Лекція 9
Лінійні моделі: статистичний погляд

§40 Задача регресії

Історія терміна

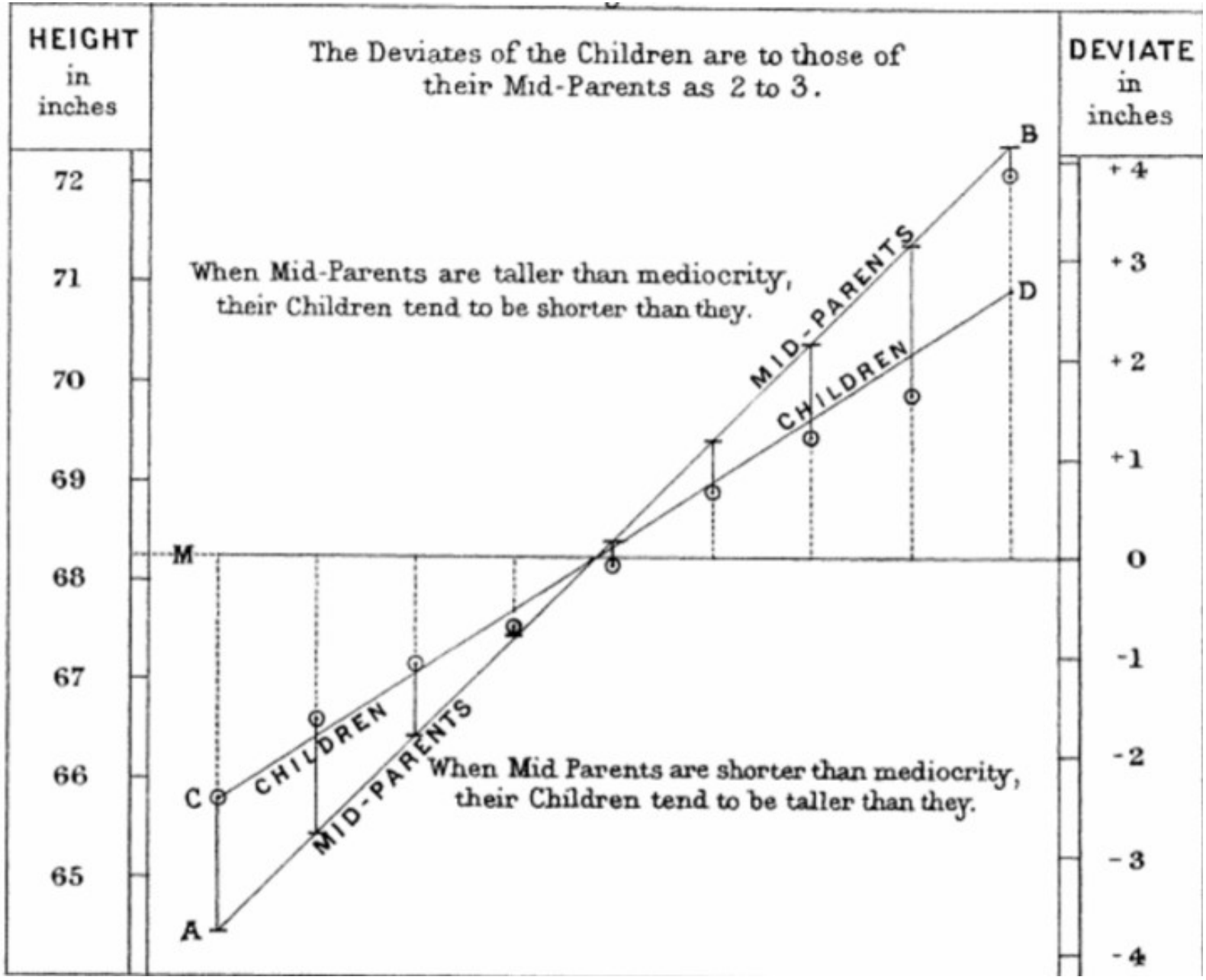
Френсис Гальтон досліджував залежність між середнім ростом дітей і середнім ростом батьків (930 сімей).

1) **Виявив, середній рост батьків і дітей приблизно однаковий 68,2** дюйми (173 см).

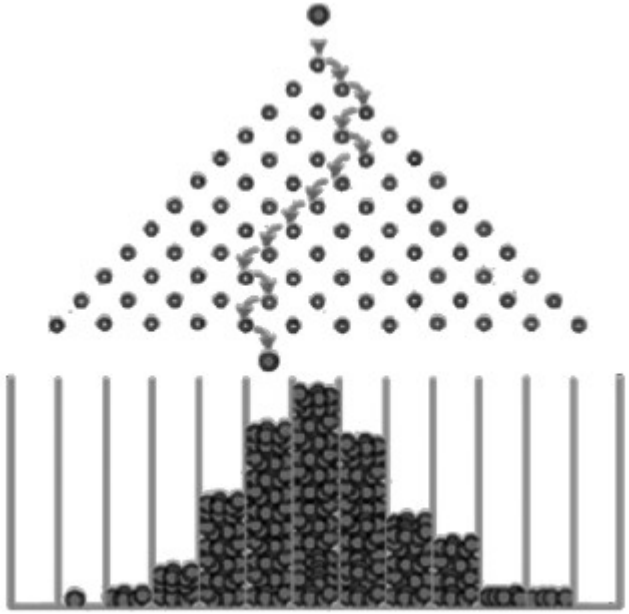
2) Він розглянув **групи батьків**, наприклад, рост яких знаходився у проміжку від **70 до 71 дюйма (відстань від середнього 70,5-68,2=2,3)**, розглянув рост **ЇХ** дітей і виявив що він дорівнює **69,5** дюймам **(відстань від середнього 69,5-68,2 =1,3)**. Отже, **рост дітей** таких батьків відрізнявся від середнього росту всіх дітей на меншу величину (**1,3 дюйми**), ніж рост їх батьків від середнього росту всіх батьків (**2,3 дюйми**), тобто, відбувалася **регресія (зменшення) цього показника**.

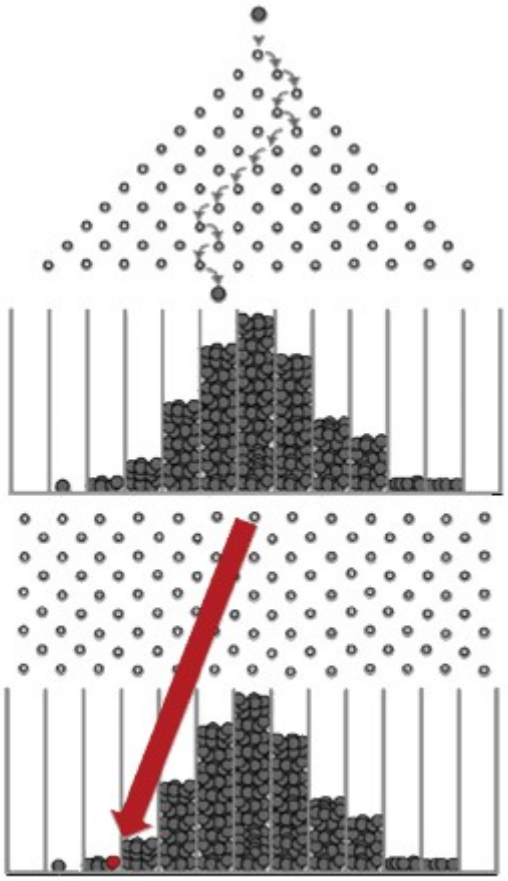
Він далі досліджує інші групи батьків (наприклад, рост яких лежить в інтералі від 71 до 72 дюйми) та рост їх дітей. Результати подає на графіку. Відклає за **горизонтальною шкалою середній рост різних груп батьків, на вертикальній осі відхилення від середнього росту цієї групи батьків та дітей цієї групи батьків.** Побачив, що експериментальні дані, які стосуються дітей **склалися в практично пряму лінію.**

Згодом лінія, яка відповідає дітям була названа лінією РЕГРЕСІЇ.


HEIGHT in inches
72
71
70
69
68
67
66
65
The Deviates of the Children are to those of their Mid-Parents as 2 to 3.
When Mid-Parents are taller than mediocrity, their Children tend to be shorter than they.
When Mid Parents are shorter than mediocrity, their Children tend to be taller than they.
MID-PARENTS
CHILDREN
CHILDREN
MID-PARENTS
A
B
C
D
M
DEVIATE in inches
+4
+3
+2
+1
0
-1
-2
-3
-4

Дошка Гальтона





Пояснення регресії до середнього

Внизу формується **така сама гауссіана**. Якщо зафіксувати якусь конкретну кулька в нижній половині ближче до краю, то виявиться, що з досить великою ймовірністю **ця кулька прийшла не з комірки, яка знаходиться у верхній половині прямо над коміркою, в якій він опинився внизу, а від комірки яка ближче до середини**. Це відбувається просто тому, що **в середині кульок більше**.

Ефект регресії до середнього проявляється в багатьох практичних задачах.

