Basic knowledge :

Random experiment.

Sample space.

Event.

Probability function.

- Return -> Mean 期望

$$\mathrm{E}[X] = \Sigma_x x\, P(X = x) = \Sigma_x x\, f(x)$$

Linearity

- Risk -> Variance 方差

$$\mathrm{Var}[X] = \Sigma_x (x - \mathrm{E}[X])^2 f(x)$$

$$\mathrm{Var}(X) = \mathrm{E}(x^2) - \mathrm{E}^2(x)$$

standard deviation 标准差

Disjoint ⇄ Independent ⇄ uncorrelated

P(A)=0 does not imply event A is impossible !

Formula 1: Linearity

$$\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y]$$

$$\begin{aligned} E[X+Y] &= \sum_x \sum_y (x+y) f(x,y) \\ &= \sum_x \sum_y x f(x,y) + \sum_x \sum_y y f(x,y) \\ &= \sum_x x \sum_y f(x,y) + \sum_y y \sum_x f(x,y) \\ &= \sum_x x f_X(x) + \sum_y y f_Y(y) \\ &= E[X] + E[Y] \end{aligned}$$

Formula 1: Linearity

Some useful variants:

$$\mathrm{E}[X + a] = \mathrm{E}[X] + a \quad \text{(a is a constant)}$$

$$\mathrm{E}[\Sigma_i\, X_i] = \Sigma_i\, \mathrm{E}[X_i] \quad \text{(For >2 random variables)}$$

$$\mathrm{E}[\Sigma_i\, c_i X_i] = \Sigma_i\, c_i \mathrm{E}[X_i] \quad (c_i \text{ are constants})$$

Note $\mathrm{E}[cX] = c\mathrm{E}[X]$

Formula 2

$$\mathrm{E}[g(X)] = \Sigma_x\, g(x) P(X = x) = \Sigma_x\, g(x) f(x)$$

Some useful variants:

$$\mathrm{E}[g(X) + h(X)] = \mathrm{E}[g(X)] + \mathrm{E}[h(X)]$$

Proof:

$$\begin{aligned} \mathrm{E}[g(X) + h(X)] &= \Sigma_x \big(g(x) + h(x)\big) P(X = x) \\ &= \Sigma_x\, g(x) f(x) + \Sigma_x\, h(x) f(x) \\ &= \mathrm{E}[g(X)] + \mathrm{E}[h(X)] \end{aligned}$$

Formula 3: Variance

$$\begin{aligned} \mathrm{Var}[X] &= \mathrm{E}[(X - E[X])^2] \\ &= \mathrm{E}[X^2 - 2X \times E[X] + (E[X])^2] \\ &= \mathrm{E}[X^2] - 2E[X] \times E[X] + (E[X])^2] \\ &= \mathrm{E}[X^2] - (E[X])^2] \end{aligned}$$

a. If X is a non-negative and continuous random variable, and $E(X^n)$ exists, then $E(X) = \int_0^{+\infty} P(X > x)\mathrm{d}x$, $E(X^n) = \int_0^{+\infty} n x^{n-1} P(X > x)\mathrm{d}x$

b. (Formula of Total Probability) $P(A) = \sum_{i=1}^{n} P(A \mid B_i)\, P(B_i)$ when the probability of an event is difficult to obtain, it can be transformed into the sum of the probability of occurrence under a series of conditions.

c. (Standardized Normal Distribution) If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$

Formula 4: variance and covariance

$\mathrm{Var}[X] = \mathrm{E}[(X - E[X])^2] = \mathrm{E}[(X - E[X]) \times (X - E[X])] = \mathrm{Cov}(X, X)$

A useful result:

$\mathrm{Var}[X + Y] = \mathrm{Cov}(X + Y, X + Y) = \mathrm{E}[(X + Y - E[X] - E[Y]) \times (X + Y - E[X] - E[Y])]$

$= \mathrm{E}[(X - E[X])^2] + \mathrm{E}[(Y - E[Y])^2] + 2\,\mathrm{E}[(X - E[X]) \times (Y - E[Y])]$

$= \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}(X, Y)$

>2 RVs: $\mathrm{Var}[X_1 + \cdots + X_n] = \sum_{i=1}^{n} Var(X_i) + 2\sum_{i<j} Cov(X_i, X_j)$

Hat Check Again

n people go to a party and leave their hat with a hat-check person. At the end of the party, she returns hats randomly since she doesn't care about her job. Let X be the number of people who get their original hat back. What is $E[X]$?

For $i = 1, \ldots, n$, let $X_i = \begin{cases} 1, & \text{if } i^{th} \text{ person got hat back} \\ 0, & \text{otherwise} \end{cases}$. Then $X = \sum_{i=1}^{n} X_i$.

NOT "INDEPENDENT" RVS

We will use linearity of expectation.

$$E[X_i] = 1 \cdot P(X_i = 1) + 0 \cdot P(X_i = 0) = P(X_i = 1) = P(i^{th} \text{ person got hat back}) = \frac{1}{n}$$

$$E[X] = E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} \frac{1}{n} = n \cdot \frac{1}{n} = 1$$

Now let's compute $Var(X)$. Recall each $X_i \sim Ber\left(\frac{1}{n}\right)$. By previous proof,

$$Var(X) = Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) + 2\sum_{i<j} Cov(X_i, X_j)$$

$$X_i X_j \in \{0,1\}$$

$$E[X_i X_j] = P(X_i X_j = 1) = P(X_i = 1, X_j = 1) = P(X_i = 1)P(X_j = 1 | X_i = 1) = \frac{1}{n}\left(\frac{1}{n-1}\right)$$

$$Cov(X_i, X_j) = E[X_i X_j] - E[X_i]E[X_j] = \frac{1}{n}\left(\frac{1}{n-1}\right) - \frac{1}{n^2} = \frac{n}{n^2(n-1)} - \frac{n-1}{n^2(n-1)} = \frac{1}{n^2(n-1)}$$

$$Var(X_i) = \left(\frac{1}{n}\right)\left(1 - \frac{1}{n}\right)$$

$$Var(X) = n\left(\frac{1}{n}\right)\left(1 - \frac{1}{n}\right) + (n^2 - n)\left(\frac{1}{n^2(n-1)}\right) = 1 - \frac{1}{n} + \frac{1}{n} = 1$$

Conditional Probability 条件概率

$$f(xy) = f(x)\,f(Y = y \mid X = x)$$

$$f(Y = y \mid X = x) = \frac{f(xy)}{f(x)}$$

Infected and Positive 0.0099	Infected and negative 0.0001
Well and Positive 0.0099	Well and Negative 0.9801

P(Infected) = 0.01
P(well) = 0.99
P(positive | Infected) = 0.99,
P(negative | Infected) = 0.01,
P(positive | well) = 0.01,
P(negative | well) = 0.99,

Different!

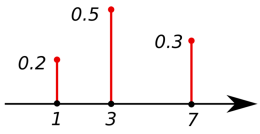
P(A| B) = P(B and A)/P(B)

P(infected | positive) = 0.50

probability function 概率函数

probability mass function 概率质量函数（pmf）

- ✓ P(ω): gives the probability for each outcome ω ∈ Ω



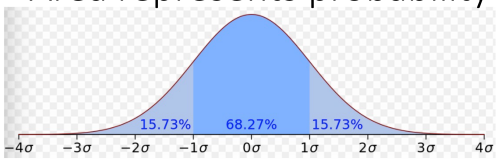
for discrete variable only

cumulative distribution probability 累积分布函数（cdf）

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i)$$

probability density function 概率密度函数（pdf）

- ✓ Area represents probability



概念密度函数不是概率，f(x)表示概率密度.
you should consider the rounding in real observation.

for continuous variable only

properties：

- ✓ $f(x) \geq 0$ for all x.
- ✓ $\int_{-\infty}^{\infty} f(x)dx = 1$.
- ✓ $P(a \leq X \leq b) = \int_a^b f(x)dx$ = area under $f(x)$ from a to b.

Distribution 分布

1. Bernoulli distribution 伯努利分布

- Take value 1 with probability p and 0 with probability 1-p.

☆ Binomial distribution 二项分布

- N Bernoulli trials.

$Pr(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$.

3. Geometric distribution 几何分布（几何级数）

- continuously draw Bernoulli R.V.
- the X sample is the first success.
- X follows geometric distribution.

$Pr(X=k) = (1-p)^{k-1} p$

4. Poisson distribution 泊松分布

λ 表示单位时间内事件发生的次数

https://zhuanlan.zhihu.com/p/139114702

- 泊松分布解释 单位时间内随机事件的累积发生次数的概率分布.
- 将过程分为 $n \to \infty$ 段，每段为伯努利试验.
- $Pr(X=k) = \binom{n}{k} p^k \cdot (1-p)^{n-k} \xrightarrow{\lambda := E(X) = np} \frac{n!}{(n-k)!k!} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k}$

$= \frac{\lambda^k e^{-\lambda}}{k!}$ ☆ $(n \to \infty)$

↑ Discrete 离散

↓ Continuous 连续

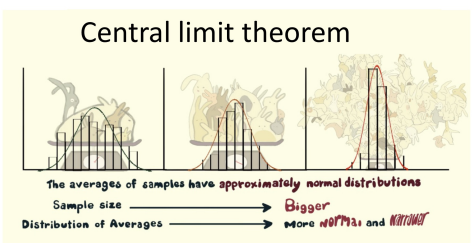
5. uniform distribution 均匀分布

☆ 6. normal distribution 正态分布 (Gaussian distribution)

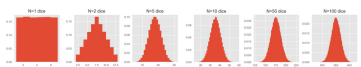
$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

$\mu = 0, \sigma = 1$ 标准正态分布

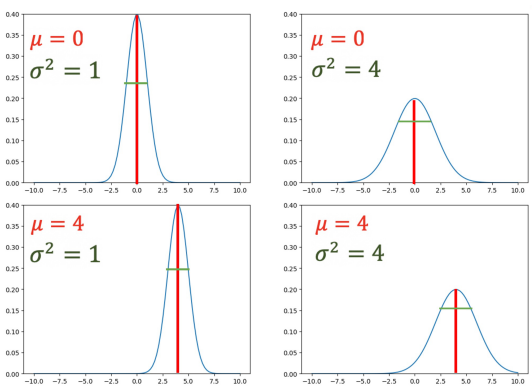
Normal Distribution



Demo



Examples



7. Expotienal distribution 指数分布

https://www.zhihu.com/question/24796044

- 指数分布解释 两事件发生（时间）间隔在无限长时间下的概率分布.

"要等到一个随机事件的发生，需要多长时间."

Parameter: λ

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & , x < 0. \end{cases}$$

λ 表示单位时间内事件发生的次数

$$F(x;\lambda) = \begin{cases} 1 - e^{-\lambda x} & , x \geq 0, \\ 0 & , x < 0. \end{cases}$$

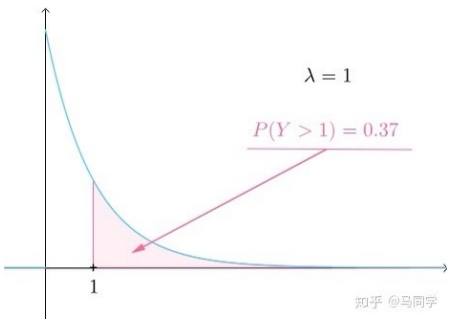
- 对一时间跨度 T，分为 $n \to +\infty$ 段，求出无事件发生（每段变量都为0）的概率.

