2, \dots, \sigma_n^2\right) \cdot X \left(X^T X\right)^{-1}$ — д.о.к.;
   $\widehat{\Sigma} = \left(X^T X\right)^{-1} X^T \cdot \operatorname{diag}\left(\widehat{\sigma}_1^2, \dots, \widehat{\sigma}_n^2\right) \cdot X \left(X^T X\right)^{-1}$ — о.д.о.к..
3. Наличие автокорреляций — более сложный случай, при котором зависимы элементы выборки.
   Ипользуются кластерное представление ковариационной матрицы или модели временных рядов.

Устойчивые оценки дисперсии

Если автокорреляции отсутствуют, используются **оценка Уайта**

White's heteroscedasticity-consistent estimator (HCE)

$$\widehat{\Sigma} = \left(X^T X\right)^{-1} X^T \cdot \mathrm{diag}\left(\widehat{\sigma}_1^2, \ldots, \widehat{\sigma}_n^2\right) \cdot X \left(X^T X\right)^{-1}$$

Варианты определения $\widehat{\sigma}_i^2$

1. HC0: $\widehat{e}_i^2$ — оценка Уайта
2. Модификации МакКиннона-Уайта

$$\text{HC1:}\quad \frac{n}{n-d}\widehat{e}_i^2, \qquad \text{HC2:}\quad \frac{\widehat{e}_i^2}{1-H_{ii}}, \qquad \text{HC3:}\quad \frac{\widehat{e}_i^2}{(1-H_{ii})^2}$$

Точнее оценивают при малых выборках.

Как ее применять?

Если автокорреляции отсутствуют, то выполнена асимптотическая нормальность оценки коэффициентов

$$\sqrt{n}\left(\widehat{\theta} - \theta\right) \xrightarrow{d} \mathcal{N}(0, B),$$

НСЕ дает хорошую (состоятельную) оценку на матрицу B:

$$n\widehat{\Sigma} \xrightarrow{\mathsf{P}} B$$

Данный факт позволяет строить асимптотические дов. интервалы для коэффициентов моделей и таргета, а также критерий Вальда для проверки линейных гипотез $\mathsf{H}_0 : T\theta = \tau$.

Обобщенная модель линейной регрессии

Вспомним простую линейную регрессию с норм. шумом

Данные $Y = X\theta + \varepsilon \sim \mathcal{N}(X\theta, \sigma^2 I_n)$

Строим модель вида $y(x) = x^T\theta$.

Что мы предсказываем?

Пусть x_0 новый объект.

Тогда $Y_0 = x_0^T\theta + \varepsilon_0 \sim \mathcal{N}(x_0^T\theta, \sigma^2)$.

Т.е. в качестве предсказания оцениваем $\mathsf{E} Y_0$ — *ожидаемый отклик*.

Итог

$y \in \mathbb{R}$ — значения наблюдаемого отклика

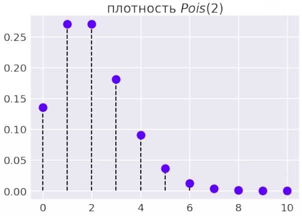
$\mathsf{E}_x Y$ — ожидаемый отклик

$Y_i \sim \mathcal{N}(x_i^T\theta, \sigma^2)$ — наблюдаемый отклик

Пуассоновское распределение

$$Pois(\lambda): \quad p(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \; x \in \mathbb{Z}_+$$

Смысл: число событий, произошедших за единицу времени



Условия:

1. события происходят с фиксированной интенсивностью λ.
2. независимо друг от друга.

Утверждение: время между двумя событиями имеет распр. $Exp(\lambda)$

Примеры:

1. число клиентов в час
2. число запросов на сервер за минуту

События разной интенсивности

Интенсивность может зависеть от каких-то факторов.

$X_1, ..., X_n$ — факторы интервала времени

$Y_1, ..., Y_n$ — число событий, произошедших за интервал времени

Тем самым имеется $\lambda(x)$ — интенсивность событий для факторов x.

Получаем $Y_i \sim Pois(\lambda(x_i))$

Что предсказывать?

Нет смысла предсказывать сам Y_i,

т.к. помимо $\lambda(x_i)$ он содержит непрогнозируемый шум.

Тогда оценим $\mathsf{E} Y_i = \lambda(x_i)$ — *ожидаемый отклик*.

Как параметризовать $\lambda(x)$ для *линейной* модели?

Определимся с требованиями

Пусть значению $x_0^T\theta = 0$ соответствует интенсивность $\lambda_0 = 1$.
Значению $x_1^T\theta$ сопоставим интенсивность $\lambda_1 = 5$ событий в час.
Хотим чтобы значению $-x_1^T\theta$ соответствовала интенсивность $1/\lambda_1 = 0.2$ событий в час.

Линеаризация

Соответственно, нужно взять $\lambda_\theta(x) = \exp(x^T\theta)$.

Тогда $\ln \lambda_\theta(x) = x^T\theta$.

$g(z) = \ln z$ — **линеаризация** ожидаемого отклика.

Итог

$y \in \mathbb{Z}_+$ — значения наблюдаемого отклика

$\lambda_\theta(x) = \mathsf{E}_x Y$ — ожидаемый отклик

$Y_i \sim Pois(\lambda_\theta(x_i))$ — наблюдаемый отклик

Это **пуассоновская регрессия**.

Случай бинарной классификации

$X_1, ..., X_n$ — признаки объекта

$Y_1, ..., Y_n$ — бинарный класс

Тем самым имеется $\rho(x)$ — вероятность класса 1 для объекта x.