Приклад, якщо дати студентам дуже складний тест, більшу роль у тому, наскільки добре вони його пройдуть, будуть грати не тільки їх знання з предмету, але й везіння, тобто **випадковий фактор**. Тому, **якщо ізолювати 10% студентів, які пройшли тест краще всіх (набрали найбільше балів) і дати їм ще один варіант тесту, то середній бал у цій групі швидше за все впаде.** Просто тому що люди, яким повезло в перший раз, швидше за все вже не будуть так щасливі в другий - у цьому й складається ефект регресії до середини.

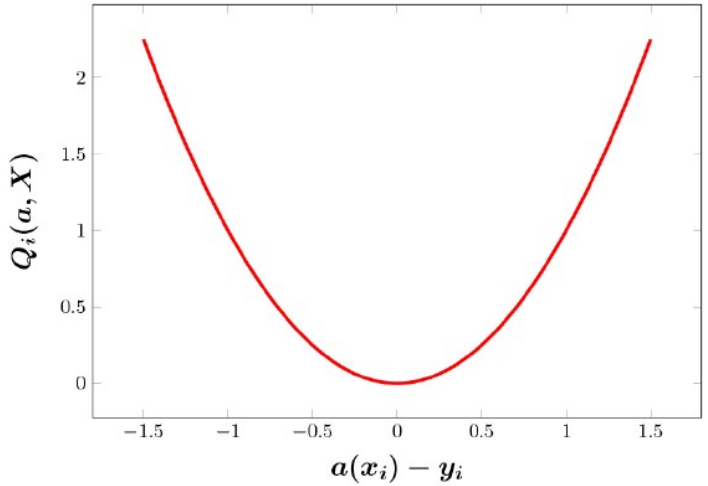
Регресія

Найчастіше під **регресією розуміють залежність** $a(x)$**, яка найкраще описує величини** y**,** тобто отримується мінімізацію середньоквадратичної похибки: квадратів відхилень відкликів y від їх прогнозних значень $a(x)$.

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

$$a_*(x) = \underset{a}{\operatorname{argmin}}\, Q(a, X)$$

(Метод найменших квадратів (МНК))



Лінійна регресія

$$Q(\mathrm{w}, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle \mathrm{w}, x_i \rangle - y_i)^2$$

$$\mathrm{w}_*(x) = \operatorname*{argmin}_{\mathrm{w}} Q(\mathrm{w}, X)$$

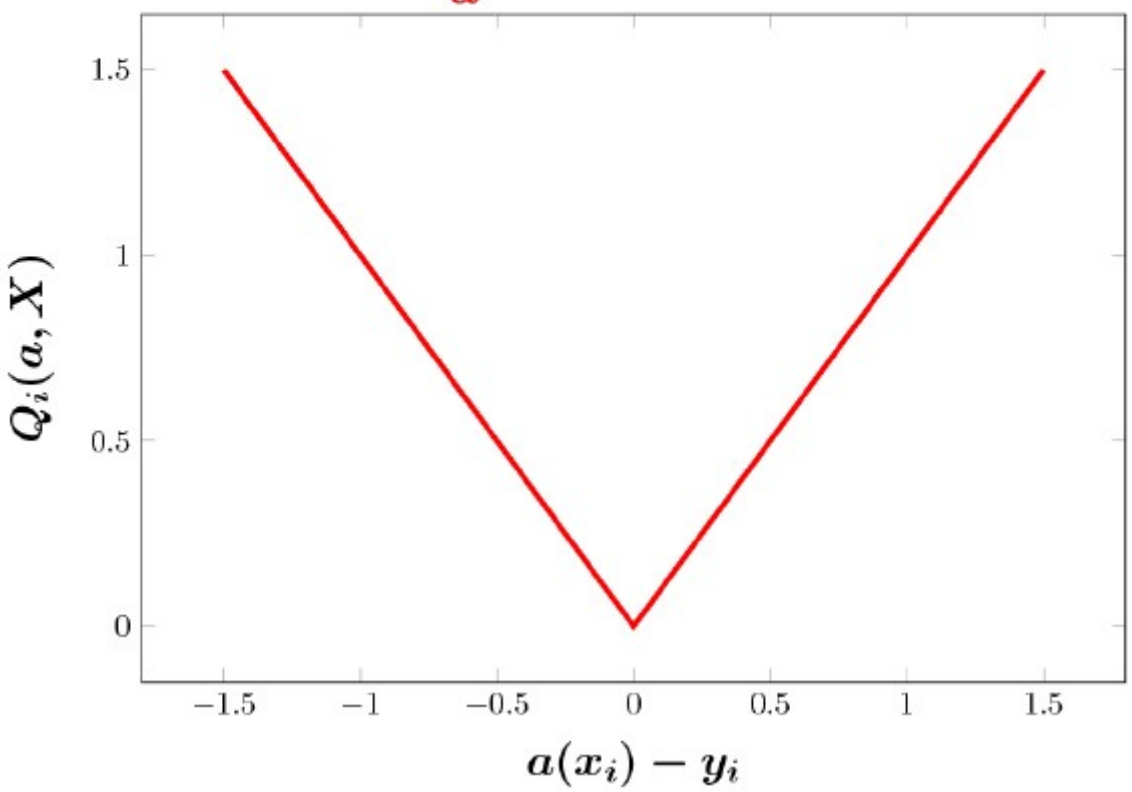
Має аналітичний розв’язок

$$\mathrm{w}_*(x) = \operatorname*{argmin}_{\mathrm{w}} Q(\mathrm{w}, X) = (X^T X)^{-1} X^T y$$

Середня абсолютна похибка (квантильна регресія)

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} |a(x_i) - y_i|$$

$$a_*(x) = \underset{a}{\operatorname{argmin}}\, Q(a, X)$$



§41 Метод максимізації правдоподібності

Нехай X — деяка випадкова величина з функцією розподілу $F(X,\theta)$, що залежить від невідомого параметра θ, а $X^n = (X_1,...,X_n)$ — вибірка розміру n, яка згенерована з розподілу $F(X,\theta)$. Необхідно оцінити за цією вибіркою невідомий параметр θ.

Метод максимізації правдоподібності: приклад

Дані про кількість смертей кавалеристів у результаті загибелі під ними коня

Кільк. загиблих	0	1	2	3	4	5	Усього
Кільк. повідомлень	109	65	22	3	1	0	200

Випадкова величина – лічильник, її необхідно моделювати розподілом Пуассона:

$$P(X=k)=\frac{\lambda^k e^{-\lambda}}{k!}$$

$\lambda - ?$

Вибірка **складається з незалежних**, однаково розподілених випадкових величин,

$$P(X = X_i) = \frac{\lambda^{X_i} e^{-\lambda}}{X_i!}$$

.

Імовірність одержання строго певної вибірки дорівнює добутку ймовірностей одержання кожного з елементів цієї вибірки

$$P(X^n, \lambda) = \prod_{i=1}^{n} \frac{\lambda^{X_i} e^{-\lambda}}{X_i!} \equiv L(X^n, \lambda)$$

Функція L залежить від невідомого параметра λ і називається **ПРАВДОПОДІБНІСТЮ ВИБІРКИ**. **Як оцінку для λ можна взяти таке значення, що максимізує функцію правдоподібності**:

$$\hat{\lambda}_{\text{ОМП}} = \underset{\lambda}{\operatorname{argmax}}\, L(X^n, \lambda)$$

Ця оцінка називається **оцінкою максимальної правдоподібності**.

В розглянутій задачі

$$L=\left(\frac{\lambda^0 e^{-\lambda}}{0!}\right)^{109}\cdot\left(\frac{\lambda^1 e^{-\lambda}}{1!}\right)^{65}\cdot\left(\frac{\lambda^2 e^{-\lambda}}{2!}\right)^{22}\cdot\left(\frac{\lambda^3 e^{-\lambda}}{3!}\right)^{3}\cdot\left(\frac{\lambda^4 e^{-\lambda}}{4!}\right)^{1}\cdot\left(\frac{\lambda^5 e^{-\lambda}}{5!}\right)^{0}=$$

$$=\left(\frac{1}{0!}\right)^{109}\cdot\left(\frac{1}{1!}\right)^{65}\cdot\left(\frac{1}{2!}\right)^{22}\cdot\left(\frac{1}{3!}\right)^{3}\cdot\left(\frac{1}{4!}\right)^{1}\cdot\left(\lambda^{0\cdot 109+1\cdot 65+2\cdot 22+3\cdot 3+4\cdot 1}e^{-\lambda\cdot 200}\right).$$

$$\ln(L)=\ln\left[\left(\frac{1}{0!}\right)^{109}\cdot\left(\frac{1}{1!}\right)^{65}\cdot\left(\frac{1}{2!}\right)^{22}\cdot\left(\frac{1}{3!}\right)^{3}\cdot\left(\frac{1}{4!}\right)^{1}\right]+$$

$$+\left(0\cdot 109+1\cdot 65+2\cdot 22+3\cdot 3+4\cdot 1\right)\ln\lambda-\lambda\cdot 200.$$

Берем похідну

$$\frac{d(\ln L)}{d\lambda}=+\left(0\cdot 109+1\cdot 65+2\cdot 22+3\cdot 3+4\cdot 1\right)\frac{1}{\lambda}-200=0.$$

Таким чином,

$\hat{\lambda}_{\text{ОМП}}=\overline{X}_n=0{,}61=(0\cdot 109+1\cdot 65+2\cdot 22+3\cdot 3+4\cdot 1)/200$ (кількість загиблих на кількість повідомлень)

Метод максимізації правдоподібності: загальний вигляд

Нехай X — деяка випадкова величина з функцією розподілу $F(X,\theta)$, що залежить від невідомого параметра θ, а $X^n = (X_1,...,X_n)$ — вибірка розміру n, яка згенерована з розподілу $F(X,\theta)$. Необхідно оцінити за цією вибіркою невідомий параметр θ.

