HISTOGRAMS

Aug 28, 2024
Anuran Makur

① Approximate means:

Suppose $X_1,\dots,X_n$ are $\overbrace{\text{indep. and identically dist.}}^{\text{iid}}$ random variables.

$X_i \in \{-1,1\}$, $\quad \mathbb{P}(X_i=1)=p$

$$\mathbb{E}[X_i] = \underline{1\,\mathbb{P}(X_i=1)+(-1)\,\mathbb{P}(X_i=-1)} = 2p-1$$

$$p \approx \frac{1}{n}\sum_{i=1}^{n} \underset{\uparrow\ \text{id. func.} \begin{cases} =1, \text{ if } X_i=1 \\ =0, \text{ ow} \end{cases}}{\mathbb{1}\{X_i=1\}} \qquad\qquad 1-p \approx \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{X_i=-1\}$$

$$\begin{aligned}
\mathbb{E}[X_i] &\approx 1\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{X_i=1\}\right) + (-1)\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{X_i=-1\}\right) \\
&= \frac{1}{n}\Big[1\,(\text{no. of } X_i\text{'s}=1) + (-1)\,(\text{no. of } X_i\text{'s}=-1)\Big] \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i
\end{aligned}$$

We can generalize this!

$X_1,\dots,X_n$ iid with PDF (prob. density func) $f(x)$.

$$\mathbb{E}[X_i] = \int_{-\infty}^{\infty} x\, f(x)\,dx \approx \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\mathbb{E}[g(X_i)] = \int_{-\infty}^{\infty} g(x) f(x)\,dx \approx \frac{1}{n}\sum_{i=1}^{n} g(X_i) \qquad \text{for any func. } g:\mathbb{R}\to\mathbb{R}$$

② Decomposing ISE:

$$L(w) = \int_{-\infty}^{\infty}\left(\hat{f}_m(x) - f(x)\right)^2 dx = \underbrace{\int_{-\infty}^{\infty}\hat{f}_m(x)^2\,dx - 2\int_{-\infty}^{\infty}\hat{f}_m(x)\overset{\text{unknown}}{\overset{\swarrow}{f(x)}}\,dx}_{\hookrightarrow\ \text{approximate this!}} + \overbrace{\int_{-\infty}^{\infty} f(x)^2\,dx}^{\text{constant}}$$

③ [Rudemo 1982] Leave-one-out:

$$\int_{-\infty}^{\infty}\hat{f}_m(x) f(x)\,dx \approx \frac{1}{m}\sum_{i=1}^{m}\hat{f}_m(X_i) \qquad \text{where } X_1,\dots,X_m \text{ are iid } f(x)$$

Let $\hat{f}_{m,\neg i}$ be the histogram with samples $X_1,\dots,X_{i-1},\overset{\text{leave } X_i \text{ out}}{\swarrow} X_{i+1},\dots,X_m$ $\quad(\hat{f}_{m,\neg i}\approx\hat{f}_m)$

$$\int_{-\infty}^{\infty} \hat{f}_m(x) f(x)\,dx \approx \frac{1}{m}\sum_{i=1}^{m} \hat{f}_{m,\neg i}(X_i)$$ ← less biased!

④ Algebra...

$$\int_{-\infty}^{\infty} \hat{f}_m(x)^2\,dx - 2\int_{-\infty}^{\infty} \hat{f}_m(x) f(x)\,dx$$

$$= \sum_{k=1}^{n} w\left(\frac{\hat{p}_k}{w}\right)^2 - 2\,\frac{1}{m}\sum_{i=1}^{m} \underbrace{\hat{f}_{m,\neg i}(X_i)}$$

