

Прикладная статистика в машинном обучении

Лекция 2

Вероятностный взгляд на ML

И. К. Козлов
(Мехмат МГУ)

2022

Что такое ML?

1 Что такое ML?

2 Линейная алгебра

3 ML решение

4 Задача классификации

5 Пара инженерных проблем

6 Вероятностный подход



“ Machine learning is the study of computer algorithms that allow computer programs to automatically improve through experience.

~ Tom Mitchell,
Machine Learning, McGraw Hill, 1997

Carnegie Mellon University
Machine Learning

Что такое ML? Какие в нём задачи?

Обучение с учителем

Одна из типичных задач машинного обучения — обучение с учителем.

- Даны **обучающая выборка** — набор пар $(x_1, y_1), \dots, (x_n, y_n)$
- $x_i \in X$ — это объекты, $y_i \in Y$ — их метки (классы).
- **Наша задача** — построить **решающую функцию**

$$a : X \rightarrow Y,$$

как можно лучше предсказывающую метки.

Регрессия и классификация

Типичные задачи машинного обучения:

- Регрессия $Y = \mathbb{R}$.
- Классификация $Y = \{1, \dots, M\}$

Конечно, есть и другие задачи.

Ранжирование.

Google

Bing

YAHOO!

Yandex

Ask

Bai du 百度



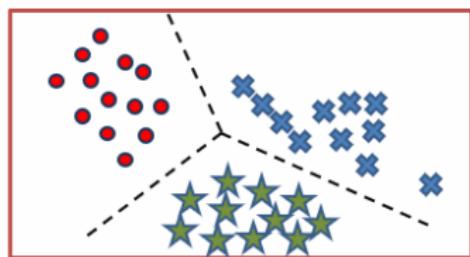
DuckDuckGo

wow

Кластеризация

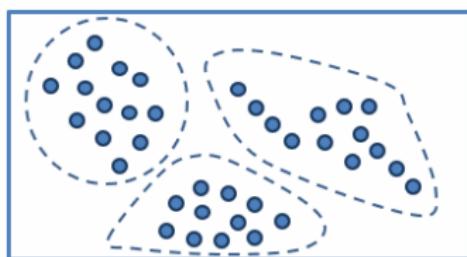
Кластеризация (обучение без учителя).

Classification



Supervised learning

Clustering



Unsupervised learning

RL

Обучение с подкреплением = RL



Боты в играх



RL в медицине



Беспилотники

Признаки

Каждый объект x_i задаётся своими признаками x_{i1}, \dots, x_{ip} .

Для простоты будем считать, что все признаки — числа $f_i \in \mathbb{R}$ ([количественные признаки](#)).

На практике многие признаки — [категориальные](#), т.е. принимают конечное число значений (мужчина/женщина, группа крови, регион и т.д.)

По матрице предсказываем таргет

Итак, нам дана [матрица объектов-признаков](#)

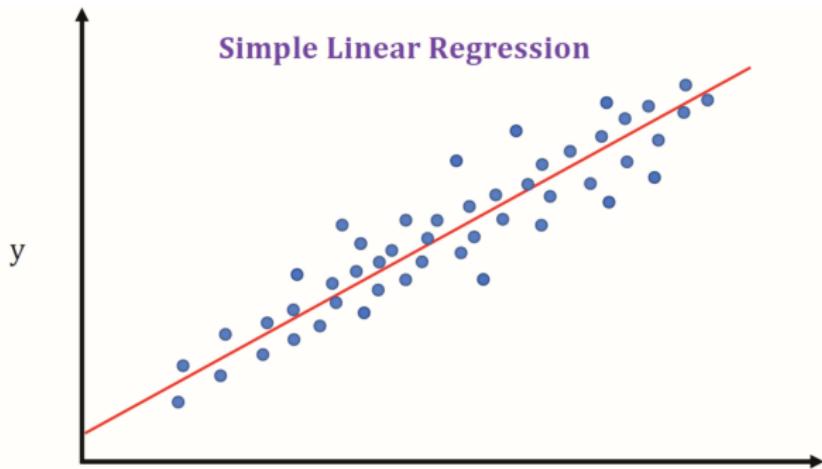
$$\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

и мы пытаемся по ней предсказать метки y_i .