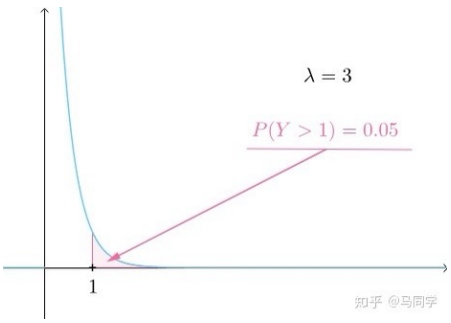
$$P_T = \lim_{n\to+\infty} (1 - \frac{\lambda T}{n})^n = e^{-\lambda T}$$ （泊松过程）

设随机变量 Y = 两事件发生的间隔（全局）

$$P(X \leq Y) = 1 - P_Y = 1 - e^{-\lambda Y}$$ （cdf form）

- 对其求导得到每点的概率分布函数 $p(x) = \begin{cases} \lambda e^{-\lambda x} & , x \geq 0 \\ 0 & , x < 0 \end{cases}$





Discrete

$E(X^2) - E^2(X)$

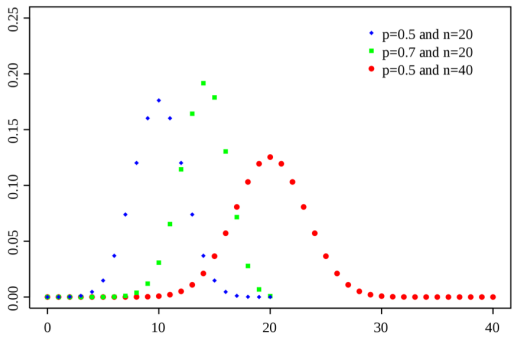
	p.m.f	Mean	Variance
Bernoulli; Ber(p)	$P(X=1) = p$ $P(X=0) = 1-p$	p	$p(1-p)$
Binomial; Bin(N,p)	$\Pr(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$	Np	$Np(1-p)$
Geometric; Geo(p)	$\Pr(X=k) = (1-p)^{k-1} p$	$\frac{1}{p}$	$\frac{1-p}{p^2}$?
Poisson; Poi(λ)	$P(X=k) = \frac{\lambda^k}{k!} e^{-\lambda}, k = 0, 1, \cdots$	λ	λ ?

proof of the mean of Geometric distribution:

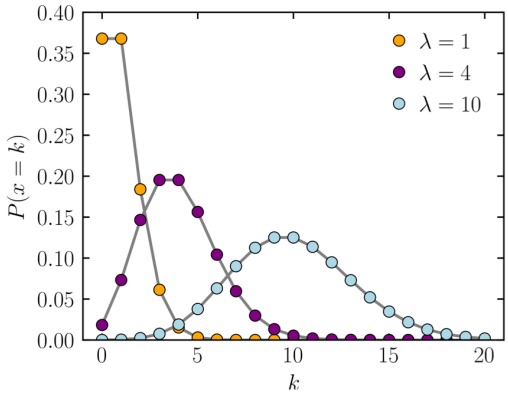
$$E(X) = \sum_{k=1}^{\infty} k \cdot (1-p)^{k-1} p$$
$$= p \cdot \sum_{k=1}^{\infty} k \cdot (1-p)^{k-1}$$
$$= p \cdot \frac{1}{p^2} = \frac{1}{p}$$

$$S(x) = \sum_{n=1}^{\infty} n \cdot x^{n-1}$$
$$\int S(x)\,dx = \sum_{n=1}^{\infty} \int n \cdot x^{n-1} dx$$
$$= \sum_{n=1}^{\infty} x^n \Big|_0^x = \sum_{n=1}^{\infty} x^n$$
$$= \frac{x}{1-x} \quad (0 < x < 1)$$
$$S(x) = \left(\frac{x}{1-x}\right)' = \frac{1}{(1-x)^2}$$

Discrete - visualization



Comparison of different binomial

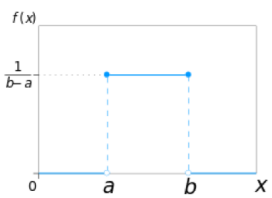


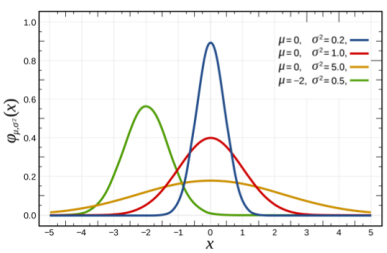
Comparison of different Poisson

我要和数学同归于尽

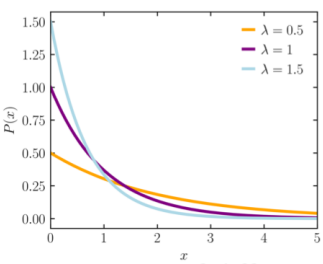
Continuous

	p.d.f	Mean	Variance
Uniform; Unif[a,b]	$f(x) = \frac{1}{b-a}, a \le x \le b$	$\frac{a+b}{2}$	$\frac{1}{12}(b-a)^2$
Normal; $N(\mu, \sigma^2)$	$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	μ	σ^2
Exponential; Exp(λ)	$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \ge 0, \\ 0 & , x < 0. \end{cases}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$





Comparison of different normal



Comparison of different exp

* Joint distribution 联合分布

- $f(x,y) = P(X=x, Y=y)$
- $\rho = \frac{Cov(x,y)}{\sqrt{Var(x)}\sqrt{Var(y)}}$, $\rho \in [-1,1]$
- $Cov(x,y) = E[(x-E[x])(Y-E[Y])] = E[xY] - E[x]\cdot E[Y]$ 协方差

  covariance indicates the correlation between x, y.

  Note: Correlation does not imply causality

  ☆ (线性)相关

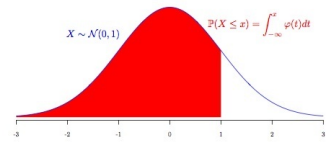
- $Var(\sum_{i=1}^{n} x_i) = Cov(\sum_{i=1}^{n} x_i, \sum_{j=1}^{n} x_j)$

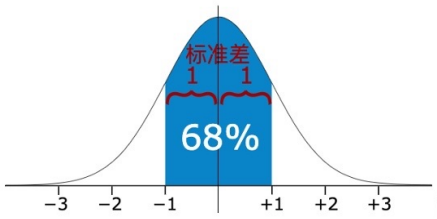
  $= \sum_{i=1}^{n}\sum_{j=1}^{n} Cov(x_i, x_j)$

  $= \sum_{i=1}^{n} Var(x_i) + 2\sum_{i<j} Cov(x_i, x_j)$

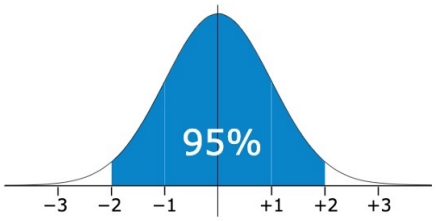
Additional Normal Distribution

z变换 $X \sim N(\mu, \sigma^2) \xrightarrow{z} \frac{X-\mu}{\sigma} \sim N(0,1)$

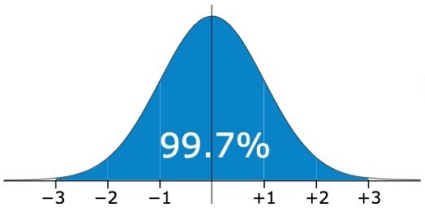




68%的数值是在离平均值1个标准差之内



95%的数值是在离平均值2个标准差之内



99.7%的数值是在离平均值3个标准差之内

知乎 @Brick何

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

z值表

正态分布的系数中为何会有π? - 知乎
https://zhuanlan.zhihu.com/p/74735645

对数正态分布

Exercise2-Discrete Random Variable

Let's assume that the number of customers X in a shopping mall in a day follows the Poisson Distribution with parameter λ, and the probability of each customer coming to the shopping mall for shopping is p. Please find which distribution the number of customers in the shopping mall for shopping in a day follows.

$Y = k. \quad k = 0, \dots, \infty$

$$P(Y=k) = \sum_i P(X=i)\, P(Y=k \mid X=i)$$

$$= \sum_{i=k}^{\infty} \frac{\lambda^i e^{-\lambda}}{i!} \times \binom{i}{k} p^k (1-p)^{i-k}$$

$$= \sum_{i=k}^{\infty} \frac{\lambda^i e^{-\lambda}}{i!} \times \frac{i!}{(i-k)!\,k!}\, p^k (1-p)^{i-k}$$

$$= \frac{p^k e^{-\lambda}}{k!} \sum_{t=0}^{\infty} \frac{(1-p)^t}{t!}\, \lambda^{t+k}$$

$$= \frac{p^k \lambda^k e^{-\lambda}}{k!} \sum_{t=0}^{\infty} \frac{1}{t!} [(1-p)\lambda]^t$$

$$= \frac{p^k \lambda^k e^{-\lambda}}{k!}\, e^{(1-p)\lambda}$$

$$= \frac{(p\lambda)^k}{k!}\, e^{-p\lambda}$$

Poisson : $p\lambda$.

Disjoint v.s. Independent v.s. Uncorrelated

Let A and B be two events, and $X \in \mathbb{R}$ and $Y \in \mathbb{R}$ be two random variables.

- **disjoint (mutually exclusive)**: $A \cap B = \emptyset$
  - ✓ A and B are joint does not imply that $\mathrm{P}(A \cap B) > 0$.
    counterexample:
    If $X \sim Unif[0,2]$, $A = \{0 \le X \le 1\}$, $B = \{1 \le X \le 2\}$, then $A \cap B = \{X = 1\} \neq \emptyset$, but $\mathrm{P}(A \cap B) = \mathrm{P}(X = 1) = \int_1^1 \frac{1}{2} \mathrm{d}x = 0$.
    note :
    Given event C, $C = \emptyset$ implies $\mathrm{P}(C) = 0$, but $\mathrm{P}(C) = 0$ cannot imply $C = \emptyset$, i.e. $\mathrm{P}(C) = 0$ does not mean that event C is impossible.

- **independent**:
  - for events, P(AB)=P(A)P(B);
  - for random variables, $f_{X,Y}(x,y) = f_X(x) f_Y(y)$.

- **uncorreated**: $\mathrm{Cov}(X,Y) = 0 \Leftrightarrow \mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$.

disjoint $\not\rightleftarrows$ independent $\not\rightleftarrows$ uncorrelated

1 If events A and B are joint, then $\mathrm{P}(A \cup B) < \mathrm{P}(A) + \mathrm{P}(B)$ always holds.

False. As mentioned before, even if A and B are joint, it is still possible that $\mathrm{P}(A \cap B) = 0$. Since $\mathrm{P}(A \cap B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cup B)$, i.e. it is possible that $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$.

4 For some distributions, independence is equivalent to uncorrelatedness.

True. For example, joint normal distribution and Bernoulli distribution.

5 If random variables X and Y are correlated, then X and Y are also dependent.

True. The contrapositive statement of "independence implies uncorrelatedness" is "correlatedness implies dependence".

The nine players on a basketball team consists of 2 centers, 3 forwards, and 4 backcourt players. If the players are paired up at random into three groups of size 3 each, find the expected value of the number of triplets consisting of one of each type of player.