Тоді функція правдоподібності має вигляд:

$$L(X^n,\theta) \equiv \prod_{i=1}^{n} P(X = X_i,\theta)$$

Оскільки при логарифмуванні не змінюються положення максимумів функції,

$$\ln L(X^n,\theta) = \sum_{i=1}^{n} \ln P(X = X_i,\theta)$$

Оцінкою максимальної правдоподібності називається величина:

$$\hat{\theta}_{\text{ОМП}} = \underset{\theta}{\operatorname{argmax}} \ln L(X^n,\theta)$$

У випадку неперервної випадкової величини метод максимальної правдоподібності записується аналогічно:

$$L(X^n, \theta) \equiv \prod_{i=1}^{n} f(X_i, \theta)$$

$$\hat{\theta}_{\text{ОМП}} = \underset{\theta}{\operatorname{argmax}}\, L(X^n, \theta)$$

Властивості методу максимальної правдоподібності

- Обґрунтованість, тобто одержані оцінки при збільшенні об'єму вибірки починають прямувати до істинних значень:

$$\text{коли } n \to \infty \text{ то } \hat{\theta}_{\text{ОМП}} \to \theta$$

- Асимптотична нормальність, тобто зі зростанням об'єму вибірки, оцінки максимальної правдоподібності усе краще описуються нормальним розподілом із середнім, яке дорівнює істинному значенню θ, і дисперсією, яка дорівнює величині, що зворотна до інформації Фішера:

$$n \to \infty \quad \hat{\theta}_{\text{ОМП}} \sim N(\theta, I^{-1}(\theta))$$

§42 Регресія як максимізація правдоподібності

Модель шуму: нормальний розподіл

З'ясуємо, **що ми отримуємо, коли ми мінімізуємо середню квадратичну похобку?**

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

$$a_*(x) = \operatorname*{argmin}_a Q(a, X)$$

При вирішенні задачі регресії значення цільової функції можна записати у вигляді

$$y = a(x) + \varepsilon$$

ε – випадковий шум

Якщо цей випадковий шум **має нормальний розподіл** з нульовим середнім і дисперсією σ^2, виявляється, що задача мінімізації середньоквадратичної похибки

$$a_*(x) = \operatorname*{argmin}_a \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

дає оцінку максимальної правдоподібності для регресійної функції $a(x)$.

Тобто два різних підходи дають один і той же результат.

Висновок

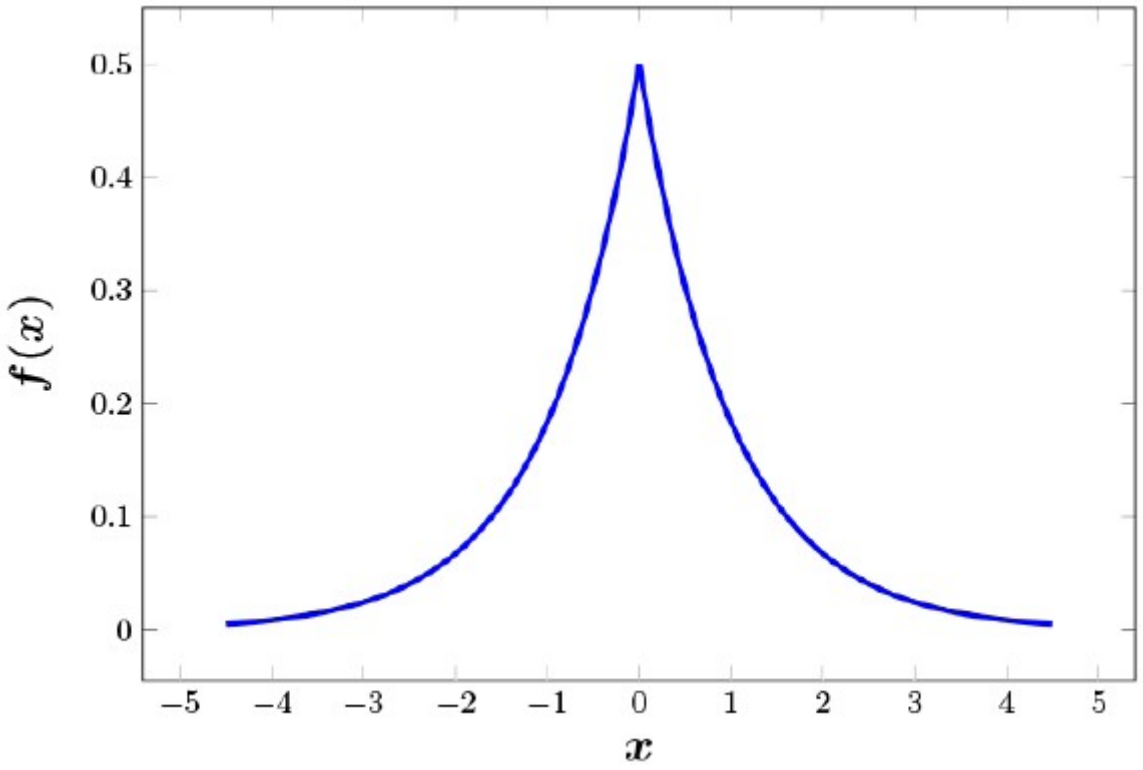
Використання в задачі з регресії властивостей методу максимальної правдоподібності дозволяє:

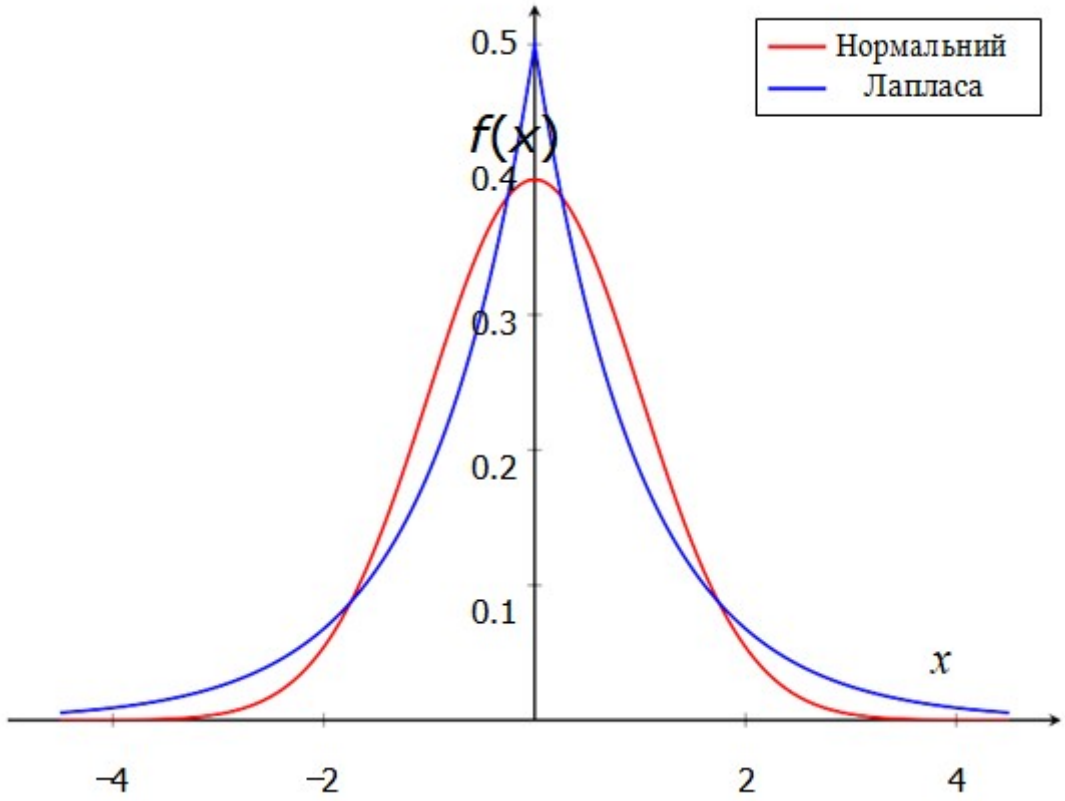
- використовуючи асимптотичну нормальність, можна визначати значимість ознак x^j у моделі й робити їх відбір.
- можна будувати довірчі інтервали для значення відгуку на нових об'єктах, яких немає в навчальній вибірці.