Линейная регрессия

Рассмотрим простейшую задачу линейной регрессии

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \dots, n,$$



Линейная регрессия

Вместе все эти уравнения можно записать в матричном виде

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Подчеркнём — чтобы учесть свободный член, в 1ом столбце матрицы \mathbf{X} стоят 1:

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix},$$

Эволюция на сегодня

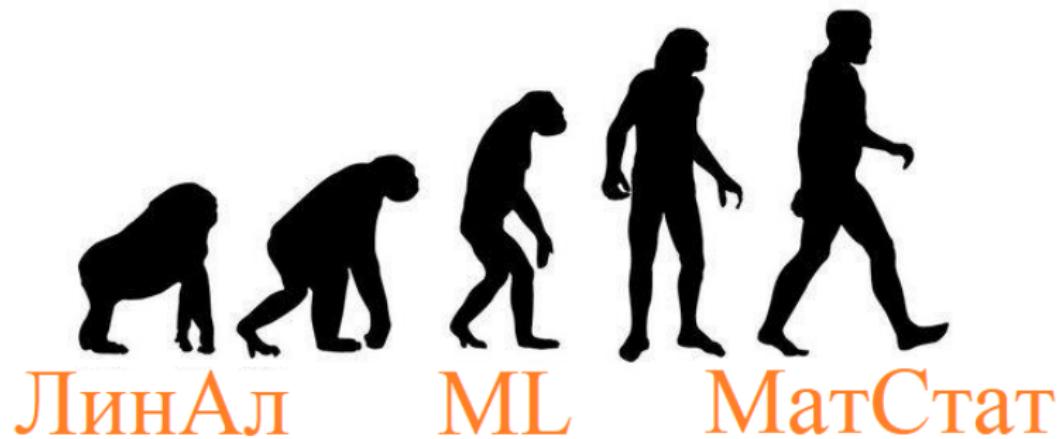


Рис.: Предстоящая эволюция в понимании задачи

Линейная алгебра

1 Что такое ML?

2 Линейная алгебра

3 ML решение

4 Задача классификации

5 Пара инженерных проблем

6 Вероятностный подход

Начнём с доисторических методов (1ый курс).



СЛУ

Q: Как решать СЛУ

$$Ax = b?$$

A: Правильно,

$$x = A^{-1}b.$$

Q2: А если матрица A не квадратная?

Псевдорешение

Для любого ЛСУ

$$Ax = b$$

существует **псевдорешение**

$$A^+b.$$

Это

- ① решение методом наименьших квадратов

$$\|Ax - b\| \rightarrow \min,$$

- ② наименьшей длины

$$\|x\| \rightarrow \min.$$

Псевдообратная матрица

Псевдообратную матрицу A^+ легко определить через **SVD-разложение**. С его помощью доказываются многие теоретические факты.

Потом приведём более простые формулы для A^+ в частных случаях.

Сингулярное разложение

SVD-разложение:

$$A = U \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix} V^T$$

Любая матрица $m \times n$ Ортогональная $m \times m$ Диагональная $m \times n$ Ортогональная $n \times n$

$$\sigma_1 \geq \dots \geq \sigma_r > 0$$

Напомним, матрица U ортогональная, если $U^T U = E$.

Сингулярное разложение

SVD-разложение над \mathbb{C} аналогично:

$$A = U\Sigma\bar{V}^T,$$

- ① U, V – унитарные матрицы

$$U\bar{U}^T = E, \quad V\bar{V}^T = E.$$

- ② все значения диагональной матрицы Σ вещественны и неотрицательны:

$$\sigma_1 \geq \dots \geq \sigma_r > 0.$$

SVD разложение

Теорема

Для любой (вещественной или комплексной) $n \times m$ -матрицы A существует SVD разложение.

У нас матрицы будут вещественными.

Доказательство (или ссылки на него) см.

Wiki:https://en.wikipedia.org/wiki/Singular_value_decomposition.

Псевдообратная через SVD

Псевдообратная через SVD

$$A = U\Sigma V^T.$$

Тогда псевдообратная матрица:

$$A^+ = V\Sigma^+ U^T,$$

где

$$\Sigma^+ = \left(\frac{1}{\sigma_1}, \dots, \frac{1}{\sigma_k}, 0, \dots, 0 \right).$$

Псевдообратная через SVD

Псевдообратная через SVD

$$A = U\Sigma V^T.$$

Тогда псевдообратная матрица:

$$A^+ = V\Sigma^+ U^T,$$

где

$$\Sigma^+ = \left(\frac{1}{\sigma_1}, \dots, \frac{1}{\sigma_k}, 0, \dots, 0 \right).$$

Q: A — это $m \times n$ матрица, какой размер A^+ ?