Solution 2

Let X_i denote

$$X_i = \begin{cases} 1, & \text{the } i\text{th triple consists of one of each type of player,} \\ 0, & \text{otherwise.} \end{cases} . \tag{1}$$

Then we can calculate the probabilities of $X_i = 1$ and $X_i = 0$,

$$\mathrm{P}(X_i = 1) = \frac{\binom{2}{1}\binom{3}{1}\binom{4}{1}}{\binom{9}{3}} = \frac{2}{7}, \ \mathrm{P}(X_i = 0) = 1 - \mathrm{P}(X_i = 1) = \frac{5}{7}, \tag{2}$$

which also means that X_i follows a Bernoulli distribution with the parameter $2/7$.

Recall the properties of Bernoulli distribution, if the probability mass function of X is $P(X = 1) = p$ $(0 \leq p \leq 1)$ and $P(X = 0) = 1 - p$, the X follows a Bernoulli distribution with the parameter p and

$$\mathrm{E}(X) = p, \ \mathrm{Var}(X) = p(1-p). \tag{3}$$

It follows that $\mathrm{E}(X_i) = 2/7$. Hence, we have

$$\mathrm{E}\left(\sum_{i=1}^{3} X_i\right) = \sum_{i=1}^{3} \mathrm{E}(X_i) = \frac{6}{7}. \tag{4}$$

Sampling 采样 calculate E[g(x)]

Part 1. simple X, complicated g.

理论: $E[g(x)] = \sum_{w \in \Omega} g(w)P(w) = \frac{1}{N}\sum_{i=1}^{N} g(x_i)$ as $N \to \infty$.

Demo 3

$\int_0^2 e^{x+\cos x}\,dx = \int_0^2 \frac{e^{x+\cos x}}{g(x)} g(x)\,dx$ → pdf. 满足 $\int_0^2 g(x)dx = 1$

- How to calculate
- $\int_0^2 e^{x+\cos(x)} dx = \int_0^2 \frac{2\, e^{x+\cos(x)}}{2} dx = \int_0^2 2e^{x+\cos(x)} f(x)dx$ → P(x)

uniform distribution

- Construct a random variable uniformly distributed in [0,2]
- Notice that the pdf is now 1/2. (As the total probability is 1)

trick

1、$X_1, X_2, \cdots, X_n$ in uniform [0,2] (use python build-in function)

2、$\frac{1}{N}\sum_{i=1}^{N} 2e^{x_i+\cos x_i}$

$E[2e^{X+\cos(X)}]$

- How to calculate
- $\int_0^\infty e^{-x+\cos(x)}dx = \int_0^\infty e^{\cos x} \cdot e^{-x} dx = E[e^{\cos(x)}]$

exponential distribution

pdf.

- Construct a random variable with pdf $f(x) = e^{-x}$
  - Exponential distribution
  - Check $\int_0^\infty e^{-x}dx$=1

× 用 Exponential distribution 生成样本空间.

- $E[e^{\cos(X)}]$

Part 2. complicated X, simple g.

Case 1: know how the data is generated, but PDF/CDF is hard to compute.

simulation.

Case 2: know PDF/CDF, but they are complicated.

1) know a nontrival CDF. Inverse Transform Method (ITM)

性质: 记随机变量 X 的 cdf 为 F_X, 则 $X = F_X^{-1}(U)$, $U \sim U(0,1)$.

proof: $P(F_X^{-1}(U) \le x) = P(U \le F_X(x)) = F_U(F_X(x)) = F_X(x)$

∴ 随机变量 X 与 $F_X^{-1}(U)$ 同分布. $F_X^{-1}(u) \triangleq \inf\{t : f(t) = u\}$.

方法: 用 U 生成样本 Ω_U.

$\Omega_U \xrightarrow{F_X^{-1}(U)} \Omega_X$.

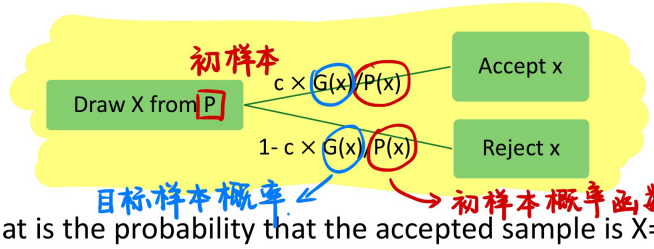
2) know a nontrival PDF. The Acceptance / Rejection Method. (ARM)

Principle

(初样本) X: simple
(目标样本) G: complicated

- There is a generator for Random variable X with pmf $P(x)$.

Step 1: Draw a sample X from P
Step 2: If X=x, I accept it with probability c × G(x)/P(x);
otherwise, I return to step 1 until I accept a sample.



- What is the probability that the accepted sample is X=x? Ans: G(x).

  - For each sample X=x, I accept it with probability c × G(x)/P(x).
  - Then how many samples with X=x will be accepted?
  - N × $P(x)$ × c × G(x)/P(x) = N × c × G(x).

- Among the accepted samples, the portion of X=x?

Σ_x G(x) =1

$$\frac{N \times c \times G(x)}{\Sigma_x N \times c \times G(x)} = \frac{G(x)}{\Sigma_x G(x)} = G(x)$$

Remark: $c \times G(x)/P(x) < 1$, c is a arbitrary constant.

选择合适的初样本很重要.

Statistics inference 统计推断

Point estimator : $\hat{\theta}$

样本随机变量的函数

And we call $\hat{\theta}$ the statistic. 统计量 $\hat{\theta}$ is also a random variable.

- $\hat{\theta}$ is a function of samples $(X_1, X_2 \ldots, X_n)$
- $\hat{\theta}$ extracts some (useful) information from the samples.

Maximum likelihood estimate (MLE)

Given a model with an unknown parameter θ, the probability that the model generate the samples is called likelihood 似然. $L(\theta) = P(X_1, X_2, \ldots, X_n \mid \theta)$

$\hat{\theta} = \text{argmax}(L(\theta))$

e.g. 1 Assume whether a drug could cure a disease is Bernoulli RV. p is unknown. For given N drug experiments, M successes.

$$L(p) = p^M (1-p)^{N-M}$$

$$\ln L(p) = M \ln p + (N-M)\ln(1-p)$$

$$[\ln L(p)]' = \frac{M}{p} - \frac{N-M}{1-p} = 0 \quad \Rightarrow \hat{p} = \frac{M}{N}$$

$$[\ln L(\hat{p})]'' < 0 \quad \therefore \hat{p} \text{ is the best.}$$

e.g. 2

- Suppose we aim to use normal distribution to fit the data.
  - Given parameter μ, the model is $N(\mu, 1)$ Mean μ, variance 1
- Samples: $X_1, X_2 \ldots, X_n$

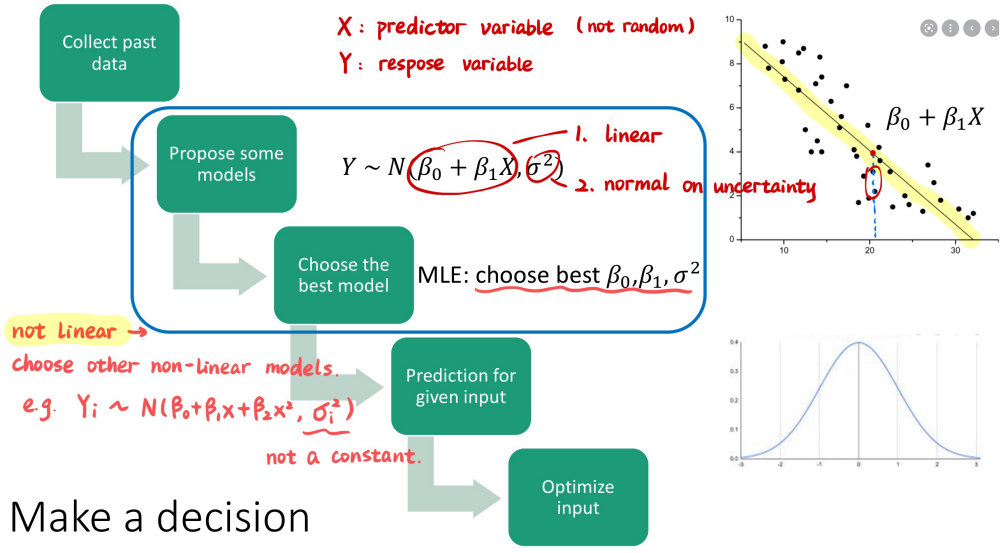
$$L(\mu) = \prod_{i=1}^{n} f(x_i, \mu) = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

$$\ln L(\mu) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$[\ln L(\mu)]' = 0 \Rightarrow \mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

线性回归 linear regression

1) propose models



2) choose the best parameters (MLE).

- For the model with $\beta_0, \beta_1, \sigma^2$, the likelihood is

似然函数： $$\frac{1}{(\sqrt{2\pi})^n \sigma^n} \exp\left[-\frac{1}{2}\frac{\Sigma_i(Y_i-\beta_1 X_i-\beta_0)^2}{\sigma^2}\right]$$

- Given σ^2, to maximize the likelihood, we only need to minimize

$$\Sigma_i(Y_i-\beta_1 X_i-\beta_0)^2$$

- Taking derivative over β_0 and β_1, we have (partial derivative)

$$\frac{\partial f}{\partial \beta_1} = \Sigma_i(Y_i-\beta_1 X_i-\beta_0)X_i = 0$$

$$\frac{\partial f}{\partial \beta_0} = \Sigma_i(Y_i-\beta_1 X_i-\beta_0) = 0$$

MLE:

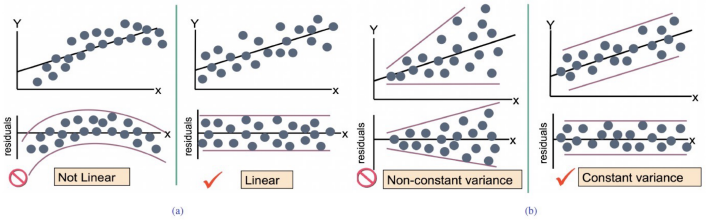
$$\widehat{\beta_1} = \frac{\Sigma_i(X_i-\bar{X})(Y_i-\bar{Y})}{\Sigma_i(X_i-\bar{X})^2} = \frac{\Sigma X_i Y_i - N\bar{x}\bar{y}}{\Sigma x_i^2 - N\bar{x}^2}$$

$$\widehat{\beta_0} = \bar{Y} - \widehat{\beta_1}\bar{X}$$

3) Residual Analysis (check assumptions)

Residual: $e_i = Y_i - \beta_0 - \beta_1 X_i$

- Check the assumptions of regression by examining the residuals
  - Examine for linearity assumption: e_i *does not depend on* X_i
  - Evaluate constant-variance assumption: variance of e_i *does not depend on* X_i
- Graphical Analysis of Residuals: Can plot residuals vs. X



Hypothesis Testing
假设检验.(没细讲)

* Additional example

Baseball Team

- The weight for a baseball team players are

  {150, 143, 132, 160, 175, 190, 123, 154}

- Assume their weights are **uniformly** distributed over an interval [a, b]

- What are good estimators for a? for b?