↳ If X_i is in bin k, $\frac{m\hat{p}_k - 1}{w(m-1)} = \hat{f}_{m,\neg i}(X_i)$

$$= \frac{1}{w}\sum_{k=1}^{n} \hat{p}_k^2 - 2\,\frac{1}{\cancel{m}}\sum_{k=1}^{n} \underbrace{\cancel{m}\hat{p}_k}_{\text{no. of } X_i\text{'s in bin } k} \cdot \frac{m\hat{p}_k - 1}{w(m-1)}$$

$$= \frac{1}{w}\sum_{k=1}^{n} \hat{p}_k^2 - \frac{2m}{w(m-1)}\sum_{k=1}^{n} \hat{p}_k^2 + \frac{2}{w(m-1)} \cancelto{=1}{\sum_{k=1}^{n} \hat{p}_k}$$

$$= \underbrace{\boxed{\frac{2}{w(m-1)} - \frac{(m+1)}{w(m-1)}\sum_{k=1}^{n} \hat{p}_k^2}}_{J(w)}$$

$$L(w) \approx J(w) + \underbrace{\int_{-\infty}^{\infty} f(x)^2\,dx}_{\text{constant}}$$

PROBABILITY

Aug 30, 2024
Anuran Makur

- **Probability Space:** $(\Omega, \mathcal{F}, \mathbb{P})$

1) Sample space Ω - set of outcomes of random experiment

   e.g: $\{H,T\}, \{1,2,3,4,5,6\}, [0,1], \mathbb{R}$

2) σ-algebra $\mathcal{F}$ - collection of subsets/events of Ω to which we will assign prob.s
   - $\phi \in \mathcal{F}$
   - $A \in \mathcal{F} \Rightarrow \overset{\text{"complement"}}{A^c} \in \mathcal{F}$
   - $A_1, A_2, A_3, \ldots \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

3) Probability measure $\mathbb{P}: \mathcal{F} \to [0,1]$ (σ-additive set func.)
   - $\mathbb{P}(\Omega) = 1$ [normalization]
   - $A_1, A_2, A_3, \ldots \in \mathcal{F}$ s.t. $A_i \cap A_j = \phi$ (disjoint) [σ-addivity]

     $$\mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

   Kolmogorov axioms

Example: (Coin toss) $\Omega = \{H, T\}$, $\mathcal{F} = \{\phi, \{H\}, \{T\}, \{H,T\}\}$

$\mathbb{P}(\phi) = 0$, $\mathbb{P}(H) = \frac{1}{2}$ $\mathbb{P}(T) = \frac{1}{2}$ $\mathbb{P}(\Omega) = 1$

- **Properties:**

1) Monotonicity: $A \subseteq B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$
2) Inclusion-Exclusion: $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
3) Union Bound (Boole): $\mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i)$

   ↑ need not be disjoint

4) Continuity: $A_1 \overset{\text{"is a subset of"}}{\subseteq} A_2 \subseteq A_3 \subseteq \cdots \Rightarrow \lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big)$

   $A_1 \supseteq A_2 \supseteq A_3 \supseteq \cdots \Rightarrow \lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\Big(\bigcap_{i=1}^{\infty} A_i\Big)$

   equivalent to σ-additivity axiom

- **Conditional Probability:** If A is an event with $\mathbb{P}(A) > 0$, then the cond. prob. of B given A is $\mathbb{P}(B|A) \triangleq \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)}$.

→ $(A, \{A \cap B : B \in \mathcal{F}\}, \mathbb{P}(\cdot|A))$ is a prob. space!

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B)\,\mathbb{P}(B)}{\mathbb{P}(A)}$$ [Bayes Rule]

Aug 30, 2024
Anuran Makur

- Independence: (occurrence of event A does not affect event B)

A and B are independent events if $\underbrace{\mathbb{P}(A|B)}_{\frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)}} = \mathbb{P}(A).$

$$\downarrow$$

$$\boxed{\mathbb{P}(A\cap B) = \mathbb{P}(A)\,\mathbb{P}(B)} \leftarrow \text{$\mathbb{P}(A)$ or $\mathbb{P}(B)$ could be 0!}$$

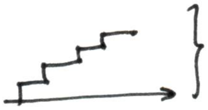
- Random Variable: A random variable is a function $X : \overset{\text{sample space}}{\Omega} \to \overset{\text{real no.s}}{\mathbb{R}}$ such that for every $c \in \mathbb{R}$, $\{\omega\in\Omega : X(\omega)\leq c\} \in \mathcal{F}$.