Псевдообратная через SVD

Псевдообратная через SVD

$$A = U\Sigma V^T.$$

Тогда псевдообратная матрица:

$$A^+ = V\Sigma^+ U^T,$$

где

$$\Sigma^+ = \left(\frac{1}{\sigma_1}, \dots, \frac{1}{\sigma_k}, 0, \dots, 0 \right).$$

Q: A — это $m \times n$ матрица, какой размер A^+ ?

A^+ — это $n \times m$ матрица.

Псевдообратная матрица

Алгоритмы нахождения псевдообратной матрицы (в частных случаях):

- Если A — квадратная и обратимая

$$A^+ = A^{-1}$$

Псевдообратная матрица

- Пусть столбцы A линейно независимы. Тогда $A^T A$ обратима.

Домножаем уравнение $Ax = b$ на A^T , получаем

$$A^T A x = A^T b.$$

Формула для псевдообратной матрицы

$$A^+ = (A^T A)^{-1} A^T.$$

Псевдообратная матрица

- Аналогично, если строчки A линейно независимы, то AA^T обратима и

$$A^+ = A^T (AA^T)^{-1}.$$

Кратко: если любая из 2 формул выше вычислима, то она верна.

ML решение

- 1 Что такое ML?
- 2 Линейная алгебра
- 3 ML решение
- 4 Задача классификации
- 5 Пара инженерных проблем
- 6 Вероятностный подход

Инженерный подход

Теперь обсудим инженерный подход.



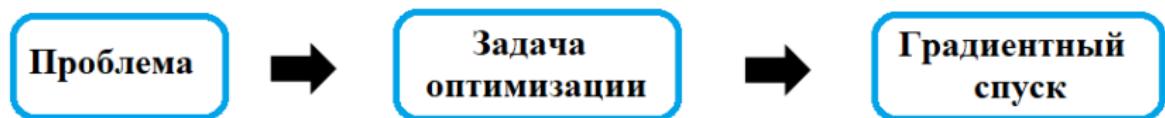
Погрешность вычислений

Формулу $A^+ = (A^T A)^{-1} A^T$ трудно вычислять на практике.

Много делений/умножений \Rightarrow большая погрешность вычислений.

Стандартный пайплайн

Стандартный пайплайн в вычислительных задачах:



Стандартный пайплайн

Насколько хорошо наше предсказание?

Вводим квадратичную функцию потерь¹

$$\mathcal{L} = \frac{1}{n} \sum_{(x_i, y_i)} (y_i - prediction(x_i))^2.$$

Для линейной регрессии решение — минимум этой функции:

$$\hat{\beta} = \arg \min_{\beta} \mathcal{L} = \arg \min_{\beta} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \beta)^2$$

¹ MSE = mean squared error

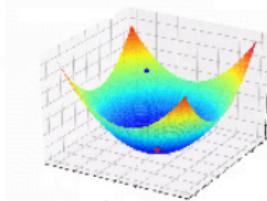
Не перепутайте:

Метрика



Чего хотим
побольше

Функция потерь



Что можем
оптимизировать

Градиент

Утверждение. Градиент — направление наибольшего роста функции.

Дана функция $f(x^1, \dots, x^n) : \mathbb{R}^n \rightarrow \mathbb{R}$. Её градиент:

$$\left(\frac{\partial f}{\partial x^1}, \dots, \frac{\partial f}{\partial x^n} \right).$$

Производная по направлению v (в нуле):

$$\frac{d}{dt} f(tv^1, \dots, tv^n) = \sum_{i=1}^n \frac{\partial f}{\partial x^i} v^i = (\text{grad } f, v).$$

Градиент

Утверждение. Градиент — направление наибольшего роста функции.

Дана функция $f(x^1, \dots, x^n) : \mathbb{R}^n \rightarrow \mathbb{R}$. Её градиент:

$$\left(\frac{\partial f}{\partial x^1}, \dots, \frac{\partial f}{\partial x^n} \right).$$

Производная по направлению v (в нуле):

$$\frac{d}{dt} f(tv^1, \dots, tv^n) = \sum_{i=1}^n \frac{\partial f}{\partial x^i} v^i = (\text{grad } f, v).$$

Напомним:

$$(u, v) = |u| |v| \cos \varphi$$

Поэтому скалярное произведение максимально, если векторы сонаправлены.