This example will show that the **MLE could be complicated to solve**, e.g., the equation $l'(\theta) = 0$ may be difficult to solve, or it may **not** always to possible to use calculus methods **directly** to find the maximum of L(θ).

MLE: Uniform

Let *X* be a Uniform random variable on the interval [0, θ]

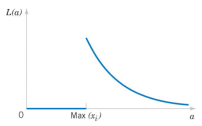
$$f(x;\theta) = \begin{cases} \frac{1}{\theta}, & \text{for } 0 \le x \le \theta, \\ 0, & \text{otherwise,} \end{cases} = \frac{1}{\theta}\mathbf{1}_{\{0\le x\le\theta\}}$$

indicator function $\mathbf{1}_A(x)$

$$\mathbf{1}_A(x) = \begin{cases} 1, & \text{if } x \in A, \\ 0, & \text{otherwise} \end{cases}$$

The likelihood function of a random sample of size *n* is:

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta) = \frac{1}{\theta^n}\prod_{i=1}^{n}\mathbf{1}_{\{0\le x_i\le\theta\}} = \begin{cases} \frac{1}{\theta^n}, & \text{if } \theta \ge max\{X_1, X_2, \dots, X_n\} \\ 0, & \text{if } \theta < max\{X_1, X_2, \dots, X_n\} \end{cases}$$



$$\hat{\theta} = max\{X_1, X_2, \dots, X_n\}$$

Calculus methods don' t work here because $L(\theta)$ is maximized at the **discontinuity**. Clearly, θ cannot be smaller than max(x_i), thus the MLE is $max\{X_1, X_2, \dots, X_n\}$.

1. **(25 points)** In the case study of linear regression problems, we assumed that the value of each X_i (e.g., the heights of parents) will affect the value of the response variable Y_i (e.g., the height of their adult children). We use the normal distribution:

$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2), \quad i = 1, 2, \ldots, N,$$

to model the distribution of Y_i given each X_i. We also assume that $Y_1, \ldots, Y_N$ are independent, and the variance of $Y_i - \beta_0 - \beta_1 X_i$ for every $i = 1, \ldots, N$ is the same (i.e., all equal to σ^2).

Which of the following statements is false?

A. When we derive the maximum likelihood estimate (MLE) of β_0 and β_1, maximizing the likelihood function is equivalent to maximizing the log-likelihood function.
B. The linear regression model assumes that, for $i = 1, \cdots, N$, $Y_i - \beta_0 - \beta_1 X_i$ are independent.
C. After we obtain the MLE $\hat{\beta}_0$ and $\hat{\beta}_1$, we may use residual analysis to check our assumptions. If our assumptions are correct, the residuals $e_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$, $i = 1, \cdots, N$, are independent.
D. The linear regression model assumes that, for $i = 1, \cdots, N$, the relationship between Y_i and X_i is linear.

Solution: C.
Residuals $e_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$ are dependent with each other. Note that residuals $Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$ and errors $Y_i - \beta_0 - \beta_1 X_i$ are different.

Confidence interval 置信区间

中心极限定理：

给定一个任意分布的总体，每次从中随机抽取 n 个样本取平均值，多组抽样的分布渐近于正态分布.

设一组样本为 $X_1, X_2, \ldots, X_n$，服从均值为 μ，方差为 σ^2 的分布.

则 $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.

$$Z = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \sim N(0,1)$$ 平移伸缩变换

$$\Rightarrow \quad P\left(-z_{\alpha/2} \le \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \le z_{\alpha/2}\right) = 1-\alpha$$

$$\Rightarrow \quad P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

With probability 1 - α, μ is within the interval $\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$

再作一次近似？

You can also write as $\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

There are four assumptions associated with a linear regression model:

1. **Linearity**: The relationship between X and the mean of Y is linear.
2. **Homoscedasticity**: The variance of residual is the same for any value of X.
3. **Independence**: Observations are independent of each other.
4. **Normality**: For any fixed value of X, Y is normally distributed.

Linear Regression Model (1/2)

- **Samples**: $(X_1, Y_1), \cdots, (X_n, Y_n)$
- **Likelihood Function**: Since

$$Y_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2), \tag{1}$$

the likelihood function and its logarithm can be given as

$$L(\beta_1, \beta_0) = \Pi_{i=1}^{n} f(Y_i - \beta_1 X_i - \beta_0) \tag{2}$$

$$= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left[-\frac{1}{2}\frac{\sum_{i=1}^{n}(Y_i - \beta_1 X_i - \beta_0)^2}{\sigma^2}\right], \tag{3}$$

$$\log[L(\beta_1, \beta_0)] = \underbrace{-\frac{n}{2}\log(2\pi) - n\log(\sigma)}_{constant} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_1 X_i - \beta_0)^2. \tag{4}$$

Linear Regression Model (2/2)

- **First-order derivatives**

  Calculate the first-order derivatives of $\log[L(\beta_1, \beta_0)]$ with respect to β_1 and β_0, respectively, and set these derivatives be zero, we have

$$\sum_{i=1}^{n}(Y_i - \hat{\beta}_1 X_i - \hat{\beta}_0)X_i = 0, \tag{5}$$

$$\sum_{i=1}^{n}(Y_i - \hat{\beta}_1 X_i - \hat{\beta}_0) = 0. \tag{6}$$

- **MLEs of β_1 and β_0**

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}. \tag{7}$$

Central limit theorem: No matter what the true distribution is, the sample mean will be very close to the normal distribution, as long as the sample size is large.

Optimization 最优化

研究的问题：

$$\begin{aligned} &\text{minimize} && f(x) \quad \rightarrow \text{objective function} \\ &\text{subject to} && g_i(x) \le 0,\ i=1,2,\dots,n \\ & && h_i(x) = 0,\ i=1,2,\dots,n. \end{aligned} \quad \Big\} \text{constraints.}$$

decision variables ←

feasible set (可行解)

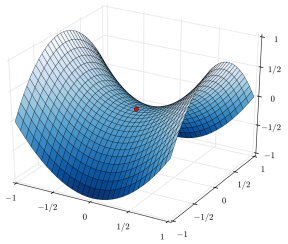
local minimizer (x^*) : if $\exists\, \varepsilon>0$, for any $y \in S \cap B(x,\varepsilon)$, s.t $f(x) \le f(y)$.

Convex Optimization 凸优化

凸函数的优势：局部最优 = 全局最优

convex set 凸集 for any $x_1, x_2 \in C$, $\theta x_1 + (1-\theta)x_2 \in C$, $0 \le \theta \le 1$.

Lemma : if S_1, S_2 are convex sets, then $S_1 \cap S_2$ is also a convex set.

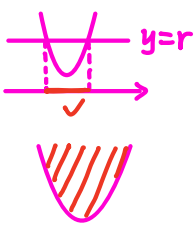


convex function 凸函数

A function $f(x): R^n \to R$ is called convex if

① the domain of f is convex set.

② for any $x, y \in dom(f)$ and $0 \le \lambda \le 1$, $f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y)$



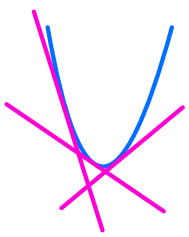
Fact 1: $C = \{x : f(x) \le r\}$ is a convex set if $f(x)$ is a convex function.

Fact 2: $C = \{(x,y) : y \ge f(x)\}$ is a convex set if $f(x)$ is a convex function.

if epi f is a convex set, f is a convex function.

Lemma: A negative concave function is a convex function and vice versa.

凸函数判定：



(FOC) (1) dom(f) is convex.

(2) $\forall\, x, y \in f$. $y = x + s\vec{e}$

$$f(y) \ge f(x) + s \cdot \frac{df(x+\theta\vec{e})}{d\theta}\Big|_{\theta=0}.$$

(SOC) (1) dom(f) is convex.

(2) For $\forall\, \theta$ and unit vector $\vec{e}$, $\frac{d^2(x+\theta\vec{e})}{d\theta^2} \ge 0$

Additional slides

FOC参考证明：

Necessity

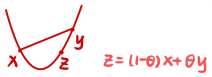
convex → 绕论

$y = x + se$

$$f(y) \geq f(x) + s\frac{d\,f(x+\theta e)}{d\theta}|_{\theta=0}$$

For any $x, y \in$ **dom f,**

- denote e as the unit vector from x to y
- let $y = x + s\,e$



$z = (1-\theta)x + \theta y$

$(1-\theta)f(x) + \theta f(y) \geq f(z)$

$f(y) \geq f(x) + \frac{f(z)-f(x)}{\theta} = f(x) + \frac{f(x+\theta s\vec{e}) - f(x)}{\theta}$

$= f(x) + s\cdot\frac{df(x+\theta\vec{e})}{d\theta}$ $(\theta \to 0)$

Let $z = (1-\theta)x + \theta y = x + \theta\, s\, e$, by convexity of f, we have

$$(1-\theta)\,f(x) + \theta f(y) \geq f(z)$$

$$f(y) \geq f(x) + \frac{f(z)-f(x)}{\theta} = f(x) + \frac{f(x+\theta\, s\, e)-f(x)}{\theta}$$

When $\theta \longrightarrow 0$,

$$s\,\frac{d\,f(x+\theta e)}{d\theta}|_{\theta=0}$$

Sufficiency

$y = x + se$

$$f(y) \geq f(x) + s\frac{d\,f(x+\theta e)}{d\theta}|_{\theta=0}$$

- Denote e as the unit vector from x to y
- Let $y = x + s\,e$
- Let $z = (1-t)x + t\,y = x + \mathrm{t}\,\mathrm{s}\,e$,

$\begin{cases} (1-t)f(x) \geq (1-t)f(z) - (1-t)ts\,\frac{df(z+\theta e)}{d\theta}|_{\theta=0} \\ t\,f(y) \geq t f(z) + st(1-t)\,\frac{df(z+\theta e)}{d\theta}|_{\theta=0} \end{cases}$

$\Rightarrow (1-t)f(x) + tf(y) \geq f(z)$.

$\Leftrightarrow$ f is convex function.

As a result, $x = z - \mathrm{t}\,\mathrm{s}\,e$, $y = z + (s - t\,s)\,e$,

distance

$$f(x) \geq f(z) - ts\,\frac{d\,f(z+\theta e)}{d\theta}|_{\theta=0} \text{ and } f(y) \geq f(z) + s(1-t)\,\frac{d\,f(z+\theta e)}{d\theta}|_{\theta=0}$$

Multiplying the first inequality by 1-t, the second by t, and adding them yields

$$(1\text{-}t)\,f(x) + t\,f(y) \geq f(z)$$

SOC参考证明：

Necessity

已知 g 为凸函数，证明 $g'' > 0$.

$y = x + se$

$g(\theta; x, e) = f(x + \theta e)$

$f(y) \geq f(x) + sg'(0; x, e)$

For any x **and** θ**,**

- let $y = x + (s+\theta)e$ and $z = x + \theta\, e$ with s >0
- $y = z + \mathrm{s}\,e$ and $z = y - s\,e$

By FOC of f, we have

$$f(y) \geq f(z) + \mathrm{s}\,g'(0; z, e), f(z) \geq f(y) - sg'(0; y, e)$$

Sum both sides together,

$$g'(0; y, e) - g'(0; z, e) \geq 0, \text{ then } \frac{g'(0; y, e) - g'(0; z, e)}{s} \geq 0$$

As g(t; y, e) = $f(y + te) = f(x + (t + s + \theta)e) = g(t + s + \theta; x, e)$, g'(t; y, e)= g'(t + $s + \theta$; x, e). Similarly, g'(t; z, e)= g'(t+θ; x, e). Accordingly,

$$0 \leq \frac{g'(0; y, e) - g'(0; z, e)}{s} = \frac{g'(s+\theta; x, e) - g'(\theta; x, e).}{s}$$

When s is approaching 0, $g''(\theta; x, e) \geq 0$

Sufficiency

$y = x + se$

$g(\theta; x, e) = f(x + \theta e)$

$f(y) \geq f(x) + sg'(0; x, e)$

已知 $g'' > 0$，证明 g 为凸函数.