The "probability dist." of X is characterized by the probabilities of $\underbrace{\{\omega\in\Omega : X(\omega)\leq c\}}_{\{X\leq c\}}$. (notation)

Cumulative Distribution Function (CDF): For rand. var. X, its CDF is $F_X : \mathbb{R}\to[0,1]$, $\quad F_X(x) \triangleq \mathbb{P}(X\leq x)$. ← All rand. var.s have CDF!

3 "pure" rand. vars: A pure rand. variable X and its CDF F has one of 3 forms:

① Discrete rand. var: X takes countably many values with >0 probability and F is discontinuous with jumps, & piecewise constant.



② Continuous rand. var: X takes uncountably many values (e.g. $[0,1]$) and F absolutely continuous. ← (Think: continuously differentiable)

$\exists$ a derivative $f = F'$ s.t. $F(x) = \int_{-\infty}^{x} f(t)\,dt$.

We refer to f as the probability density function (PDF).

③ Singular rand. var: F is continuous but not abs. cont. } not for exam!
eg: Cantor function (Devil's staircase)

Sep 4, 2024
Anuran Makur

Decomposition Thm: The probability dist: of a general rand. var., $\mathbb{P}_X$, can be uniquely decomposed into a mixture of the 3 pure components:

$$\mathbb{P}_X = \lambda_1 \mathbb{P}_{\text{disc.}} + \lambda_2 \mathbb{P}_{\text{cont.}} + \lambda_3 \mathbb{P}_{\text{sing.}}$$

$$\left. \lambda_1, \lambda_2, \lambda_3 \geq 0, \quad \lambda_1 + \lambda_2 + \lambda_3 = 1 \right\} \text{convex weights}$$

- **Continuous rand. variables:**

Prop: A function $f: \mathbb{R} \to \mathbb{R}$ is a PDF of some rand. var. iff

① $f(x) \geq 0$ for all $x \in \mathbb{R}$,

② $\int_{-\infty}^{\infty} f(x)\, dx = 1$.

- **CDF:** **Prop:** A function $F: \mathbb{R} \to [0,1]$ is a CDF of some rand. var. iff

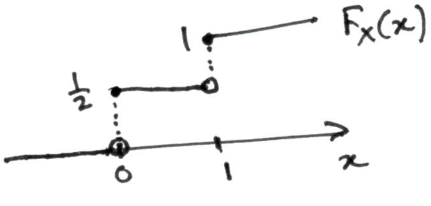
① $\lim_{x \to -\infty} F(x) = 0$,

② $\lim_{x \to \infty} F(x) = 1$,

③ If $x \leq y$, then $F(x) \leq F(y)$,

④ F is right-continuous.

Example $X \in \{0,1\}$, $X \sim \text{Ber}(\frac{1}{2})$



- **Discrete rand. var:** X takes values in a set $\mathcal{X} \subseteq \mathbb{R}$ ← countable

Prop: A function $f: \mathcal{X} \to [0,1]$ is a PMF iff

① $f(x) \geq 0$ for all $x \in \mathcal{X}$,

② $\sum_{x \in \mathcal{X}} f(x) = 1$.

Sep 4, 2024
Anuran Makur

- Expected Value:

① Discrete: $X \in \overset{\text{range}(X)}{\mathcal{X}}$ $\quad \mathbb{E}[X] = \sum_{x \in \mathcal{X}} x \overset{\text{PMF}}{f_X(x)}$, $\overset{\text{Function } g(\cdot)}{\mathbb{E}[g(X)]} = \sum_{x \in \mathcal{X}} g(x) \overset{\text{PMF}}{f_X(x)}$

② Continuous: $\quad \mathbb{E}[X] = \int_{-\infty}^{\infty} x \underset{\text{PDF}}{f_X(x)}\, dx$, $\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) \underset{\text{PDF}}{f_X(x)}\, dx$