Градиентные методы

Поэтому для поиска оптимума можно использовать различные градиентные методы

$$\beta_j \rightarrow \beta_j - \alpha \frac{\partial}{\partial \beta_j} \mathcal{L}.$$

На практике формулы посложнее.

Градиентные методы

Поэтому для поиска оптимума можно использовать различные градиентные методы

$$\beta_j \rightarrow \beta_j - \alpha \frac{\partial}{\partial \beta_j} \mathcal{L}.$$

На практике формулы посложнее.

Упр. Проверить, что градиент для функции

$$\mathcal{L}(\beta, X, y) = \|y - X\beta\|^2$$

имеет вид

$$\nabla_{\beta} \mathcal{L} = X^T (X\beta - y).$$

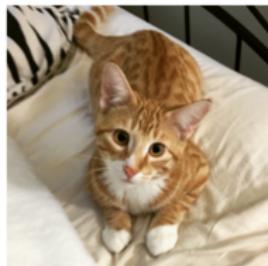
Задача классификации

- 1 Что такое ML?
- 2 Линейная алгебра
- 3 ML решение
- 4 Задача классификации
- 5 Пара инженерных проблем
- 6 Вероятностный подход

Классификация

Поговорим про задачу классификации. Пусть классов всего два: 0 и 1.

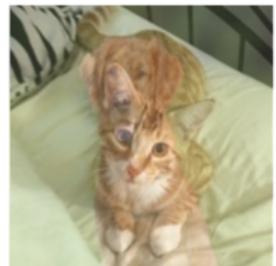
Для классификации достаточно предсказывать вероятность класса $\mathbb{P}(y_i = 1|x)$.



[1.0, 0.0]
cat dog



[0.0, 1.0]
cat dog



[0.7, 0.3]
cat dog

Рис.: Предсказываем вероятности классов

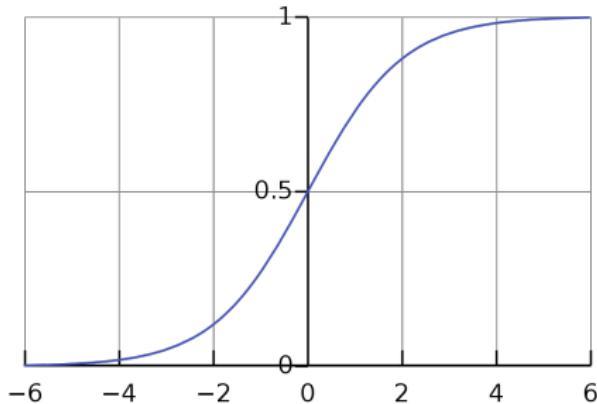
Сигмоида

Переход: регрессия → классификация.

Достаточно отобразить предсказания из \mathbb{R} в вероятности из $[0, 1]$.

Удобной оказывается **сигмоида**

$$\sigma(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1}$$



Логистическая регрессия

Логистическая регрессия:

$$p = \frac{1}{1 + b^{-(\beta_0 + \sum \beta_i x_i)}}$$

Логистическая регрессия

Логистическая регрессия:

$$p = \frac{1}{1 + b^{-(\beta_0 + \sum \beta_i x_i)}}$$

В качестве функции потерь берётся [логлосс](#) (logloss):

$$L_{\text{log}}(y, p) = -(y \log p + (1 - y) \log(1 - p)).$$

Другое название: [перекрёстная кросс-энтропия](#) (Cross Entropy).

Причину использования этого лосса [обсудим на семинаре](#).

Катарсис



Рис.: Сейчас будет катарсис

Deep Learning

Q: Знаете ли Вы, что такое нейронная сеть/Deep Learning?

Q: Знаете ли Вы, что такое нейронная сеть/Deep Learning?



Yann LeCun

24 декабря 2019 г. ·



Some folks still seem confused about what deep learning is. Here is a definition:

DL is constructing networks of parameterized functional modules & training them from examples using gradient-based optimization. That's it.

This definition is orthogonal to the learning paradigm: reinforcement, supervised, or self-supervised.