Модель шуму: розподіл Лапласа

Розподіл шуму **НЕ** обов'язково повинен бути нормальним і може бути якимсь іншим. Наприклад, можна спробувати описати його **розподілом Лапласа з нульовим середнім**:

$$f(x) = \frac{\alpha}{2} e^{-\alpha|x|}$$





Порівняно з нормальним розподілом, **розподіл Лапласа має більш важкі хвости**, тобто для нього більш ймовірні більші значення ε. Інакше кажучи, якщо моделювати шум розподілом Лапласа, то спостереження можуть сильніше відхилятися від обраної моделі. За рахунок цього **отримуємо рішення, що більше стійке до викидів**.

Виявляється, що якщо шум дійсно описується розподілом Лапласа, то **до оцінки максимальної правдоподібності** $a(x)$ приводить до мінімізація середніх абсолютних відхилень:

$$y = a(x) + \varepsilon$$

$$a_*(x) = \operatorname*{argmin}_a \frac{1}{\ell} \sum_{i=1}^{\ell} |a(x_i) - y_i|$$

Висновок:

- Регресія з методом найменшиї квадратів дає оцінку максимальної правдоподібності для $a(x)$, коли шум нормальний
- регресія зі середньою абсолютною похибкою дає оцінку максимальної правдоподібності для $a(x)$, коли шум лапласівський

§43 Регресія як оцінка середнього

Середньоквадратична похибка

Сутність методу найменших квадратів

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

$$a_*(x) = \operatorname*{argmin}_a Q(a, X)$$

Розглянемо спочатку **випадок**:

- a — константа
- y є випадковою функцією із густиною розподілу $f(t)$.

У такому випадку середньоквадратична похибка має вигляд:

$$Q(a) = \int_t (a-t)^2 f(t) dt$$

($l = \infty$, тобто у нас не вибірка з y, а уся випадкова величина)

Неважко показати, що:

$$a_* = \underset{a}{\operatorname{argmin}} Q(a) = \mathbb{E}y$$

тобто **найкраща константа**, що апроксимує значення y у сенсі середньоквадратичної похибки — це **математичне очікування** y.

Якщо $a(x)$ — довільна функція ознак x, функціонал середньоквадратичної похибки має вигляд:

$$Q(a(x), X) = \int_t (a(x) - t)^2 f(t) dt$$

а його мінімум буде відповідати **умовному математичному очікуванню**:

$$a_* = \operatorname*{argmin}_a Q(a) = \mathbb{E}(y|x)$$

(середнє значення y, коли x має певні значення)

($l = \infty$, тобто у нас не вибірка з y, а уся випадкова величина)

У **випадку з скінченною вибіркою**:

$$Q(a(x), X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

оцінка, яка одержана при мінімізації середньоквадратичної похибки:

$$a_*(x) = \operatorname*{argmin}_a Q(a, X)$$

є кращою апроксимацією умовного математичного очікування

$$\mathbb{E}(y|x)$$

У **випадку лінійної регресії**, тобто коли відгук моделюється лінійною комбінацією $\langle w, x_i \rangle$:

$$Q(\mathbf{w}, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle \mathbf{w}, x_i \rangle - y_i)^2$$

$$\mathbf{w}_* = \underset{\mathbf{w}}{\operatorname{argmin}} Q(\mathbf{w}, X)$$

вираз $\langle w_*, x_i \rangle$ є найкращою лінійною апроксимацією умовного математичного очікування

$$\mathbb{E}(y|x)$$

Отриманий результат **узгоджується з інтуїтивними уявленнями**. Дійсно, нехай $y_i = 2$, **графік залежності похибки** на цьому об'єкті залежно від передбачення алгоритму $a(x)$ виглядає так:

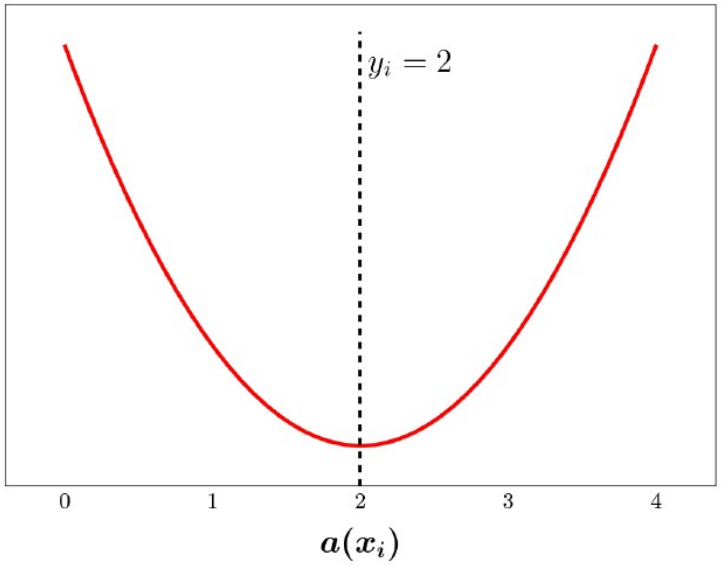


Рис. 5.2: Залежність похибки від передбачення алгоритму у випадку середньоквадратичної похибки.

За графіком видно, що однаково штрафуються відхилення передбачення як у більший, так і в менший бік від істинного значення y_i. Тому не дивно, що функція, що відповідає мінімуму функції похибок, є якесь середнім.

Дивергенція Брегмана

Однак виявляється, що умовне математичне очікування відповідає мінімуму не тільки для середньоквадратичної похибки, але й більше широкого класу функцій втрат, які називаються **дивергенціями Брегмана**.

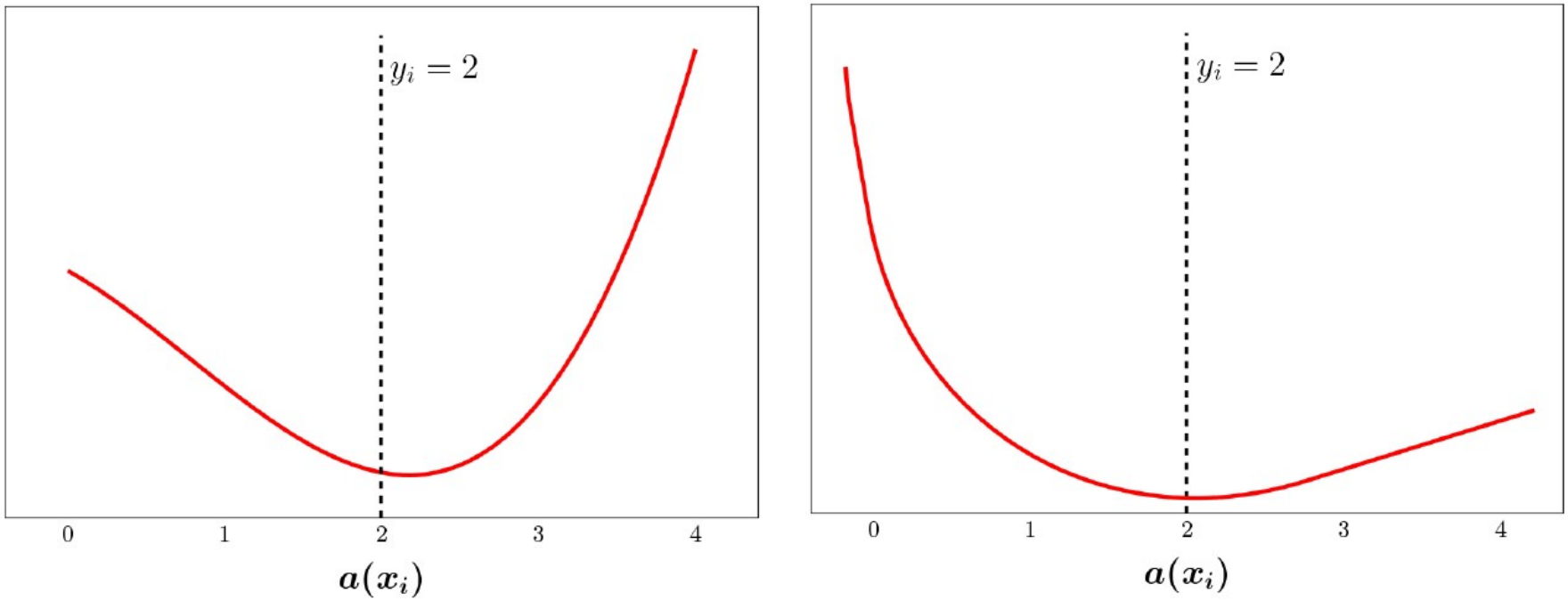
Дивергенції Брегмана породжуються будь-якою неперервною диференційованою опуклою функцією φ:

$$Q(a, X) = \\ = \varphi(y) - \varphi(a(X)) - \varphi'(a(X))(y - a(X))$$

Середньоквадратична похибка є частковим випадком дивергенції Брегмана. Мінімізуючи будь-яку дивергенцію Брегмана, ми одержуємо оцінку для умовного математичного очікування:

$$a_*(x) = \operatorname*{argmin}_a Q(a, X)$$

є кращою апроксимацією умовного математичного очікування $\mathrm{E}(y \mid x)$



Декілька функцій втрат із класу дивергенцій Брегмана

Цей результат уже є трохи дивним, оскільки в сімействі дивергенцій Брегмана можна знайти, у тому числі, несиметричні відносно *y* функції. Такі функції більше штрафують за відхилення моделі в більший або менший бік. Це результат може бути трохи контрінтуїтивним і отриманий не дуже давно.