Получаем $Y_i \sim Bern(\rho(x_i))$

Что предсказывать?

Оцениваем $\mathsf{E}Y = \rho(x)$ — *ожидаемый отклик*.

Как параметризовать $\rho(x)$ для *линейной* модели?

Определимся с требованиями

Пусть значению $x_0^T\theta = 0$ соответствует вероятность 0.5.

Значению $x_1^T\theta$ сопоставим вероятность $\rho_0 = 0.9$.

Хотим чтобы значению $-x_1^T\theta$ соответствовала вероятность 0.1.

Что такое в 2 раза более/менее вероятно?

Возможно, 0.95 и 0.05, но это не точно.

Сведение к параметру масштаба

Заметим, что $\frac{\rho}{1-\rho}$ — насколько чаще выпадает класс 1

по сравнению с классом 0. Тем самым это параметр масштаба.

Линеаризация

Логит $g(z) = \ln\frac{z}{1-z}$ — линеаризация ожидаемого отклика.

Т.е. параметризация $\rho_\theta(x)$ должна быть такой, что $\ln\frac{\rho_\theta(x)}{1-\rho_\theta(x)} = x^T\theta$.

Тогда нужно взять $\rho_\theta(x) = \frac{1}{1+\exp(x^T\theta)}$.

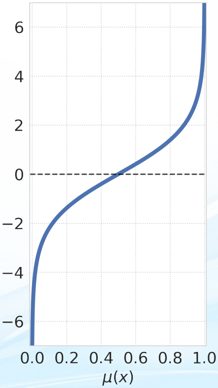
Бинарный отклик

$y \in \{0, 1\}$ — значения наблюдаемого отклика

$\rho_\theta(x) = \mathsf{P}_x(Y = 1)$ — ожидаемый отклик

$Y_i \sim Bern(\rho_\theta(x_i))$ — наблюдаемый отклик

Это **логистическая регрессия**.



Обобщенная модель линейной регрессии

Гауссовская линейная модель

Ожидаемый отклик:

$y = \mu_\theta(x) = x^T\theta.$

Наблюдаемый отклик:

$Y_i = x_i^T\theta + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$

или $Y_i \sim \mathcal{N}(\mu_\theta(x_i), \sigma^2)$

Оценка отклика:

$\widehat{y} = x^T\widehat{\theta}.$

Generalized Linear Models (GLM)

Ожидаемый отклик:

$y = \mu_\theta(x)$, причем $g(\mu_\theta(x)) = x^T\theta$,

т.е. g — линеаризация ожид. отклика

Наблюдаемый отклик:

$Y_i \sim \mathsf{P}_{\mu_\theta(x_i)}$,

где $\{\mathsf{P}_\psi \mid \psi \in \Psi\}$ — семейство распр.

Оценка отклика:

$\widehat{y} = g^{-1}\left(x^T\widehat{\theta}\right).$

Свойства GLM

В качестве $\widehat{\theta}$ берется ОМП (ищется численно)

$$L_X(\theta) = \prod_{i=1}^{n} p_{\mu_\theta(x_i)}(Y_i) = \prod_{i=1}^{n} p_{g^{-1}(x_i^T\theta)}(Y_i) \longrightarrow \max_\theta$$

Если $\{\mathsf{P}_\psi \mid \psi \in \Psi\}$ лежит в экспоненциальном классе, то $\widehat{\theta}$:

1. существует и единственна;
2. асимптотически нормальна: $\sqrt{I(\theta)}\left(\widehat{\theta} - \theta\right) \xrightarrow{d} \mathcal{N}(0, I_d)$,

   где $I(\theta) = \left(-\mathsf{E}\frac{\partial^2 \log L_X(\theta)}{\partial\theta_j \partial\theta_k}\right)_{jk}$ — информационная матрица Фишера.

Частные случаи:

1. Линейная (гауссовская): $I(\theta) = \sigma^{-2} X^T X$.
2. Логистическая: $I(\theta) = X^T \cdot \text{diag}\left[\sigma\left(x_i^T\theta\right)\left(1 - \sigma\left(x_i^T\theta\right)\right)\right] \cdot X$.
3. Пуассоновская: $I(\theta) = X^T \cdot \text{diag}\left[\exp\left(x_i^T\theta\right)\right] \cdot X$.

Примечание. Свойства работают если верны предположения модели.

Асимпт. доверительные интервалы в GLM

Для параметров ($\Longrightarrow$ критерий для гипотезы $\mathsf{H}_0 : \theta_j = 0$)

$$\theta_j \in \left(\widehat{\theta}_j \pm z_{1-\alpha/2}\sqrt{\left(I^{-1}(\widehat{\theta})\right)_{jj}}\right)$$

Для преобразованного ожидаемого отклика

$$x_0^T\theta \in \left(x_0^T\widehat{\theta} - \delta, x_0^T\widehat{\theta} + \delta\right)$$

Для ожидаемого отклика

$$\mu(x_0) = g^{-1}(x_0^T\theta) \in \left[g^{-1}\left(x_0^T\widehat{\theta} - \delta\right), g^{-1}\left(x_0^T\widehat{\theta} + \delta\right)\right],$$

Например, в случае лог. регрессии можем не просто говорить, что вероятность болезни 20%, но и оценивать интервал для нее: вероятность болезни от 16% до 27%.

Обозначения: $\delta = z_{1-\alpha/2}\sqrt{x_0^T I^{-1}(\widehat{\theta})x_0}$, а z_α — α-квантиль $\mathcal{N}(0,1)$.

Примечание. Для линейной регрессии вместо σ^2 нужно взять ее несмещ. оценку.


ВСЁ!