Регрессии — нейронки

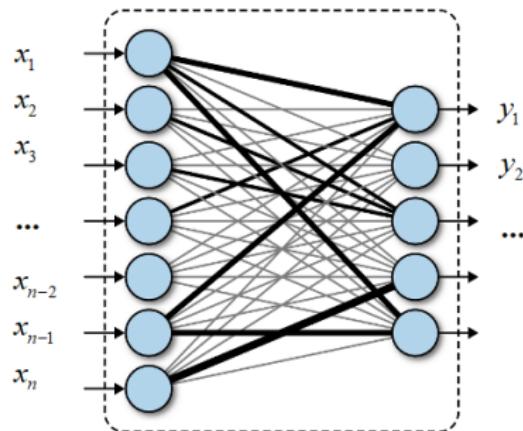
Линейная регрессия — это однослойная нейронная сеть.

“Док-во”. Полносвязанный слой — это линейное отображение (см. сайт pytorch).

LINEAR ↗

```
CLASS torch.nn.Linear(in_features, out_features, bias=True, device=None, dtype=None) [SOURCE]
```

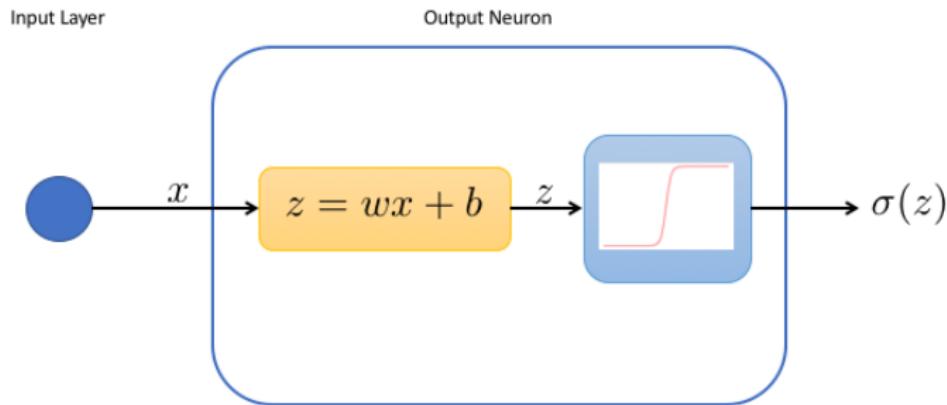
Applies a linear transformation to the incoming data: $y = xA^T + b$



Регрессии — нейронки

Логистическая регрессия — тоже нейронная сеть:

полносвязный слой + сигмоида



Сигмоида играет роль **функции активации** в нейроне.

Перерыв

Перерыв

Пара инженерных проблем

- 1 Что такое ML?
- 2 Линейная алгебра
- 3 ML решение
- 4 Задача классификации
- 5 Пара инженерных проблем
- 6 Вероятностный подход

Много вычислений

С линейной регрессией возникает пара проблем:

- ➊ Долго вычислять градиент.

На каждом шаге нужно искать градиент функции потерь $\nabla \mathcal{L}$.

А это сумма большого числа слагаемых

$$\mathcal{L} = \frac{1}{n} \sum_{i=1}^n \mathcal{L}(\beta, x_i, y_i).$$

Стохастический градиентный спуск

Общий подход в программировании:

Хотим ускорить процесс — распараллеливаем его.



Стохастический градиентный спуск

Стохастический градиентный спуск: заменяет градиент

$$\nabla \mathcal{L} = \frac{1}{n} \sum_{i=1}^n \nabla \mathcal{L}(\beta, x_i, y_i)$$

на одно слагаемое

$$\nabla_j = \nabla \mathcal{L}(\beta, x_j, y_j).$$

Стохастический градиентный спуск

Стохастический градиентный спуск: заменяет градиент

$$\nabla \mathcal{L} = \frac{1}{n} \sum_{i=1}^n \nabla \mathcal{L}(\beta, x_i, y_i)$$

на одно слагаемое

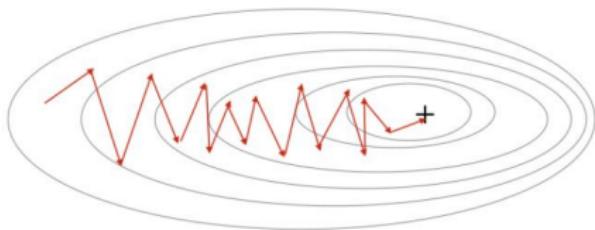
$$\nabla_j = \nabla \mathcal{L}(\beta, x_j, y_j).$$