- Denote e as the unit vector from x to y (chosen arbitrarily)
- Let $y = x + \mathrm{s}\,e$, then

$$0 \leq \int_0^s (s-t)\,g''(\mathrm{t}; x, e)\,\mathrm{dt}$$

$$= -s\,g'(0; x, e) + \int_0^s g'(\mathrm{t}; x, e)\,\mathrm{d\,t}$$

$$= -s\,g'(0; x, e) + g(s; x, e) - g(0; x, e)$$

$$= -\,s\,g'(0; x, e) + f(y) - f(x)$$

$$\int_a^b u(x)v'(x)\,dx = \Big[u(x)v(x)\Big]_a^b - \int_a^b u'(x)v(x)\,dx$$

$$= u(b)v(b) - u(a)v(a) - \int_a^b u'(x)v(x)\,dx.$$

Integration by parts

- Then $f(y) \geq f(x) + s\,g'(0; x, e)$

FOC:

Suppose f is differentiable (*i.e.*, its gradient ∇f exists at each point in **dom** f, which is open). Then f is convex if and only if **dom** f is convex and

$$f(y) \geq f(x) + \nabla f(x)^T(y - x) \tag{3.2}$$

holds for all $x,\ y \in$ **dom** f. This inequality is illustrated in figure 3.2.

SOC:

We now assume that f is twice differentiable, that is, its *Hessian* or second derivative $\nabla^2 f$ exists at each point in **dom** f, which is open. Then f is convex if and only if **dom** f is convex and its Hessian is positive semidefinite: for all $x \in$ **dom** f,

$$\nabla^2 f(x) \succeq 0.$$

凸函数的性质：（假设 f 是凸函数，则 g 为凸函数）

1. Composition — $g = f(ax+b) - c$
2. Elementwise maximum — $g = \max\{f_1(x), f_2(x), \ldots, f_m(x)\}$.
3. Nonnegative weighted sums — $g = w_1 f_1 + w_2 f_2 + \ldots + w_n f_n \quad (w > 0)$.

These properties extend to infinite sums and integrals. For example if $f(x, y)$ is convex in x for each $y \in \mathcal{A}$, and $w(y) \geq 0$ for each $y \in \mathcal{A}$, then the function g defined as

$$g(x) = \int_{\mathcal{A}} w(y) f(x, y)\, dy$$

integration form.

is convex in x (provided the integral exists).

4. (special) Minimization

If f is convex in (x, y), and C is a convex nonempty set, then the function

$$g(x) = \inf_{y \in C} f(x, y)$$

is convex.

We prove this by verifying Jensen's inequality for $x_1,\ x_2 \in \mathbf{dom}\, g$. Let $\epsilon > 0$. Then there are $y_1,\ y_2 \in C$ such that $f(x_i, y_i) \leq g(x_i) + \epsilon$ for $i = 1,\ 2$. Now let $\theta \in [0, 1]$. We have

$$\begin{aligned} g(\theta x_1 + (1-\theta)x_2) &= \inf_{y \in C} f(\theta x_1 + (1-\theta)x_2, y) \\ &\leq f(\theta x_1 + (1-\theta)x_2, \theta y_1 + (1-\theta)y_2) \\ &\leq \theta f(x_1, y_1) + (1-\theta) f(x_2, y_2) \\ &\leq \theta g(x_1) + (1-\theta) g(x_2) + \epsilon. \end{aligned}$$

Since this holds for any $\epsilon > 0$, we have

$$g(\theta x_1 + (1-\theta)x_2) \leq \theta g(x_1) + (1-\theta) g(x_2).$$

Example 3.16 *Distance to a set.* The distance of a point x to a set $S \subseteq \mathbf{R}^n$, in the norm $\|\cdot\|$, is defined as

$$\mathbf{dist}(x, S) = \inf_{y \in S} \|x - y\|.$$

The function $\|x-y\|$ is convex in (x, y), so if the set S is convex, the distance function $\mathbf{dist}(x, S)$ is a convex function of x.

The Newsvendor Problem

- A newsboy needs to decide how many newspapers to buy each morning in order to generate the most expected sales profit. The demand D for newspapers in a day is random, with f(·) being its pdf and F(·) being its CDF. He needs to buy each newspaper at price c, and can sell it at price p.

买了q，已经知

$E(q) = E[\min(q, D)] \cdot P - Cq$

$E[\min(q, D)] = \sum x \cdot f(x) = \sum_{D \leq q} D f(D) + \sum_{D > q} q \cdot f(q)$

$= \int_0^q x f(x)\, dx + q(1 - F(q))$

- Let the newsboy's purchase decision be q. His objective is

$$\text{Maximize } E[\text{Profit}]_q = E[p \min(q, D)] - cq.$$

We can write

$$E[\min(q, D)]_q = \int_0^q x f(x)\, dx + q(1 - F(q)).$$

$D \leq q$ $\quad$ $D > q$

So

$$E[\text{Profit}] = p \int_0^q x f(x)\, dx + pq(1 - F(q)) - cq.$$

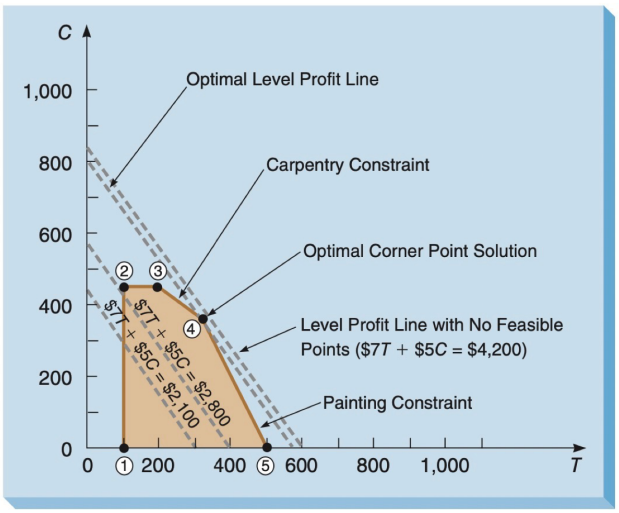
Is E[Profit] concave?

- $E[\text{Profit}] = p \int_0^q x f(x)\, dx + pq(1 - F(q)) - cq.$
- SOC: $\frac{\partial}{\partial q} E[\text{profit}] = pqf(q) + pq(-F'(q)) + p[1 - F(q)] - c = p[1 - F(q)] - c$

$\frac{\partial^2}{\partial q^2} E[\text{profit}] = p[-F'(q)] \leq 0$ (note F() is an increasing function)

$\frac{\partial}{\partial q} E[Profit] = 0$: $q^* = F^{-1}\left(\frac{p-c}{p}\right)$.

Linear programming model



$$E(q) = p\int_0^q x f(x)\,dx + pq(1-F(q)) - cq$$

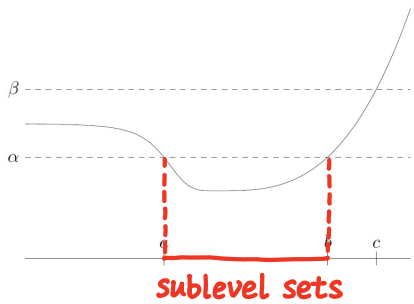
$$\frac{dE}{dq} = pqf(q) + p[(1-F(q)) + q(-f(q))] - c$$

$$\frac{d^2E}{dq^2} = p(-f(q)) = -pf(q) < 0.$$

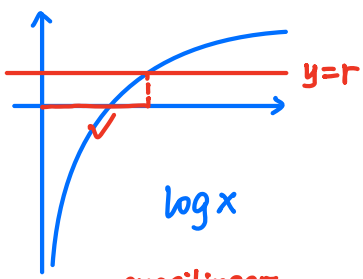
拟凸函数

- A function $f : \mathrm{R}^n \to \mathrm{R}$ is quasiconvex if all its sublevel sets are convex. i.e. if $S_\alpha = \{x \mid f(x) \le \alpha\}$ is convex for each $\alpha \in \mathrm{R}$.
  - f is quasiconcave if $-f$ is quasiconvex.
    - Equivalently, all its superlevel sets $S_\alpha = \{x \mid f(x) \ge \alpha\}$ are convex.
  - f is quasilinear if it is both quasiconvex and quasiconcave
    - Equivalently, all its sublevel and superlevel sets are halfspaces, and all its level sets are affine

definition

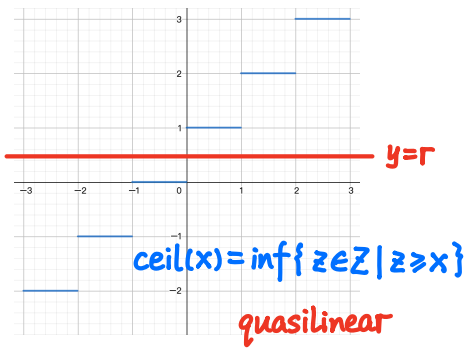


sublevel sets



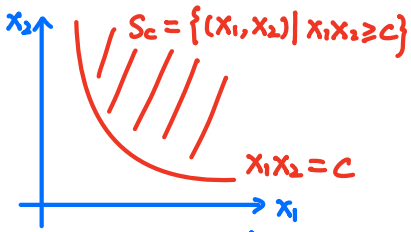
log x

quasilinear



ceil(x) = inf{z∈Z | z≥x}

quasilinear



x_1x_2 on $\mathbb{R}^2_+$

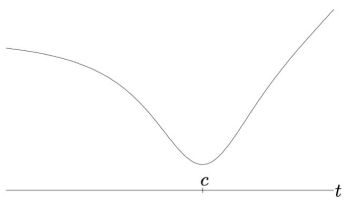
quasiconcave

拟凸函数的性质.

性质一：

A continuous function $f : \mathbf{R} \to \mathbf{R}$ is quasiconvex if and only if at least one of the following conditions holds:

- f is nondecreasing
- f is nonincreasing
- there is a point $c \in \mathbf{dom}\, f$ such that for $t \le c$ (and $t \in \mathbf{dom}\, f$), f is nonincreasing, and for $t \ge c$ (and $t \in \mathbf{dom}\, f$), f is nondecreasing.



性质二：

f is quasiconvex function if and only if:

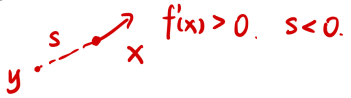
$$f(\theta x+(1-\theta)y) \le \max\{f(x), f(y)\}.$$

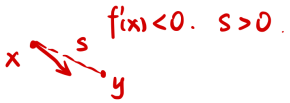
性质三：(FOC)

- A differentiable $f : \mathrm{R}^n \to \mathrm{R}$ is quasiconvex if and only if
  - For any $f(\mathbf{y}) \le f(\mathbf{x})$, letting $\mathbf{y} = \mathbf{x} + s\,\mathbf{e}$, we have

$$s\frac{d\,f(x+te)}{dt}\Big|_{t=0} \le 0$$

One dimension: $f(y) \le f(x) \Rightarrow f'(x)(y-x) \le 0$.





