Линейные модели: введение
по материалам "The Elements of Statistical Learning"

И. Куралёнок, Н. Поваров

Яндекс

СПб, 2013

Формальная постановка

Ищем решающую функцию в виде:

$$y = F(\lambda, x) = \lambda^T x$$

Такое решение кажется примитивным!

Формальная постановка

Ищем решающую функцию в виде:

$$y = F(\lambda, x) = \lambda^T x$$

Такое решение кажется примитивным!
До того как мы расскажем что такое x.

Какое x бывает

Просто фичи:

$$x \in \mathbb{R}^n$$

Мономы:

$$u \in \mathbb{R}^n x = \prod u_j$$

Произвольные функции:

$$u \in \mathbb{R}^n x : \mathbb{R}^n \to \mathbb{R}$$

В любом случае мы всегда можем посчитать значение x по входным параметрам.

Простое решение

$$\arg\min_{\lambda} \|F(X,\lambda) - y\| = \arg\min_{\lambda} \|X\lambda - y\|$$

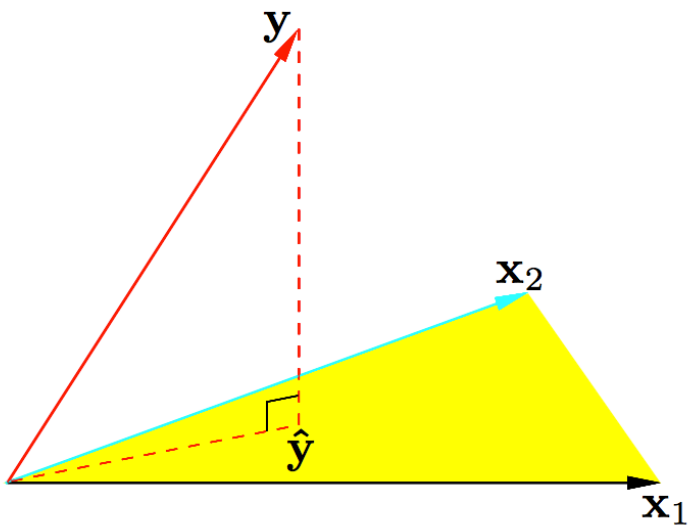
Если норма l_2, то:

$$\frac{\partial T}{\partial X} = 2X^T(X\lambda - y) = 0$$

$$\lambda_0 = (X^T X)^{-1} X^T y$$

Геометрическая интерпретация

Если посмотреть на колонки, соответствующие фичам то картинка такая:



Об этом говорит (если нам все удалось):

$$X^T(y - \hat{y}) = X^T(y - X\lambda_0) = 0$$

В случае, если $rank(X) < n$ ортогональность остается!

Статистические свойства решения

Если наблюдения независимы, $Var(y) = const$, а x вычислены точно:

$$Var(\lambda) = \left(X^T X\right)^{-1} \frac{1}{m-n-1} \|y - \hat{y}\|_2$$

А если еще и предположить, что $y = \lambda_1^T x + \epsilon$ и $\epsilon \sim N(0, \sigma)$:

$$\lambda_0 \sim N(\lambda_1, \left(X^T X\right)^{-1} \sigma^2)$$

а наблюдаемая σ для y распределена по χ^2:

$$(n-m-1)\hat{\sigma} = \|y - \hat{y}\|_2 \sim \sigma \chi^2_{m-n-1}$$

А точно $\lambda_{0_i} \neq 0$?

Введем такую штуку (Z-score):

$$z_i = \frac{\lambda_{0_i}}{\hat{\sigma}\sqrt{v_i}}$$

где v_i — диагональный элемент $\left(X^T X\right)^{-1}$. Если подумать что $\lambda_{0_i} = 0$, то:

$$z_i \sim T_{m-n-1}$$

Чем больше Z-score, тем более мы уверены, что $\lambda_{0_i} \neq 0$

Теорема Гаусса-Маркова

Theorem

Линейное приближение по MSE обладает на наименьшим разбросом из всех несмещенных линейных решений

$\Rightarrow$ для того, чтобы сделать решение более стабильным надо вводить bias

$\Rightarrow$ простым MSE нам не отделаться, надо будет менять T

Расширение на несколько целей

$$y_i \in \mathbb{R}^k$$

В этом случае задача превращается в такую:

$$\arg\min_{\Lambda} tr\left((Y - X\Lambda)^T (Y - X\Lambda)\right)$$

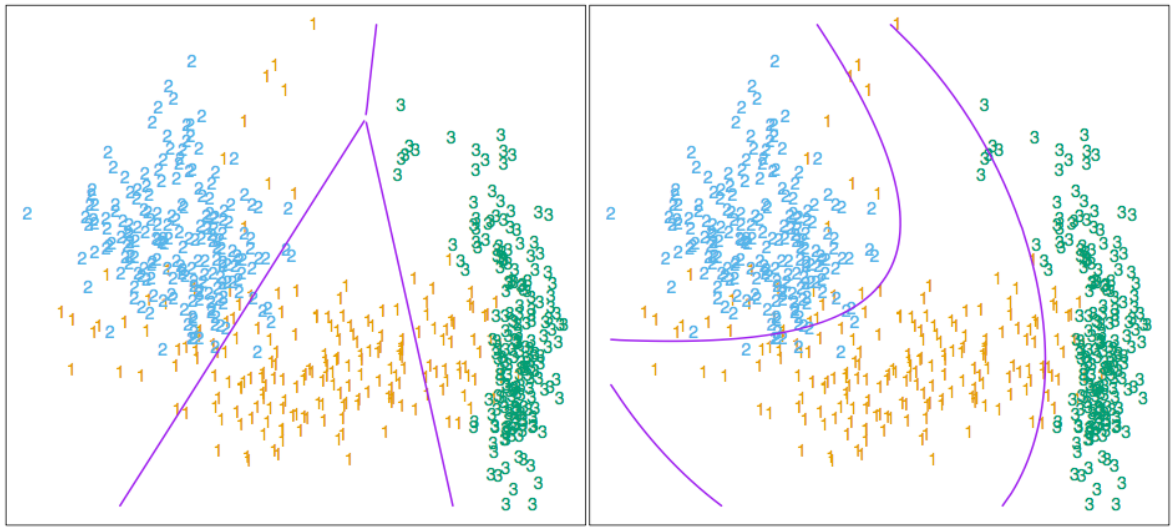
$$\Lambda_0 = \left(X^T X\right)^{-1} X^T Y$$

Если же $y = x^T\Lambda + \epsilon$, $\epsilon \sim N(0, \Sigma)$:

$$\arg\min_{\Lambda} \left((Y - X\Lambda)^T \Sigma^{-1} (Y - X\Lambda)\right)$$

Классификация

$$x \in \mathbb{R}^n, y \in \{1, \ldots, k\}$$



Введем дискриминационные функции для каждого класса. У какого класса больше, тот и молодец. Там где равны — границы решения.

NB: монотонные преобразования дискриминационным функциям не страшны

Линейное решение задачи классификации

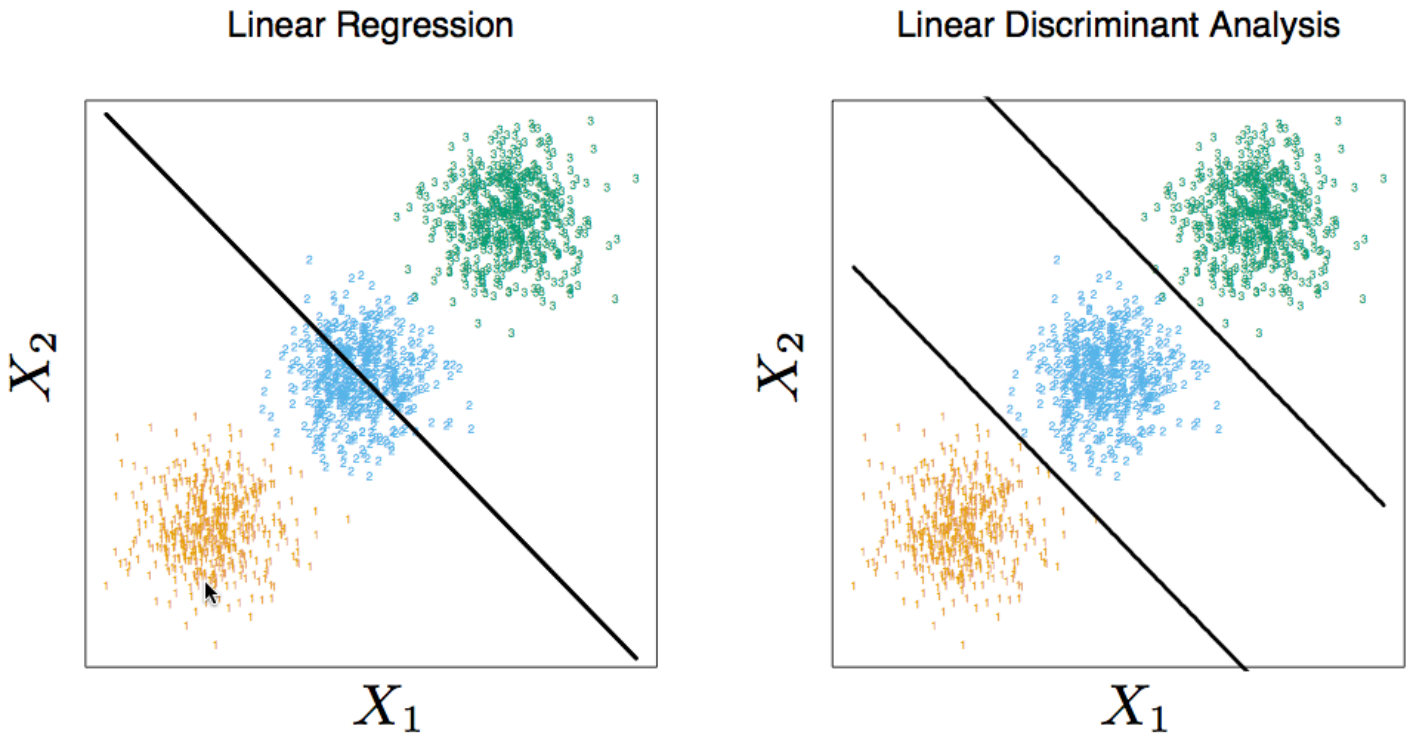
Можем пойти по-простому и решить регрессией:

$$\gamma_{ij} = \begin{cases} 1, & i = y_i \\ 0 \end{cases}$$

В терминах предсказания γ решаем:

$$\arg\min_{\Lambda} tr\left((\Gamma - X\Lambda)^T(\Gamma - X\Lambda)\right)$$

Сложности с простым решением



Линейный дискриминантный анализ (LDA)

Представим себе, что точки порождены смесью нормальных распределений по одному на класс:

$$f_j = \frac{1}{\sqrt{(2\pi)^n |\Sigma_j|}} e^{-\frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)}$$

Границы решения прямые! Если зафиксировать Σ:

$$f_j(x) = x^T \Sigma^{-1} \mu_j - \frac{1}{2} \mu_j^T \Sigma^{-1} \mu_j + log \pi_j$$

LDA: решение

Можно решать напрямую:

$$\pi_j = \frac{m_j}{m},$$
$$\mu_j = \frac{1}{m_j} \sum_{x_u : y_u = j} x_u,$$
$$\Sigma = \frac{1}{m-k} \sum_{j=1}^{k} \sum_{x_u : y_u = j} (x_u - \mu_j)(x_u - \mu_j)^T$$

Можно *EM* даже не заморачиваясь одинаковыми Σ_j

LDA: свойства

- Нормальные распределения в основе
- Решение в аналитическом виде
- Работает даже в далеких от “гауссовых” ситуаций
- Имеет расширение в квадратичные мономы (QDA)
- Часто рассматривают диагональные Σ_k для ускорения вычислений
- Можно использовать $\Sigma_k = \alpha\Sigma_0 + (1 - \alpha)\Sigma^k$

Логистическая регрессия

Будем искать не дискриминантные функции, а напрямую границы между классами:

$$log\left(\frac{P(y=u|x)}{P(y=k|x)}\right) = x^T\lambda_u$$

Преобразование справа — logit. Тогда вероятности можно найти так:

$$p(y=u|x) = \frac{x^T\lambda_u}{1+\sum_{v<k} x^T\lambda_v}, u < k$$

$$p(y=k|x) = \frac{1}{1+\sum_{v<k} x^T\lambda_v}$$

Оптимизация логистическая регрессия

Вероятности у нас есть, давайте максимизировать правдоподобие!

$$\arg\max_{\Lambda} \prod_{i} p(y_i|x_i, \Lambda)$$

$$= \arg\max_{\Lambda} \sum_{i} log \frac{x^T \lambda_u}{1 + \sum_{v<k} x^T \lambda_v}$$

Как будем искать?

Оптимизация логистическая регрессия

Вероятности у нас есть, давайте максимизировать правдоподобие!

$$\arg\max_{\Lambda} \prod_i p(y_i|x_i, \Lambda)$$

$$= \arg\max_{\Lambda} \sum_i log \frac{x^T \lambda_u}{1 + \sum_{v<k} x^T \lambda_v}$$

Как будем искать?

Когда что?

- Есть много точек, для которых нет оценок $\Rightarrow$ LDA
- Есть подозрение на близость к норамльности $\Rightarrow$ LDA
- Хотим использовать prior $\Rightarrow$ LDA
- Во всех остальных случаях логистическая регрессия, особенно если есть много outlier'ов

Результаты ДЗ второй недели

1. 6af9df
2. dccc3e
3. f33f66
4. f1015e
5. f33f2d
6. 458048
7. 93184e
8. 608072
9. 824e76
10. 88d593
11. cfd271
12. 48364c
13. c5b930
14. 080c07
15. 579569
16. 01b988
17. 68c819
18. dcb652
19. ba605a
20. 692f0b
21. 6aca1b

Домашнее задание

- SVN, howto.txt
- Две недели