На практике берут не одно слагаемое, а среднее по **мини-батчу**

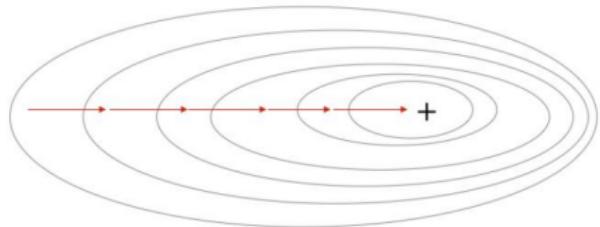
$$\nabla \mathcal{L} \approx \frac{1}{B} \sum_{t=1}^B \nabla \mathcal{L}(\beta, x_{i_t}, y_{i_t}).$$

Стохастический градиентный спуск

Stochastic Gradient Descent



Gradient Descent



Алгоритм Роббинса-Монро

Математически сходимость обосновывается алгоритмом Роббинса-Монро.

Важно, что градиент $\nabla \mathcal{L}$ заменяется на его несмещённую оценку в этой точке

$$\nabla_B = \frac{1}{B} \sum_{t=1}^B \nabla \mathcal{L}(\beta, x_{i_t}, y_{i_t}), \quad \Rightarrow \quad \mathbb{E} \nabla_B = \nabla \mathcal{L}$$

Очень грубо говоря, в среднем мы идём в правильном направлении.

Мультиколлинеарность

2ая проблема с линейной регрессией:

② Мультиколлинеарность.

Пусть признаки зависимы, скажем,

$$x_1 = x_2 + x_3.$$

Тогда к решению можно прибавлять тривиальное выражение вида

$$\alpha x_1 - \alpha(x_2 + x_3).$$

Мультиколлинеарность

Выражение

$$(X^T X)^{-1} X^T$$

нельзя посчитать, если матрица $(X^T X)$ вырождена (или близка к вырожденной).

Мультиколлинеарность

Выражение

$$(X^T X)^{-1} X^T$$

нельзя посчитать, если матрица $(X^T X)$ вырождена (или близка к вырожденной).

Q: Как выглядит следующее множество?

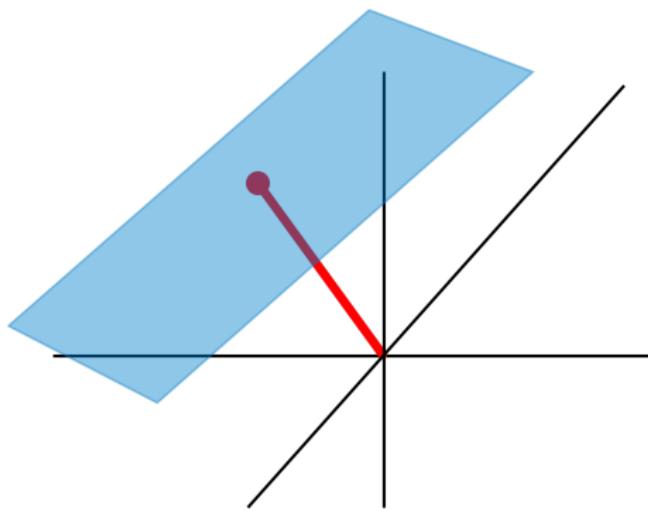
$$\|X\beta - y\| \rightarrow \min$$

Мультиколлинеарность

Линейная алгебра учит нас, что решение $X\beta = y$ — это подпространство.

Общее решение = частное решение + $\text{Ker } X$:

$$X\hat{\beta} = y, \quad X\gamma = 0 \quad \Rightarrow \quad X(\hat{\beta} + \gamma) = y.$$



Как решить проблему?



Регуляризация

Регуляризуем лосс — добавляем слагаемое, чтобы ограничить модуль решения.

① L_2 -регуляризация (Ridge regression)

$$\hat{\mathcal{L}} = \mathcal{L} + \lambda \sum_{j=1}^n \beta_j^2.$$

Регуляризация

Регуляризуем лосс — добавляем слагаемое, чтобы ограничить модуль решения.