Середня абсолютна похибка й несиметрична абсолютна похибка

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} |a(x_i) - y_i|$$

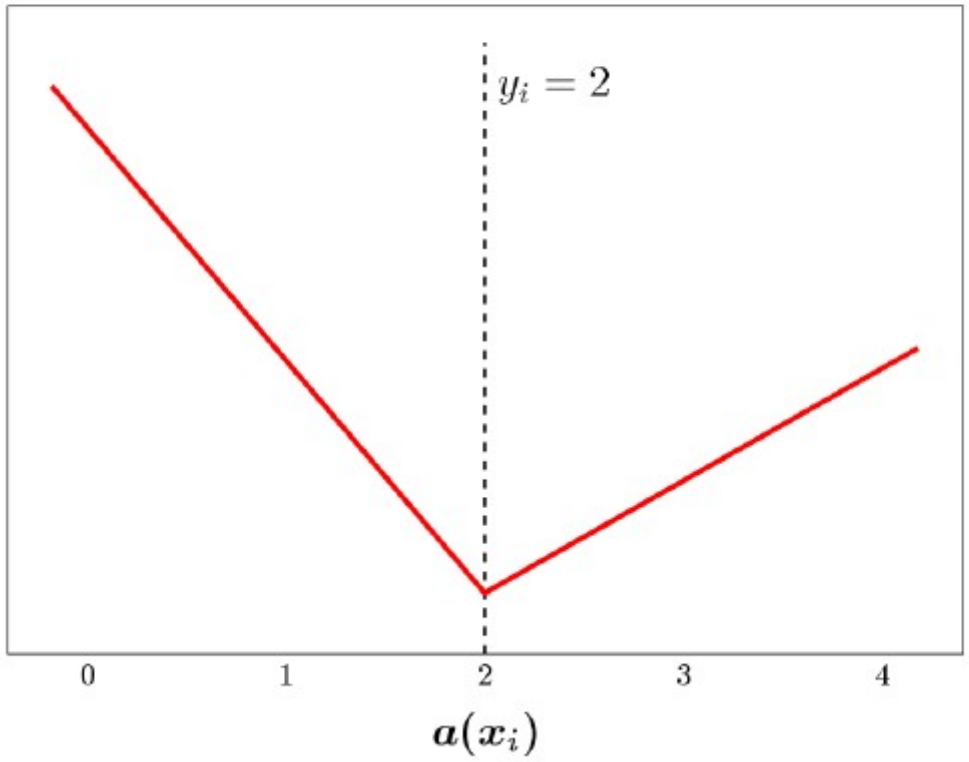
При її мінімізації отримуємо оцінку

$$a_*(x) = \operatorname*{argmin}_a Q(a, X)$$

не умовного математичного очікування, а оцінку умовної медіани:

$$\mathrm{med}(y|x)$$

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} ((\tau - 1)[y_i < a(x_i)] + \\ + \tau[y_i \geq a(x_i)])(y_i - a(x_i))$$



Графік несиметричної абсолютної функції похибок.

При мінімізації такого функціонала отримуємо

$$a_*(x) = \underset{a}{\operatorname{argmin}}\, Q(a, X)$$

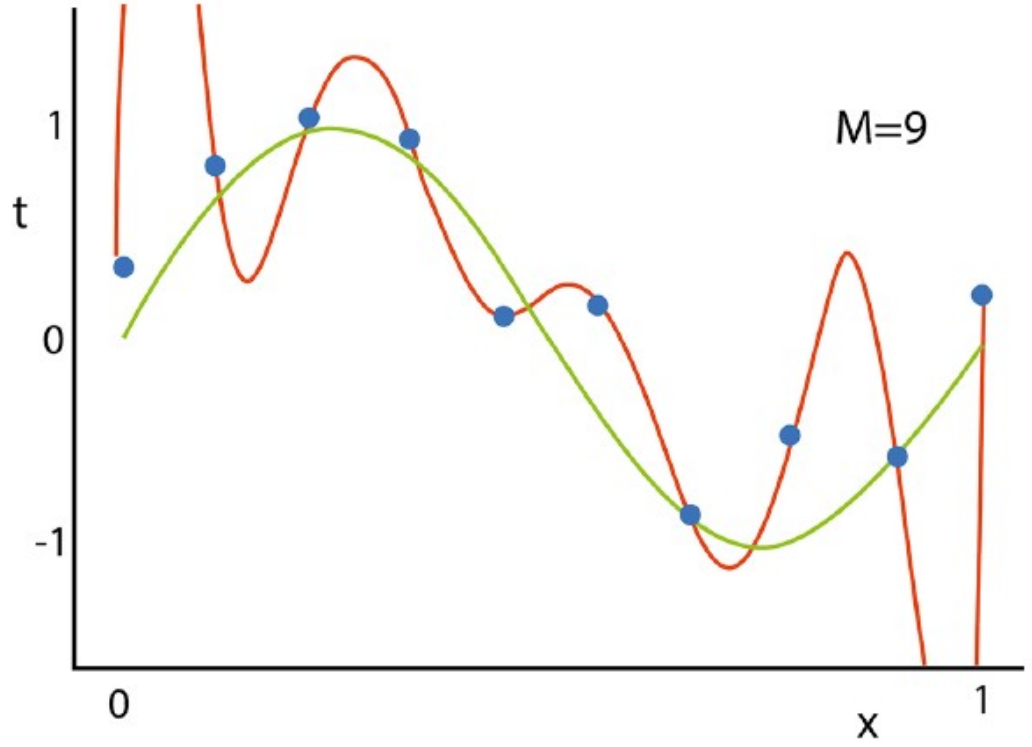
кращу оцінку $y \mid x$ для відповідного умовного квантиля (порядка τ)

Висновок

- розв'язання задачі МНК-регресії - оцінка умовного математичного очікування
- розв'язання задачі квантильної регресії при використанні середньої абсолютної помилки – оцінка умовної медіани
- розв'язання задачі квантильної регресії з несиметричною абсолютною функцію похибок - оцінка умовного квантилю;

§44 Регуляризація

Перенавчання регресійних моделей



Якщо використовується **занадто складна модель**, а даних недостатньо, щоб точно визначити її параметри, ця модель легко може вийти **перенавченою, тобто добре описувати навчальну вибірку й погано – тестову**.

Боротися із цим можна різними способами:

- **Взяти більше даних**. Такий варіант звичайно недоступний, оскільки додаткові дані коштують додаткових грошей, а також іноді недоступні зовсім. Наприклад, у задачах веб-пошуку, не дивлячись на наявність терабайтів даних, ефективний об'єм вибірки, що описує персоналізовані дані, істотно обмежений: у цьому випадку можна використовувати тільки історію відвідувань даного користувача.
- **Вибрати більше просту модель** або спростити модель, наприклад **виключивши з розгляду деякі ознаки**. Процес відбору ознак являє собою нетривіальну задачу. Зокрема, не зрозуміло, який із двох схожих ознак варто залишати, якщо ознаки містять сильний шум.
- **Використовувати регуляризацію**. Раніше було показано, що в перенавченої лінійної моделі значення ваг у моделі стають величезними й різними за знаком. Якщо обмежити значення ваг моделі, то з перенавчанням можна до якогось ступеня поборотися.

L1-регуляризація й L2-регуляризація

Є кілька способів провести регуляризацію:

- L2-регуляризатор (ridge-регресія або гребенева регресія):

$$\mathbf{w}_* = \underset{\mathbf{w}}{\operatorname{argmin}} \left(\frac{1}{\ell} \sum_{i=1}^{\ell} (\langle \mathbf{w}, x_i \rangle - y_i)^2 + \lambda \sum_{j=1}^{d} \mathbf{w}_j^2 \right)$$

- L1-регуляризатор (lasso-регресія або ласо-регресія):

$$\mathbf{w}_* = \underset{\mathbf{w}}{\operatorname{argmin}} \left(\frac{1}{\ell} \sum_{i=1}^{\ell} (\langle \mathbf{w}, x_i \rangle - y_i)^2 + \lambda \sum_{j=1}^{d} |\mathbf{w}_j| \right)$$

Важливо: константний доданок не входить у штрафи.

Модельний приклад

Зрозуміти розходження між L1 і L2 регулязаторами можна на модельному прикладі.