性质四：(SOC)

Given f is differentiable.

- if f is quasiconvex, then for any x and e.

if $\frac{df(x+te)}{dt}\Big|_{t=0} = 0$, we always have $\frac{d^2f(x+te)}{dt^2}\Big|_{t=0} > 0$.

Scaling

If f is quasiconvex and $w > 0$, then wf is also quasiconvex.

f and wf have the same sublevel sets: $wf(x) \le \alpha$ iff $f(x) \le \alpha/w$,

Composition with Nondecreasing Function

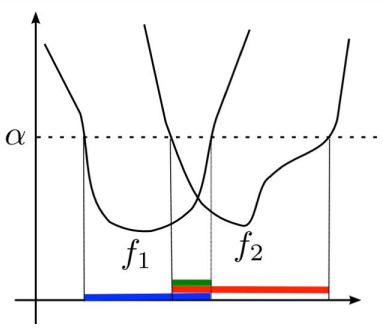
If $f : \mathbb{R}^n \to \mathbb{R}$ is quasiconvex $h : \mathbb{R} \to \mathbb{R}$ is non-decreasing, then $h \circ f$ is quasiconvex.

$h \circ f$ and f have the same sublevel sets: $h(f(x)) \le \alpha$ iff $f(x) \le h^{-1}(\alpha)$

Maximum

If f_1, f_2 are quasiconvex, then $g(x) = \max\{f_1(x), f_2(x)\}$ is also quasiconvex.

Generalizes to the maximum of any number of functions, $\max_{i=1}^k f_i(x)$, and also to the supremum of an infinite set of functions $\sup_y f_y(x)$.

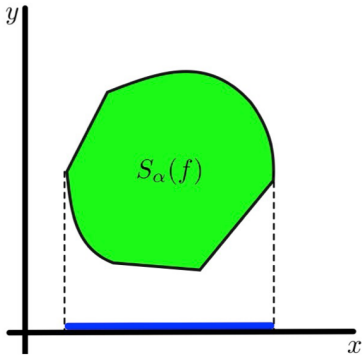


Minimization

If $f(x, y)$ is quasiconvex and $\mathcal{C}$ is convex and nonempty, then $g(x) = \inf_{y \in C} f(x, y)$ is quasiconvex.

Proof (for $\mathcal{C} = \mathbb{R}^k$)

$S_\alpha(g)$ is the projection of $S_\alpha(f)$ onto hyperplane $y = 0$.



Sum

$f_1 + f_2$ is NOT necessarily quasiconvex when f_1 and f_2 are quasiconvex.

Exercise 1

Convexity

Prove the following statements:

1. $\frac{1}{n}\sum_{i=1}^{n} x_i \geq (\Pi_{i=1}^{n} x_i)^{1/n}$
2. $f(\mathbf{x}) = (\Pi_{i=1}^{n} x_i)^{1/n}$ is concave where $x_i > 0$ for all i.

Solution:

1. As $log(t)$ is concave, then $f((1-\lambda)x+\lambda y) \geq (1-\lambda)f(x) + \lambda f(y)$

$$\log(\frac{1}{n}\sum_{i=1}^{n} x_i) \geq \frac{1}{n}\sum_{i=1}^{n} \log x_i = \frac{1}{n}\log \Pi_{i=1}^{n} x_i.$$

As a result,

$$\frac{1}{n}\sum_{i=1}^{n} x_i \geq (\Pi_{i=1}^{n} x_i)^{1/n}$$

Solution:

2.

$$\frac{f(\mathbf{x})}{f(\lambda\mathbf{x}+(1-\lambda)\mathbf{y})} = (\Pi_{i=1}^{n} \frac{x_i}{\lambda x_i + (1-\lambda)y_i})^{1/n} \leq \frac{1}{n}\sum_{i=1}^{n} \frac{x_i}{\lambda x_i + (1-\lambda)y_i}$$

Similarly

$$\frac{f(\mathbf{y})}{f(\lambda\mathbf{x}+(1-\lambda)\mathbf{y})} \leq \frac{1}{n}\sum_{i=1}^{n} \frac{y_i}{\lambda x_i + (1-\lambda)y_i}$$

$+ \quad = 1$

Now

$$\lambda\frac{f(\mathbf{x})}{f(\lambda\mathbf{x}+(1-\lambda)\mathbf{y})} + (1-\lambda)\frac{f(\mathbf{y})}{f(\lambda\mathbf{x}+(1-\lambda)\mathbf{y})} \leq \frac{1}{n}\sum_{i=1}^{n} \frac{\lambda x_i + (1-\lambda)y_i}{\lambda x_i + (1-\lambda)y_i} = 1.$$

As $f(\lambda\mathbf{x}+(1-\lambda)\mathbf{y}) > 0$, we have

$$\lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{y}) \leq f(\lambda\mathbf{x}+(1-\lambda)\mathbf{y})$$

Exercise 2

Convex Function

Suppose $f(t) \in \mathbb{R}$ for $t > 0$ is convex and differentiable. Show that

$$F(x) = \frac{1}{x}\int_0^x f(t)dt, \quad x > 0,$$

is convex. **Hint**: Use the first-order condition of convex function.

Solution 1

F is differentiable with

$$\begin{aligned} F'(x) &= -(1/x^2)\int_0^x f(t)dt + f(x)/x, \\ F''(x) &= (2/x^3)\int_0^x f(t)dt - 2f(x)/x^2 + f'(x)/x \\ &= (2/x^3)\int_0^x [f(t) - f(x) - f'(x)(t-x)]\,dt \geq 0, \end{aligned}$$

f is convex.

which is derived from the first-order condition of the convex f:

$$f(t) \geq f(x) + f'(x)(t-x) \text{ for all } x, t > 0.$$

Thus F is convex.

Solution 2

If we set $t = xz$ and let z be the integration variable, then the integration domain is changed to $[0, 1]$. As a result,

$$\begin{aligned} F(x) &= \frac{1}{x}\int_0^x f(t)dt \\ &= \frac{1}{x}\int_0^1 f(xz)d(xz) \\ &= \int_0^1 f(xz)dz \end{aligned}$$

As f is a convex, so $f''(x) \geq 0$. Then

$$F''(x) = \int_0^1 z^2 f''(xz)dz \geq 0$$

which implies the convexity of F.

The objective is equivalent to min $f(x) = -(30x_1 + 20x_2 + 40x_3 + 25x_4 + 10x_5)$. And $f(x)$ is linear so it's convex. However, since the feasible set is set of integers and thus not convex.
Therefore, it's not a convex optimization problem. □

(b) **(5 points)**Suppose $\mathcal{I}$ is a nonempty index set[may be infinite]. Then the function $F : \mathbb{R} \to \mathbb{R}$ with $F(x) = \max_{i\in\mathcal{I}} f_i(x)$ is convex if $\max_{i\in\mathcal{I}} f_i(x)$ exists for each x. Now, if $f : \mathbb{R} \to \mathbb{R}_+(f \geq 0)$ is convex and $g : \mathbb{R} \to \mathbb{R}_{++}(g > 0)$ is concave. Show that $h : \mathbb{R} \to \mathbb{R}$ with $h(x) = \frac{f(x)^2}{g(x)}$ is convex.

(b) Define that

$$h(x) = \max_{t\geq 0} \tilde{h}_t(x),$$
$$\tilde{h}_t(x) = -g(x)t^2 + 2f(x)t. \qquad \text{convex (x)}$$

Since

$$\frac{\mathrm{d}\tilde{h}_t(x)}{\mathrm{d}t} = -2g(x)t + 2f(x) = 0 \Rightarrow t = \frac{f(x)}{g(x)},$$
$$\frac{\mathrm{d}^2\tilde{h}_t(x)}{\mathrm{d}t^2} = -2g(x) < 0, \quad \text{verify}$$

we can see that

$$h(x) = \max_{t\geq 0} \tilde{h}_t(x) = \tilde{h}_t(x)|_{t=f(x)/g(x)} = \frac{f(x)^2}{g(x)}.$$

Since $f(x)$ and $-g(x)$ are both convex, and $t \geq 0$, the nonnegative weighted sum $\tilde{h}_t(x) = -g(x)t^2+2f(x)t$ is also convex. Thus, we can apply the result in sub-problem (a) and show that $h(x)$ is also convex.

□

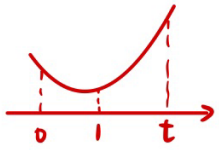
Exercise1-Convex Function

Suppose $f: \mathbb{R}^n \to \mathbb{R}$ is convex with $\operatorname{dom} f = \mathbb{R}^n$, and bounded above on $\mathbb{R}^n$. Show that f is constant.

Consider a function

$g(t) = f(x + t(y-x))$ is convex.

Assume $f(x) < f(y) \Leftrightarrow g(0) < g(1)$.



For all $t > 1$, $g(1) < \frac{t-1}{t} g(0) + \frac{1}{t} g(t)$

$\Rightarrow g(t) > t g(1) - (t-1) g(0)$

$= g(1) + (t-1) g(0)$.

Exercise2-Nonnegative weighted sums and integrals.

$f(x) = \sum_{i=1}^{r} \alpha_i x_{[i]} = \sum_{i=1}^{r} (\alpha_i - \alpha_{i+1} + \alpha_{i+1} - \alpha_{i+2} + \cdots + \alpha_{r-1} - \alpha_r + \alpha_r) x_{[i]}$

$= (\alpha_1 - \alpha_2) x_{[1]} + (\alpha_2 - \alpha_3)(x_{[1]} + x_{[2]}) + \cdots + (\alpha_{r-1} - \alpha_r)(x_{[1]} + x_{[2]} + \cdots + x_{[r]}) + \alpha_r x_{[r]}$ is convex.

(a). Show that $f(x) = \sum_{i=1}^{r} \alpha_i x_{[i]}$ is a convex function of x, where $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_r \geq 0$, and $x_{[i]}$ denotes the i th largest component of x. (You can use the fact that $f(x) = \sum_{i=1}^{k} x_{[i]}$ is convex on $\mathbb{R}^n$.)