- ➊ L_2 -регуляризация (Ridge regression)

$$\hat{\mathcal{L}} = \mathcal{L} + \lambda \sum_{j=1}^n \beta_j^2.$$

- ➋ L_1 -регуляризация (Lasso regression)

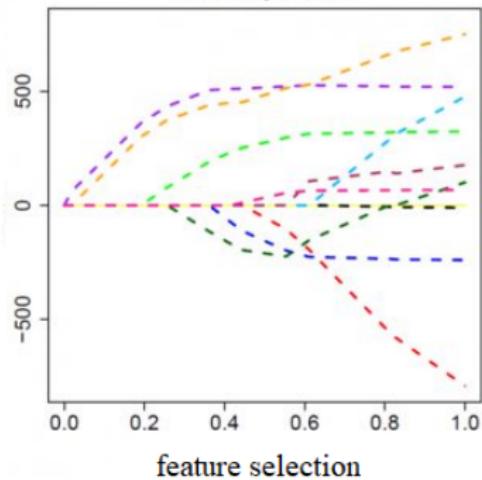
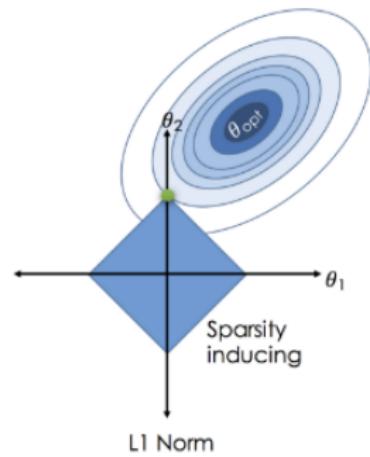
$$\hat{\mathcal{L}} = \mathcal{L} + \lambda \sum_{j=1}^n |\beta_j|.$$

λ — подбираемый гиперпараметр.

LASSO отбирает признаки

Q: Откуда название LASSO?

A: При изменении λ происходит **отбор признаков** — β_i постепенно обнуляются.



Вероятностный подход

- 1 Что такое ML?
- 2 Линейная алгебра
- 3 ML решение
- 4 Задача классификации
- 5 Пара инженерных проблем
- 6 Вероятностный подход

Вероятностные модели

Общая вероятностная постановка задачи ML:

- Есть неизвестная вероятностная плотность $p(x, y)$.
- Пытаемся аппроксимировать её плотностью $\varphi(x, y, \theta)$.
- Делаем это, по *принципу максимума правдоподобия*:

$$\mathcal{L} = \prod_{i=1}^n \varphi(x_i, y_i, \theta) \rightarrow \max.$$

Вероятностные модели

По сути стандартная задача статистики.

Q: Чем полезен вероятностный подход? Его плюсы и минусы?

Преимущества и недостатки

Недостатки:

- ① Сложнее.
- ② “Игра часто не стоит свеч” (костыли и эвристики for the win).

Преимущества и недостатки

Недостатки:

- ① Сложнее.
- ② “Игра часто не стоит свеч” (костыли и эвристики for the win).

Преимущества:

- ① Можно добавлять априорные знания.
- ② Модели допускают пропуски в данных.
- ③ Если знаем вероятность $p(y|x)$ метки y для данного объекта x , то можем только предсказать метку y .

А зная совместную плотность $p(x,y)$, можно генерировать (!) объекты.

2 типа моделей

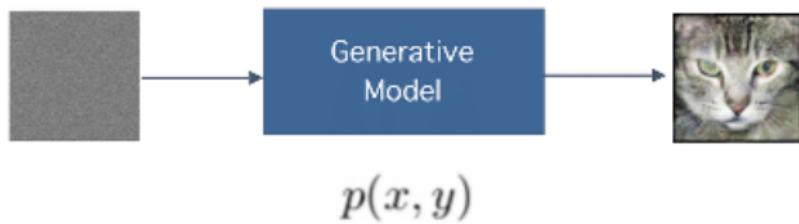
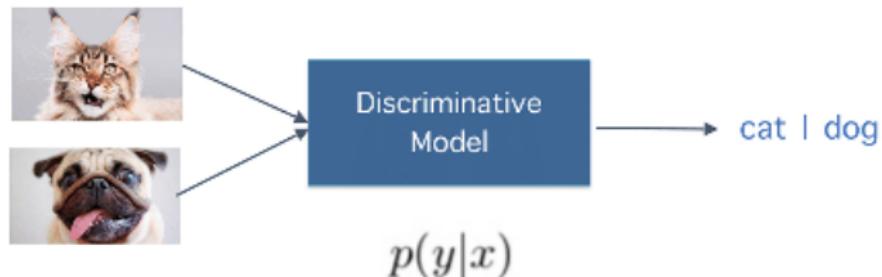