Нехай $l = d$, X – одинична матриця, константа відсутня

$$X = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{pmatrix}$$

Тоді МНК-регресію без регуляризації має вигляд:

$$\mathbf{w}_* = \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{i=1}^{\ell} (\mathbf{w}_i - y_i)^2$$

Відповідь:

$$\mathbf{w}_{*j} = y_j$$

Якщо розглянемо L2 –регуляризацію,

$$\mathbf{w}_* = \underset{\mathbf{w}}{\operatorname{argmin}} \left(\frac{1}{\ell} \sum_{i=1}^{\ell} (\langle \mathbf{w}, x_i \rangle - y_i)^2 + \lambda \sum_{j=1}^{d} \mathbf{w}_j^2 \right)$$

то компоненти вектора ваг будуть мати вигляд:

$$\mathbf{w}_{*j} = \frac{y_j}{1+\lambda}$$

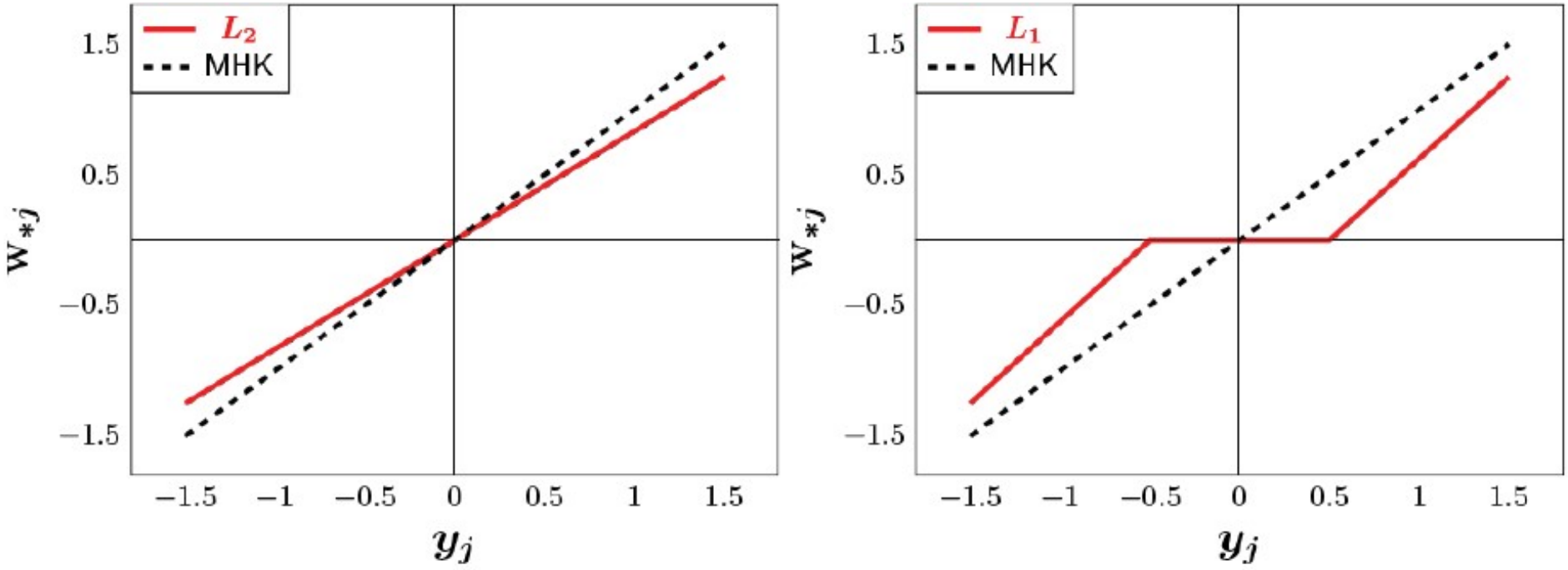
а при використанні L1 -регуляризатора (lasso):

$$\mathbf{w}_* = \underset{\mathbf{w}}{\operatorname{argmin}} \left(\frac{1}{\ell} \sum_{i=1}^{\ell} (\langle \mathbf{w}, x_i \rangle - y_i)^2 + \lambda \sum_{j=1}^{d} |\mathbf{w}_j| \right)$$

отримаємо

$$\mathbf{w}_{*j} = \begin{cases} y_j - \lambda/2, & y_j > \lambda/2 \\ y_j + \lambda/2, & y_j < -\lambda/2 \\ 0, & |y_j| \le \lambda/2 \end{cases}$$

Графіки залежності коефіцієнтів w_{*j}



- При використанні **L2 регуляризації** залежність w_{*j} від y_j усе ще лінійна, **компоненти вектора ваг ближче розташовані до нуля**.
- У випадку L1 регуляризації графік виглядає трохи інакше: **існує область (розміру λ) значень y_j , для яких w_{*j} =0**. Тобто lasso, або L1 - регуляризація, дозволяє відбирати ознаки, а саме: ваги ознак, що мають низку передбачувальну здатність, виявляються такими що дорівнюють нулю.

Зміщення і дисперсія

Можна показати, що математичне очікування квадрата похибки регресії являє собою суму трьох компонентів:

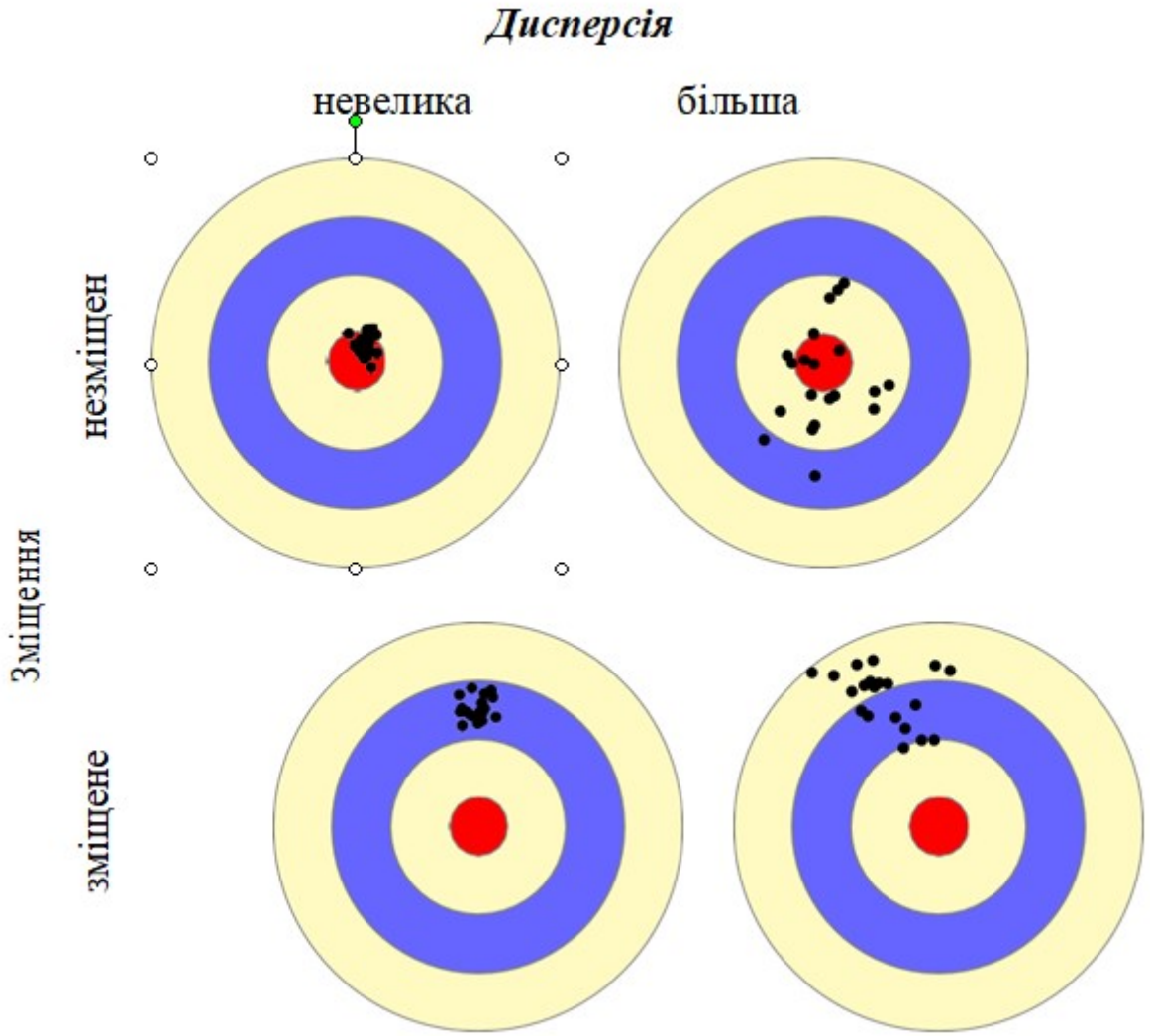
$$\mathbb{E}(a_*(x) - y)^2 = \\ = (\mathbb{E}a_*(x) - a(x))^2 + \mathbb{D}a_*(x) + \sigma^2$$

квадрат зміщення **дисперсія оцінки** **шум**

Від вибору моделі залежить квадрат зміщення й дисперсія оцінки, але не шум, що є властивістю даних, а не моделі.

- Метод найменших квадратів дає оцінки, які мають **нульове зміщення**.
- Регуляризація дає зміщені оцінки, але їх дисперсія може бути менше!

Аналогія дозволяє краще зрозуміти баланс між зміщенням і дисперсією.



При стрілянині по мішені **середнє число набраних очків** залежить від **положення середньої точки** влучення й **розкидом** відносно цього середнього.