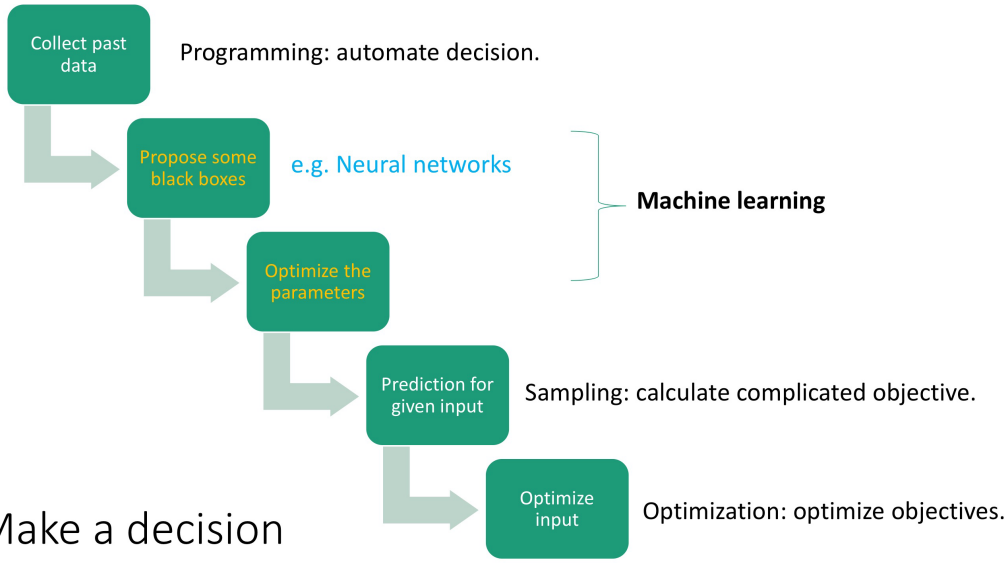
(b). Prove the following statement: if $g(x)$ is convex an $f(x)$ is non-decreasing and convex, then $h(x) = f(g(x))$ is convex.

$0 < \theta < 1$, $x < y$ $\quad h(\theta x + (1-\theta) y) \geq \theta h(x) + (1-\theta) h(y)$

$f(g(\theta x + (1-\theta) y)) \geq \theta f(g(x)) + (1-\theta) f(g(y))$

$RHS \leq f(\theta g(x) + (1-\theta) g(y)) \leq f(g(\theta x + (1-\theta) y)) = LHS$.

机器学习 machine learning



The quality of a machine learning model is dependent on two major aspects.

1. the quality of the input data.
2. the model choice itself.

How does machine learning work ?

Step 1. Choose and prepare a training data set.

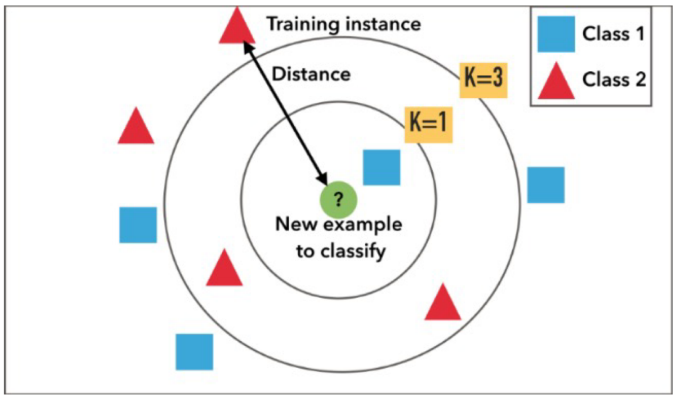
Step 2. select an algorithm to apply to the training data set.

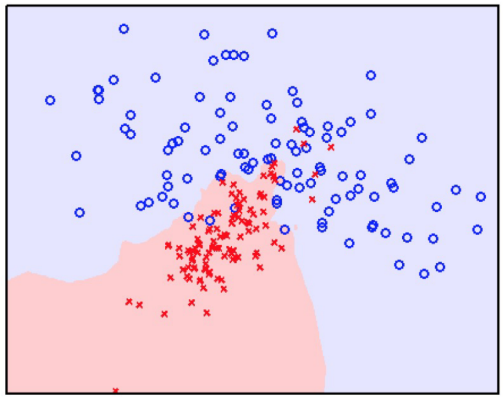
Step 3. train the algorithm to build the model.

Step 4. use and improve the model. (optimize the parameters)

监督学习 supervised learning

Model 1: k-Nearest neighbor classfier (kNN)





K = 5

- Choice of K is very critical – A small value of K means that noise will have a higher influence on the result.

噪点

- A large value of K make it kind of defeats the basic philosophy behind KNN.

无意义的分类

  - If K>total number of data, the label of any new input will be the labels that appears the most in the samples.

Model 2: Logistic regression model

$$P(y=1|x,\theta,b) = \frac{1}{1+\exp(-(\theta^T x+b))}$$

input data　parameters

x, θ 可以是向量形式.

$$P(y=0|x,\theta,b) = \frac{\exp(-(\theta^T x+b))}{1+\exp(-(\theta^T x+b))}$$

MLE: $\max_\theta l(\theta,b) = \log \prod_{i=1}^m P(y^i|x^i,\theta,b) = \sum_{i=1}^m \log P(y^i|x^i,\theta,b)$

分析: $\log P(y^i|x^i,\theta,b) = \underbrace{(y^i-1)(\theta^T x^i+b)}_{\text{convex \& concave}} - \underbrace{\log(1+\exp(-(\theta^T x^i+b))}_{\text{concave}}$

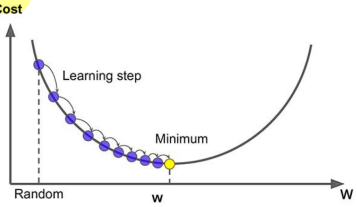
$\max_\theta l(\theta,b)$ 是凸函数，有唯一最优解.

$$\begin{cases} \frac{\partial l(\theta,b)}{\partial \theta} = \sum_{i=1}^m (y^i-1)x^i + \frac{\exp(-(\theta^T x^i+b))}{1+\exp(-(\theta^T x^i+b))} x^i = 0 \\ \frac{\partial l(\theta,b)}{\partial b} = \sum_{i=1}^m (y^i-1) + \frac{\exp(-(\theta^T x^i+b))}{1+\exp(-(\theta^T x^i+b))} = 0 \end{cases}$$

- Start with an initial point $x^{(0)}$
- Update our point by the following rule:

$$x^{(t+1)} = x^{(t)} - \alpha^{(t)} f'(x^{(t)})$$

先大后小



- Stopping criteria:
  - $|x^{(t+1)} - x^{(t)}| \le \varepsilon$
  - or $|f'(x^{(t)})| \le \varepsilon$

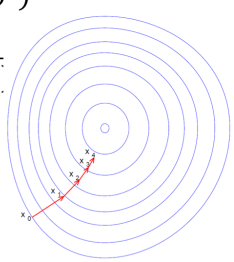
梯度下降法

gradient descend

- Initialize parameter (θ^0, b^0)
- t: the current iteration
- Do

$$\theta^{t+1} \leftarrow \theta^t + \alpha^{(t)} \frac{1}{m} \sum_i (y^i - 1)\, x^i + \frac{\exp(-\theta^t x^i - b^t) x^i}{1+\exp(-\theta^t x^i - b^t)}$$

$$b^{t+1} \leftarrow b^t + \alpha^{(t)} \frac{1}{m} \sum_i (y^i - 1) + \frac{\exp(-\theta^t x^i - b^t)}{1+\exp(-\theta^t x^i - b^t)}$$



- While $|\theta^{t+1} - \theta^t| > \epsilon$ or $|b^{t+1} - b^t| > \epsilon$

$\alpha^{(t)}$: the step size or learning rate

Unsupervised learning 无监督学习

K-Means Algorithm

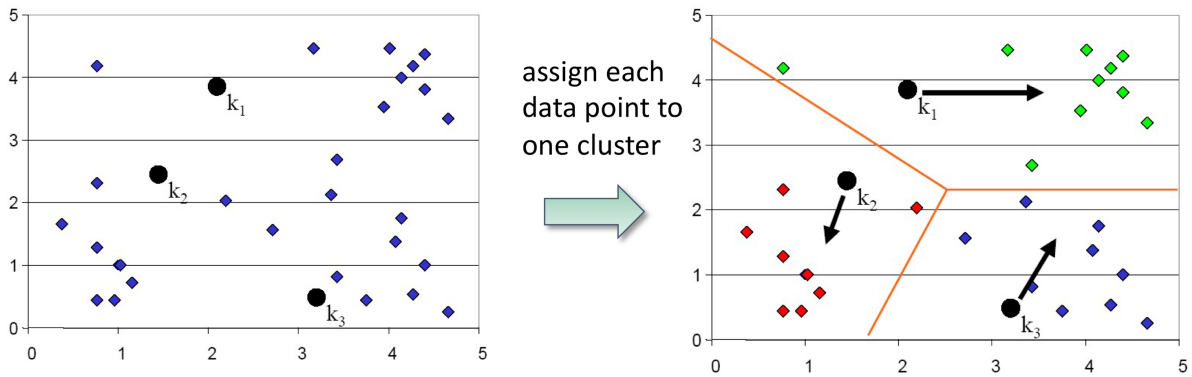
- Given m data points, $\{x^1, x^2, \cdots, x^m\}$
- Initialize k cluster centers, $\{c^1, c^2, \cdots, c^k\}$, randomly
- Do
  1. Cluster Assignment:

$$\pi(i) = \arg\min_{j=1,\cdots,k} \|x^i - c^j\|^2$$

  2. Center Adjustment:

$$c^j = \frac{1}{|\{i : \pi(i) = j\}|} \sum_{i:\pi(i)=j} x^i$$

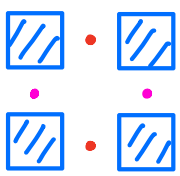
- While $|c^j - c^j_{prev}| < \epsilon$ for $j = 1, 2, \cdots, k$



- Desired properties of dissimilarity function
  - Symmetry: $d(x, y) = d(y, x)$
    - *Otherwise you could claim "Alex looks like Bob, but Bob looks nothing like Alex"*
  - Positive separability: $d(x, y) = 0$, if and only if $x = y$
    - *Otherwise there are objects that are different, but you cannot tell apart*
  - Triangular inequality: $d(x, y) \leq d(x, z) + d(z, y)$
    - *Otherwise you could claim "Alex is very like Bob, and Alex is very like Carl, but Bob is very unlike Carl"*

Remark.

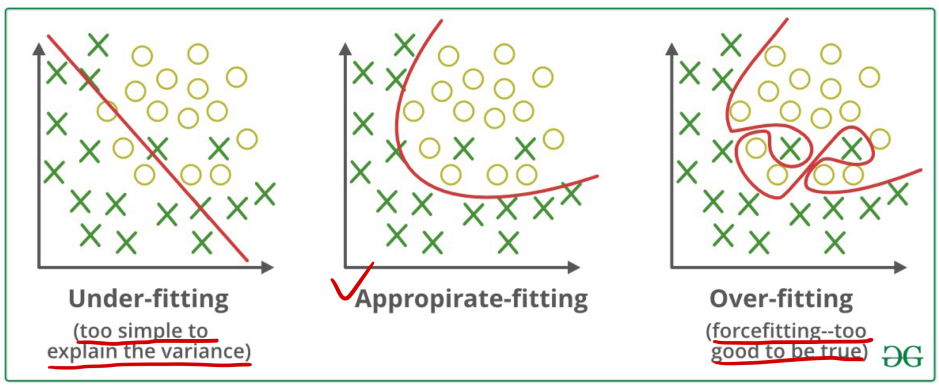
- Different initialization may lead to different result.
- The iterations will end if the tie-breaking rule is deterministic.



过拟合 / 欠拟合　overfit / underfit

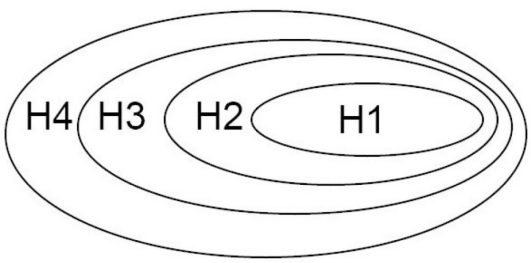
- A machine learning model is said to **overfit** the data when it learns patterns that exist only in the training set and only makes accurate prediction in the training set.