Рис.: Генеративные VS Дискриминативные модели

Зашумлённое предсказание

Считаем, что:

- Значения (таргет), которые мы наблюдаем, *зашумлены*:

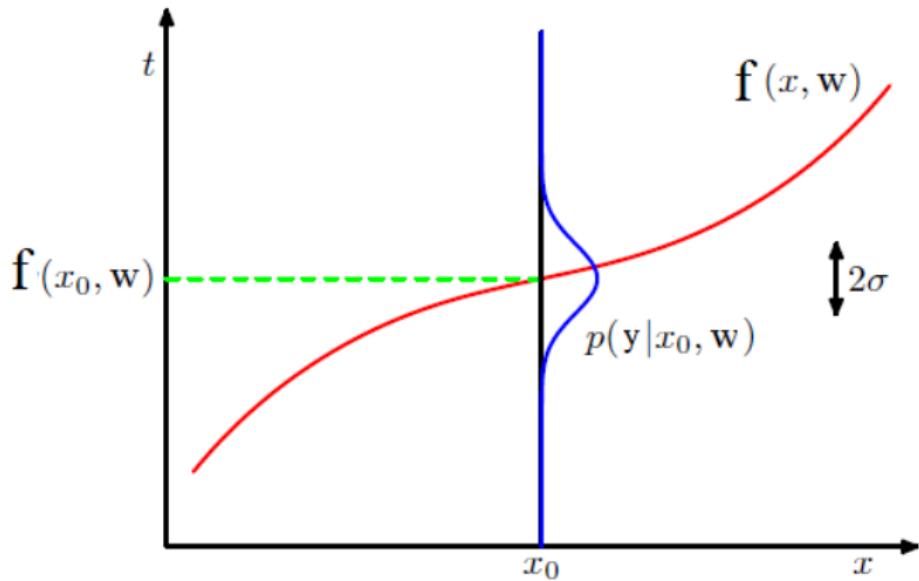
$$y_i = f(x_i, \omega) + \varepsilon_i$$

- ε_i — это гауссовский шум, т.е. он имеет нормальное распределение

$$\varepsilon_i | X \sim \mathcal{N}(0, \sigma^2).$$

- Шумы независимы и одинаково распределены.

Зашумлённое предсказание



Условная вероятность

Мы будем использовать [условные вероятности](#).

Подробно обсудим их потом. Выражение вида

$$p(y_i|x_i, \omega)$$

следует понимать как “вероятность y_i при фиксированных x_i, β ”.

Можно считать это формальным обозначением. Зафиксируем x_i, ω .

Правдоподобие

Вероятность $y_i = f(x_i, \omega) + \varepsilon_i$, такая же, как и у шума ε_i :

$$p(y_i|x_i, \omega) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - f(x_i, \omega))^2}{2\sigma^2}\right).$$

Посмотрим на оценку максимума правдоподобия.

- (Условное) правдоподобие:

$$L = \prod_i p(y_i|x_i, \omega) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - f(x_i, \omega))^2}{2\sigma^2}\right)$$

- (Условное) правдоподобие:

$$L = \prod_i p(y_i|x_i, \omega) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - f(x_i, \omega))^2}{2\sigma^2}\right)$$

- Логарифм правдоподобия:

$$\ell = - \sum_i \frac{(y_i - f(x_i, \omega))^2}{2\sigma^2} + \text{const.}$$

Q: Мы видели ранее похожее выражение?

Max Likelihood = Min MSE

Максимизация правдоподобия

$$\ell = - \sum_i \frac{(y_i - f(x_i, \omega))^2}{2\sigma^2} + \text{const}$$

эквивалентна минимизации MSE-ошибки

$$\text{MSE} = \frac{1}{n} \sum_i (y_i - f(x_i, \omega))^2.$$

Подведём итоги

В предположении нормальности:

ОМП = оценка МНК

А где в вероятностной схеме регуляризация?



Что дальше?

Обсудим во 2ой части курса.

Если ввести правильное априорное распределение на веса $p(\beta)$
и применить теорему Байеса

$$p(\beta \mid \text{Data}) = \frac{p(\text{Data} \mid \beta)p(\beta)}{p(\text{Data})}$$

получатся формулы для регуляризации.

Что дальше?

Объяснит ли теорема Байеса регуляризацию?
Какие новые модели мы узнаем?



← To Be Continued ┌