Кращий результат буде, якщо стріляти без зсуву й без розкиду. **Перенавчанню лінійних моделей відповідає стрілянина без зсуву, але з величезним розкидом.**

І часто виявляється, що **можна набрати більше очків, стріляючи не зовсім у ціль, тобто зі зсувом, але зате більш точно**. Саме це й дозволяє домогтися **регуляризація**.

Вирішення задач гребеневої регресії (L_2) й ласо

У байесовській статистиці

- **гребенева регресія** відповідає заданню нормального апріорного розподілу на коефіцієнти лінійної моделі,
- **метод ласо** відповідає заданню Лапласівського апріорного розподілу.

- Задача гребеневої регресії має аналітичне вирішення:

$$\mathbf{w}_* = (X^T X + \lambda I)^{-1} X^T y$$

- Для вирішення задачі ласо аналітичного рішення не існує, однак є дуже ефективний чисельний спосіб одержання розв'язку.

Висновки

- Регуляризація - один із способів боротьби з перенавчанням
- Дає зміщені оцінки коефіцієнтів, але помилка може бути меншою за рахунок меншої дисперсії
- Лассо ще й відбирає ознаки

§45 Логістична регресія

Логістична регресія

Нехай X — простір об'єктів, Y — простір відповідей, $X = (x_i, y_i)_{i=1}^{l}$ — навчальна вибірка, $x = (x^1, ..., x^d)$ – ознаковий опис.

Логістична регресія – це метод навчання із учителем у задачі бінарної класифікації Y = {0, 1} з використанням логістичної функції втрат.

Метод лінійного дискримінанта Фішера

Метод лінійного дискримінанта Фішера, один із самих старих **методів класифікації**, полягає в мінімізації середньоквадратичної похибки:

$$Q(\mathrm{w}, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle \mathrm{w}, x_i \rangle - y_i)^2$$

У результаті отримуємо вектор ваг:

$$\mathrm{w}_* = \operatorname*{argmin}_{\mathrm{w}} Q(\mathrm{w}, X)$$

Якщо для деякого об'єкта $\langle w, x_i \rangle > 0{,}5$, об'єкт відносять до першого класу $y = 1$, в іншому випадку, до нульового $y = 0$.

Насправді, хочеться передбачити не просто мітки класів, а і ймовірності того, що об'єкти відносяться до якогось із класів:

$$P(y = 1|x) \equiv \pi(x)$$

Хоча $\pi(x)$ збігається з умовним математичним очікуванням:

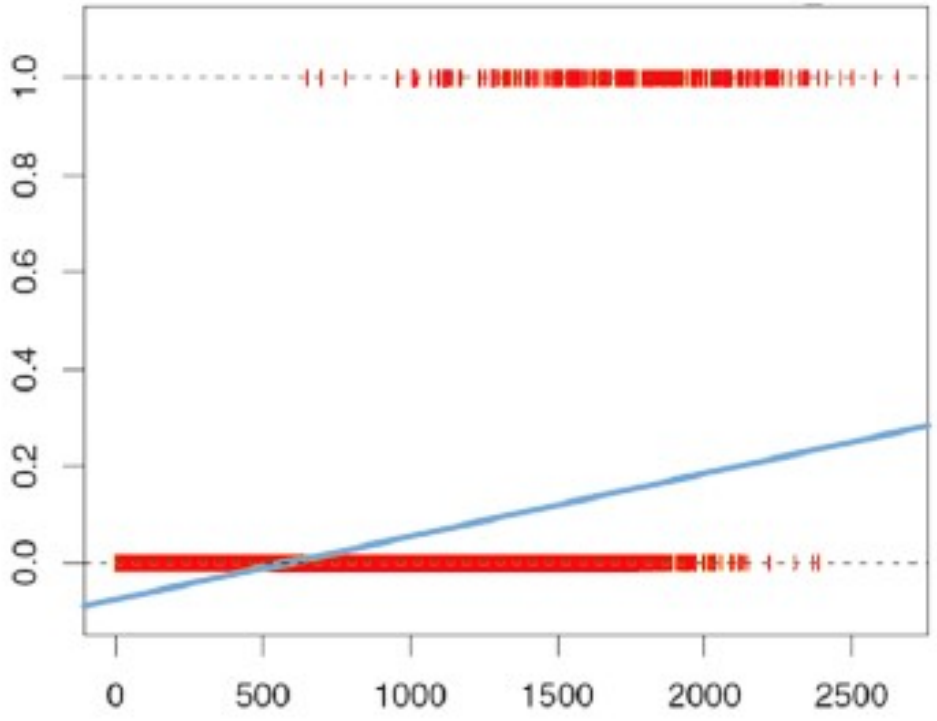
$$\pi(x) = 1 \cdot P(y = 1|x) + 0 \cdot P(y = 0|x) = \\ = \mathbb{E}(y|x)$$

використовувати для оцінки ймовірності звичайну лінійну регресію

$$\pi(x) \approx \langle \mathbf{w}, x \rangle$$

не вийде: **отримана лінійна комбінація факторів не обов'язково належить на відрізку від 0 до 1**.

Наприклад, вирішується наступна задача. Необхідно **передбачити ймовірність неповернення платежу по кредитній карті залежно від розміру заборгованості**.



За навчальною вибіркою була створена **модель лінійної регресії**. Отримано, що при заборгованості 2000$ імовірність прострочити платіж по кредиту дорівнює 0.2, при заборгованості 500$ — нулю, **а при менших значеннях і зовсім від'ємна**. Також, якщо заборгованість більше 10000$, імовірність **прострочення буде більше 1**. Не зрозуміло, як інтерпретувати цей результат.

Узагальнені лінійні моделі

Наступна ідея

Нехай функція $g:[0,1]\to \boldsymbol{R}$ **переводить інтервал [0, 1] на множину всіх дійсних чисел**, тоді можна вирішувати задачу лінійної регресії:

$$g(\mathbb{E}(y|x)) \approx \langle \mathbf{w}, x \rangle$$

у якій будується оцінка не для умовного математичного очікування E(*y*|*x*), а для *g*(E(*y*|*x*)). Що те ж саме:

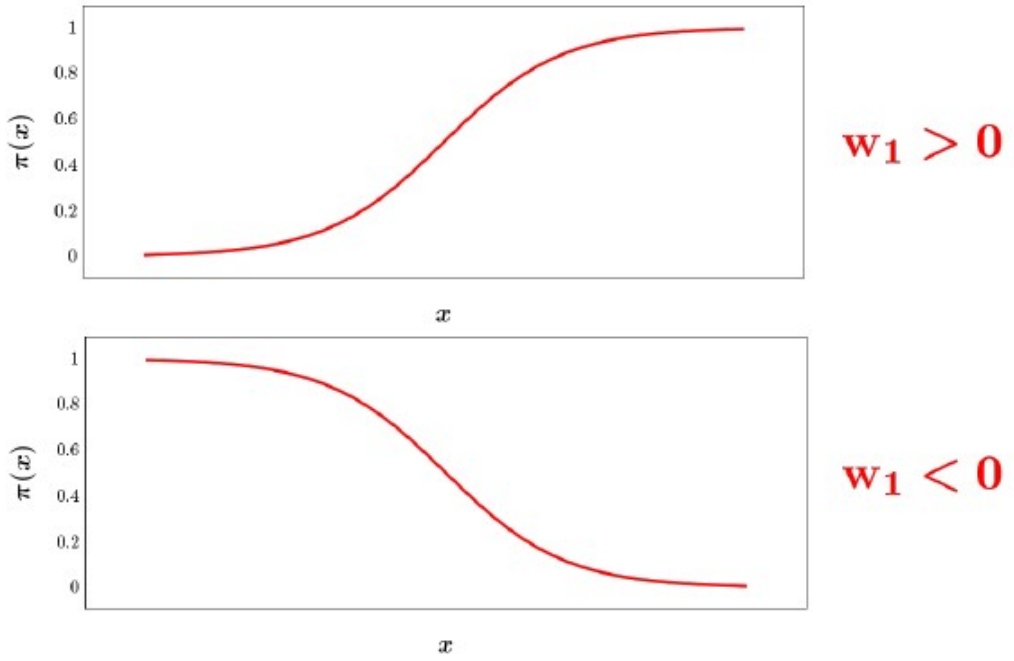
$$\mathbb{E}(y|x) \approx g^{-1}(\langle \mathbf{w}, x \rangle)$$

У статистиці таке сімейство моделей називається **узагальненими лінійними моделями (GLM)**.