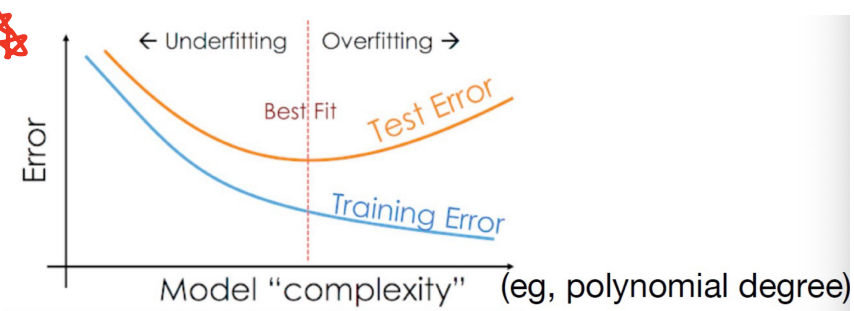
- A machine learning model is said to **underfit** if it cannot find the major patterns or relationship between variables in both training and testing data sets.



Model space

- Which model space should we choose?
- The more complex the model, the large the model space
- Eg. Polynomial function of degree 1, 2, … corresponds to space H1, H2 …





k-fold validation

$\frac{1}{k}\sum_i Err_h(\theta_i) < \frac{1}{k}\sum_i Err_g(\gamma_i) \iff$ h is better than g.

训练出的参数.

- **Parametric learning**
  - Summarizes data with a set of parameters of fixed size (independent of the number of training examples).
  - Examples: **logistic regression model**, naive Bayes, simple neural networks.
  - Benefits: simple, fast, less data required.
  - Limitations: highly constrained to a specified form, suited to simpler problems.
- **Non-parametric learning**
  - Does not make strong assumptions about the form of the mapping function. By not making assumptions, it is free to learn any functional form from the training data.
  - Examples: **K-nearest neighbor classifier**, support vector machines.
  - Benefits: flexibility, higher prediction performance.
  - Limitations: slow, more data required, overfitting.

Exercise 2

For the more general M-class classification task, the logistic regression is defined for $m = 1,\ldots,M$, as

$$P(y=m|\boldsymbol{x},\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_M,b_1,\ldots,b_M) = \frac{\exp\{\boldsymbol{\theta}_m^T\boldsymbol{x}+b_m\}}{\sum_{j=1}^{M}\exp\{\boldsymbol{\theta}_j^T\boldsymbol{x}+b_j\}}.$$

The label for a new point $\boldsymbol{x}\in\mathbb{R}^n$ is the label that maximizes the probability $P(y=m|\boldsymbol{x},\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_M,b_1,\ldots,b_M)$.

1. Prove that the region for each label is a convex set.
2. Given N samples $(\boldsymbol{x}_i,y_i)$, $i=1,\ldots,N$, prove that the log-likelihood is concave in $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_M,b_1,\ldots,b_M$.
   You can use the following two results:
   Result 1: Suppose $f:\ \mathbb{R}^n\to\mathbb{R}$, $\boldsymbol{a_i}\in\mathbb{R}^m$, $i=1,\ldots,n$ and $\boldsymbol{b}\in\mathbb{R}^n$. Define $g:\ \mathbb{R}^m\to\mathbb{R}$ by
   $$g(x) = f\left(\begin{bmatrix}\boldsymbol{a}_1^T\boldsymbol{x}+b_1\\ \vdots\\ \boldsymbol{a}_n^T\boldsymbol{x}+b_n\end{bmatrix}\right),$$
   with $\boldsymbol{dom}\ g = \{\boldsymbol{x}\,|\,\begin{bmatrix}\boldsymbol{a}_1^T\boldsymbol{x}+b_1\\ \vdots\\ \boldsymbol{a}_n^T\boldsymbol{x}+b_n\end{bmatrix}\in\boldsymbol{dom}\ f\}$. Then f is convex, so is g; if f is concave, so is g.
   Result 2: The function $f(\boldsymbol{x}) = \log(\exp\{x_1\}+\exp\{x_2\}+\cdots+\exp\{x_n\})$ is convex on $\mathbb{R}^n$.
3. Give the updating rule.

Exercise 3

We apply the gradient descent method to minimize the function $f(x) = (x-1)^2$.

1. If the update is stuck in a loop, that is, $f(x^{(t+1)}) = f(x^{(t)})$ yet $|x^{(t+1)}-x^{(t)}| > \varepsilon$, what can we say about the learning rate/step size $\alpha^{(t)}$?
2. If the function $f(x)$ keeps descending as the update goes on, what can we say about the learning rate/step size $\alpha^{(t)}$?

Solution to Exercise 2

1 The region for label m is the following set:

$$\begin{aligned} S_m &= \{\boldsymbol{x} \in \mathbb{R}^n \,|\, \frac{\exp\{\boldsymbol{\theta}_m^T \boldsymbol{x} + b_m\}}{\sum_{j=1}^M \exp\{\boldsymbol{\theta}_j^T \boldsymbol{x} + b_j\}} \geq \frac{\exp\{\boldsymbol{\theta}_k^T \boldsymbol{x} + b_k\}}{\sum_{j=1}^M \exp\{\boldsymbol{\theta}_j^T \boldsymbol{x} + b_j\}}, \text{for } m \neq k\} \\ &= \{x \in \mathbb{R}^n \,|\, \exp\{\boldsymbol{\theta}_m^T \boldsymbol{x} + b_m\} \geq \exp\{\boldsymbol{\theta}_k^T \boldsymbol{x} + b_k\}, \text{for } m \neq k\} \\ &= \{x \in \mathbb{R}^n \,|\, \boldsymbol{\theta}_{\boldsymbol{m}}^T \boldsymbol{x} + b_m \geq \boldsymbol{\theta}_k^T \boldsymbol{x} + b_k, \text{for } m \neq k\} \end{aligned}$$

which is convex.

2 The log-likelihood is

$$\begin{aligned} & l(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M, b_1, \ldots, b_M) \\ =& \log \prod_{i=1}^N \frac{\exp\{\boldsymbol{\theta}_{y_i}^T \boldsymbol{x}_i + b_{y_i}\}}{\sum_{j=1}^M \exp\{\boldsymbol{\theta}_j^T \boldsymbol{x}_i + b_j\}} = \sum_{i=1}^N (\boldsymbol{\theta}_{y_i}^T \boldsymbol{x}_i + b_{y_i}) - \sum_{i=1}^N \log \left(\sum_{j=1}^M \exp\{\boldsymbol{\theta}_j^T \boldsymbol{x}_i + b_j\} \right) \end{aligned}$$

where the first term is concave.
Now we only need to prove that $\log \left(\sum_{j=1}^M \exp\{\boldsymbol{\theta}_j^T \boldsymbol{x}_i + b_j\} \right)$ is convex in $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M, b_1, \ldots, b_M$.
According to result 1 and result 2, it is convex in $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M, b_1, \ldots, b_M$.

3 We have

$$\nabla_{\boldsymbol{\theta}_m} l(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M, b_1, \ldots, b_M) = \sum_{i=1}^N \left(\boldsymbol{x}_i \mathbf{1}_{y_i = m} - \frac{\exp\{\boldsymbol{\theta}_m^T \boldsymbol{x}_m + b_m\} \boldsymbol{x}_m}{\sum_{j=1}^M \exp\{\boldsymbol{\theta}_j^T \boldsymbol{x}_i + b_j\}} \right),$$

$$\frac{\partial l(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M, b_1, \ldots, b_M)}{\partial b_m} = \sum_{i=1}^N \mathbf{1}_{y_i = m} - \sum_{i=1}^N \frac{\exp\{\boldsymbol{\theta}_m^T \boldsymbol{x}_m + b_m\}}{\sum_{j=1}^M \exp\{\boldsymbol{\theta}_j^T \boldsymbol{x}_i + b_j\}},$$

where $\mathbf{1}_{y_i = m}$ is 1 if $y_i = m$ and 0 otherwise. The updating scheme is then:
(1) initialize $\boldsymbol{\theta}_m^{(0)}$, $b_m^{(0)}$ for $m = 1, \ldots, M$.
(2) for $m = 1, \ldots, M$, update by:
$\boldsymbol{\theta}_m^{(t+1)} = \boldsymbol{\theta}_m^{(t)} - \alpha_{\boldsymbol{\theta}_m}^{(t)} \nabla_{\boldsymbol{\theta}_m} l(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M, b_1, \ldots, b_M)|_{\boldsymbol{\theta}_m^{(t)}}$,
$b_m^{(t+1)} = b_m^{(t)} - \alpha_{b_m}^{(t)} \partial l(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M, b_1, \ldots, b_M) / \partial b_m|_{b_m^{(t)}}$
(3) Stop when $\|\boldsymbol{\theta}_m^{(t)} - \boldsymbol{\theta}_m^{(t+1)}\|_2 < \varepsilon$ and $|b_m^{(t)} - b_m^{(t+1)}| < \varepsilon$ for $m = 1, \ldots, M$.
Here $\alpha_{\boldsymbol{\theta}_m}^{(t)}$ and $\alpha_{b_m}^{(t)}$ are the learning rates/step sizes for $\boldsymbol{\theta}_m$ and b_m.

KNN Model Representation

The model representation for KNN is the entire training dataset.

It is as simple as that.

KNN has no model other than storing the entire dataset, so there is no learning required.

- **Instance-Based Learning**: The raw training instances are used to make predictions. As such KNN is often referred to as instance-based learning or a case-based learning (where each training instance is a case from the problem domain).
- **Lazy Learning**: No learning of the model is required and all of the work happens at the time a prediction is requested. As such, KNN is often referred to as a lazy learning algorithm.
- **Non-Parametric**: KNN makes no assumptions about the functional form of the problem being solved. As such KNN is referred to as a non-parametric machine learning algorithm.

Curse of Dimensionality

KNN works well with a small number of input variables (p), but struggles when the number of inputs is very large.

Each input variable can be considered a dimension of a p-dimensional input space. For example, if you had two input variables x1 and x2, the input space would be 2-dimensional.

As the number of dimensions increases the volume of the input space increases at an exponential rate.

Best Prepare Data for KNN

- **Rescale Data**: KNN performs much better if all of the data has the same scale. Normalizing your data to the range [0, 1] is a good idea. It may also be a good idea to standardize your data if it has a Gaussian distribution.
- **Address Missing Data**: Missing data will mean that the distance between samples can not be calculated. These samples could be excluded or the missing values could be imputed.
- **Lower Dimensionality**: KNN is suited for lower dimensional data. You can try it on high dimensional data (hundreds or thousands of input variables) but be aware that it may not perform as well as other techniques. KNN can benefit from feature selection that reduces the dimensionality of the input feature space.

Summary

In this post you discovered the KNN machine learning algorithm. You learned that:

- KNN stores the entire training dataset which it uses as its representation.
- KNN does not learn any model.
- KNN makes predictions just-in-time by calculating the similarity between an input sample and each training instance.
- There are many distance measures to choose from to match the structure of your input data.
- That it is a good idea to rescale your data, such as using normalization, when using KNN.