У задачі бінарної класифікації в якості g^{-1} використовується сигмоїда:

$$\pi(x) \approx \frac{e^{\langle \mathbf{w}, x \rangle}}{1 + e^{\langle \mathbf{w}, x \rangle}}$$

В одновимірному випадку значення параметр w_0 сигмоїди визначає положення її центра на числовій осі, а w_1 — форму цієї сигмоїди:



Чим більше за модулем значення w_1, тим крутіше нахил сигмоїди в області її середини

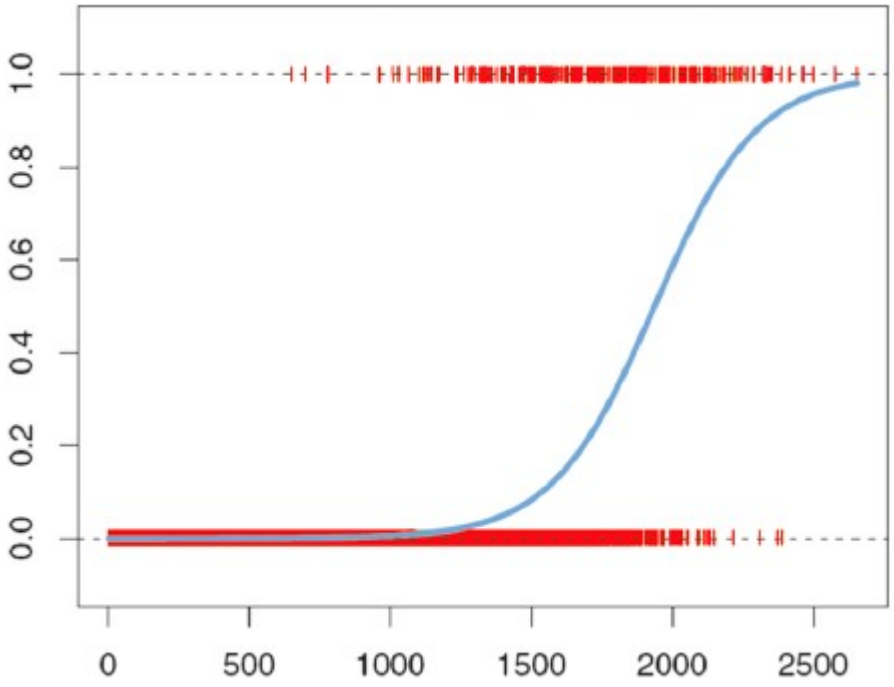
Передбачення ймовірностей

Якщо використовувати сигмоїду:

$$\pi(x) \approx \frac{e^{\langle \mathbf{w}, x \rangle}}{1 + e^{\langle \mathbf{w}, x \rangle}}$$

в узагальненій лінійній моделі в задачі логістичної регресії, результат буде адекватним:

- Імовірність $\pi(x) \in [0, 1]$, як і потрібно.
- На краях області значень x функція (імовірність) $\pi(x)$ слабко змінюється при невеликих змінах x, тоді як істотно змінюється, якщо x міститься в середині діапазону своїх значень.



Остання властивість є досить корисною. Наприклад, у вже розглянутій задачі при розмірі заборгованості **в районі 2000$ оцінка ймовірності прострочення платежу сильно змінюється при збільшенні або зменшенні заборгованості на 100$.** З іншої сторони при розмірі заборгованості в 500$ збільшення заборгованості на 100$ приводить тільки до незначних змін необхідної оцінки.

Оцінка параметрів

За функцією $\pi(x)$ можна відновити функцію g, що фігурує у визначенні узагальненої лінійної моделі:

$$\pi(x) \approx \frac{e^{\langle w,x\rangle}}{1+e^{\langle w,x\rangle}} \quad \Longleftrightarrow \quad \langle w,x\rangle \approx \underbrace{\ln \overbrace{\frac{\pi(x)}{1-\pi(x)}}^{\text{Риск}}}_{\text{Логит}}.$$

Відношення, що міститься **під логарифмом**, називається **ризиком**, а весь **логарифм називається «логіт»**. Саме тому метод називається логістичною регресією: логіт наближається лінійною комбінацією факторів. Налаштування моделі відбувається методом максимізації правдоподібності *L*(*X*):

$$L(X) = \prod_{i:y_i=1} \pi(x_i) \prod_{i:y_i=0} (1-\pi(x_i)) = \prod_i \pi(x_i)^{y_i} (1-\pi(x_i))^{1-y_i}$$

Зручніше однак не максимізувати правдоподібність, а мінімізувати мінус логарифм від правдоподібності:

$$-\ln L(X) = -\sum_{i=1}^{\ell} \Big(y_i \ln \pi(x_i) + (1 - y_i) \ln(1 - \pi(x_i)) \Big)$$

Такий функціонал також має назви **log-loss**, **крос-ентропія** й інші.

Якщо змінити мітку **нульового класу на –1**, то отримаємо логістичну функцію втрат у такому вигляді, у якому вона зустрічалася в курсі до цього:

$$L(X) = \prod_{i:y_i=1} \pi(x_i) \prod_{i:y_i=-1} (1-\pi(x_i)) = \prod_{i:y_i=1} \frac{\exp(< w, x_i >)}{1+\exp(< w, x_i >)} \prod_{i:y_i=-1} \frac{1}{1+\exp(< w, x_i >)} =$$

$$= \prod_{i:y_i=1} \frac{\exp(< w, x_i > \cdot y_i)}{1+\exp(< w, x_i > \cdot y_i)} \prod_{i:y_i=-1} \frac{1}{1+\exp(- < w, x_i > \cdot y_i)} =$$

$$= \prod_{i:y_i=1} \frac{1}{\exp(- < w, x_i > \cdot y_i)+1} \prod_{i:y_i=-1} \frac{1}{1+\exp(- < w, x_i > \cdot y_i)} = \prod_i \frac{1}{\exp(- < w, x_i > \cdot y_i)+1}.$$

Потрібно мінімізувати

$$-\ln(L(X)) = \sum_i \ln[1+\exp(-y_i < w, x_i >)]$$

Нагадаю, раніше ми говорили про мінімізацю логістичної функції втрат

$$Q(\mathbf{w}, X) = \sum_{i=1}^{\ell} \ln\left(1+\exp\left(-y_i \langle \mathbf{w}, x_i \rangle\right)\right)$$

Вирішення задачі максимізації правдоподібності

Задача **максимізації правдоподібності**

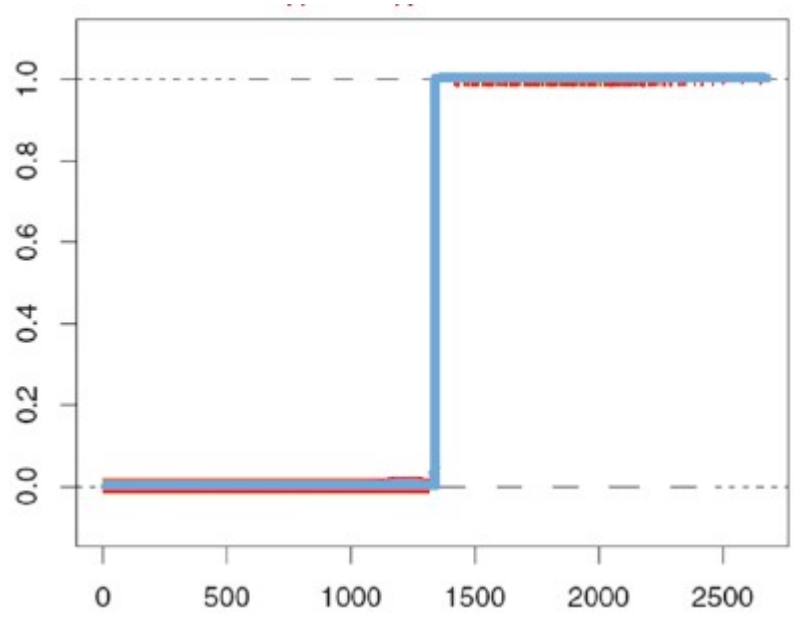
$$\ln L(X) = \sum_{i=1}^{\ell} \Big(y_i \ln \pi(x_i) + (1 - y_i) \ln(1 - \pi(x_i)) \Big)$$

в логістичній регресії дуже добре вирішується ЧИСЕЛЬНО, оскільки правдоподібність – опукла функція, а отже, вона має **єдиний глобальний максимум**. Крім того, її градієнт і гессіан можуть бути добре оцінені.

Матриця Гесе — квадратна матриця елементами якої є часткові похідні деякої функції:

$$H(f)_{ij}(x) = \frac{\partial^2 f}{\partial x_i \, \partial x_j}$$

Якщо **об'єкти з різних класів лінійно роздільні в просторі ознак**, виникає **проблема перенавчання** $\| w \| \to \infty$: сигмоїда **вироджується в «сходинку»**.



Наприклад, така ситуація виникає, якщо у вже згаданій задачі оцінці ймовірності повернути заборгованість, навчальна вибірка така, що всі клієнти із заборгованістю менш 1300$ повернули платіж вчасно, а всі клієнти із заборгованістю більше 1300$ — ні.

У таких випадках необхідно використовувати **методи регуляризації**, наприклад L1 або L2 регуляризатор.

Передбачення відгуку

- Імовірності, які дає логістична регресія, можна використовувати для класифікації, тобто для передбачення підсумкових міток класів. Для цього вибирається поріг p_0 і об'єкт відноситься до класу 1 тільки у випадку $\pi(x) > p_0$. В інших випадках об'єкт відноситься до класу 0.
- Поріг p_0 не слід вибирати завжди рівним 0.5, як це може здатися з інтуїтивних міркувань. Його необхідно підбирати для кожної задачі окремо таким чином, щоб забезпечити оптимальний баланс між точністю й повнотою класифікатора.