$g(x) = x^2 \qquad \mathbb{E}[g(X)] = \mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x)\,dx$

$g(x) = x \qquad \mathbb{E}[g(X)] = \mathbb{E}[X] = \int_{-\infty}^{\infty} x\, f_X(x)\,dx$

Linearity: For two r.v.s X, Y and two constants $a, b \in \mathbb{R}$,

$$\mathbb{E}[aX + bY] = a\,\mathbb{E}[X] + b[Y].$$

Variance:

$$\begin{aligned} \operatorname{var}(X) &\triangleq \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \geq 0 \\ &= \mathbb{E}\left[X^2 - 2X\,\mathbb{E}[X] + \mathbb{E}[X]^2\right] \\ &= \mathbb{E}[X^2] - 2\,\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2 \\ &= \boxed{\mathbb{E}[X^2] - \mathbb{E}[X]^2} \geq 0 \end{aligned}$$

Sep 6, 2024
Anuran Makur

- **Probability plots:**

① **P-P plot:** ← probability

Given two CDFs F and G, their P-P plot is the graph

$$\{(F(x), G(x)) : \underset{\uparrow \text{ hard to "grid" this}}{x \in \mathbb{R}}\}.$$

② **Q-Q plot:** ← quantile

Given two CDFs F and G, and their $\overbrace{\text{generalized inverses}}^{\text{quantile functions}}$* F^{-1} and G^{-1}, their Q-Q plot is the graph $\{(F^{-1}(q), G^{-1}(q)) : \underset{\uparrow \text{ can "grid" this!}}{q \in [0,1]}\}$.

* **Generalized inverse:** Given CDF $F : \mathbb{R} \to [0,1]$, its gen. inverse is $F^{-1} : [0,1] \to \mathbb{R} \cup \{\pm\infty\}$, $F^{-1}(q) \triangleq \inf\{x \in \mathbb{R} : F(x) \geq q\}$.

↑ "infimum" [min if set is closed]

eg: $\inf(0,\infty) = 0$ but min doesn't exist

- **Example:** $X \sim \text{Bernoulli}(p)$ ← "is dist. as"

$$\mathbb{E}[X] = 0(1-p) + 1(p) = \underline{\underline{p}}$$

$$\begin{aligned}\text{var}(X) &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\ &= \left(0^2(1-p) + 1^2(p)\right) - p^2 \\ &= p - p^2 \\ &= \underline{\underline{p(1-p)}}\end{aligned}$$

Example: $X \sim \text{Exp}(\lambda)$

$$\mathbb{E}[X] = \int_0^\infty \underbrace{x\lambda e^{-\lambda x}}_{\hookrightarrow u = \lambda x} dx = \frac{1}{\lambda}\int_0^\infty u e^{-u} du$$

$$= \frac{1}{\lambda}\Big(\underbrace{\cancelto{0}{\left[-ue^{-u}\right]_0^\infty}}_{\lim_{u\to\infty} \frac{u}{e^u} = 0} + \underbrace{\cancelto{1}{\int_0^\infty e^{-u} du}}_{\text{PDF}}\Big) \quad \left[\begin{matrix}\text{integrate} \\ \text{by parts}\end{matrix}\right.$$

$$= \underline{\underline{\frac{1}{\lambda}}}$$

$$\text{var}(X) = \underline{\underline{\frac{1}{\lambda^2}}} \leftarrow \text{Exercise}$$

Example: $X \sim \text{Binomial}(n, p)$

$X_1, X_2, \ldots, X_n$ iid Ber(p) variables

$$X = X_1 + X_2 + \cdots + X_n$$

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}[X_1 + \cdots + X_n] \\ &= \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] \\ &= \underline{\underline{np}}\end{aligned}$$

Binomial coefficient

$$\binom{n}{k} = \frac{n(n-1)(n-2)\ldots(n-k+1)(n-k)\cdots 2\cdot 1}{k!\,(n-k)\cdots 2\cdot 1}$$

$$= \frac{n!}{k!\,(n-k)!}$$

$$\text{var}(X) = \underline{\underline{np(1-p)}} \leftarrow \text{Exercise}$$

ESTIMATION:

Sep 13, 2024
Anuran Makur

① Properties of mean and variance:

1) For rand var.s X, Y and constants $a, b, c \in \mathbb{R}$,

$$\mathbb{E}[aX + bY + c] = a\,\mathbb{E}[X] + b\,\mathbb{E}[Y] + c.$$

2) For rand. var.s X, Y that are <u>independent</u>, and constants $a, b, c \in \mathbb{R}$,

$$\mathrm{var}(aX + bY + c) = a^2\,\mathrm{var}(X) + b^2\,\mathrm{var}(Y).$$

② Unbiased estimators of mean and variance:

Suppose $X_1, \ldots, X_n$ are iid rand. var.s with $\mathbb{E}[X] = \mu$ and variance $\mathrm{var}(X) = \sigma^2$.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S^2 = \overset{\text{Bessel's correction}}{\frac{1}{n-1}} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

$\bar{X}$: sample mean $\qquad$ S^2: sample variance

$$\mathbb{E}[\bar{X}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\mathbb{E}[X_i]}_{\mu} = \mu \qquad [\underline{\text{unbiased}}]$$

$\bar{X}$ is an <u>unbiased</u> estimator of μ.

$$\mathbb{E}\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \mathbb{E}\left[\sum_{i=1}^{n} X_i^2 - 2X_i\bar{X} + \bar{X}^2\right] = \mathbb{E}\left[\sum_{i=1}^{n} X_i^2 - n2\bar{X}\underbrace{\sum_{i=1}^{n} X_i}_{} + \sum_{i=1}^{n}\bar{X}^2\right]$$

$$= \mathbb{E}\left[\sum_{i=1}^{n} X_i^2 - 2n\bar{X}^2 + n\bar{X}^2\right]$$

$$= \sum_{i=1}^{n}\mathbb{E}[X_i^2] - n\,\mathbb{E}[\bar{X}^2]$$

$$= n\,\mathbb{E}[X^2] - n\,\mathbb{E}[\bar{X}^2]$$

$$= n(\sigma^2 + \mu^2) - n\Big(\underbrace{\mathrm{var}(\bar{X})}_{\sigma^2/n} + \underbrace{\mathbb{E}[\bar{X}]^2}_{\mu^2}\Big)$$

$$= n\sigma^2 + \cancel{n\mu^2} - \cancel{n}\left(\frac{\sigma^2}{\cancel{n}}\right) - \cancel{n\mu^2}$$

$$= (n-1)\sigma^2$$

$$\mathrm{var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

$$\mathrm{var}(\bar{X}) = \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}(X_i)$$

$$= \frac{1}{n^2}\, n\sigma^2$$

$$= \frac{\sigma^2}{n}$$

$$\mathbb{E}\Bigg[\underbrace{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}_{S^2}\Bigg] = \sigma^2$$

③ Tail bounds:

- **Markov's inequality:** For any rand. var. $X \geq 0$ and any constant $a > 0$,

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Pf: (continuous)

$$\mathbb{E}[X] = \int_0^\infty x \underset{\substack{\uparrow \\ \text{PDF}}}{f_X(x)}\, dx = \underbrace{\int_0^a x f_X(x) dx}_{\geq 0} + \int_a^\infty x f_X(x)\, dx \geq \int_a^\infty \underset{\substack{\uparrow \\ x \geq a}}{x} f_X(x)\, dx$$

$$\geq \int_a^\infty a f_X(x)\, dx = a \int_a^\infty f_X(x) dx = a\, \mathbb{P}(X \geq a).$$ ▨

- **Chebyshev's inequality:** For any rand. var. X and any constant $a > 0$,

$$\mathbb{P}\big(|X - \mathbb{E}[X]| \geq a\big) \leq \frac{\mathrm{var}(X)}{a^2}.$$

Pf: Let $Y = (X - \mathbb{E}[X])^2 \geq 0$. Then, by Markov,

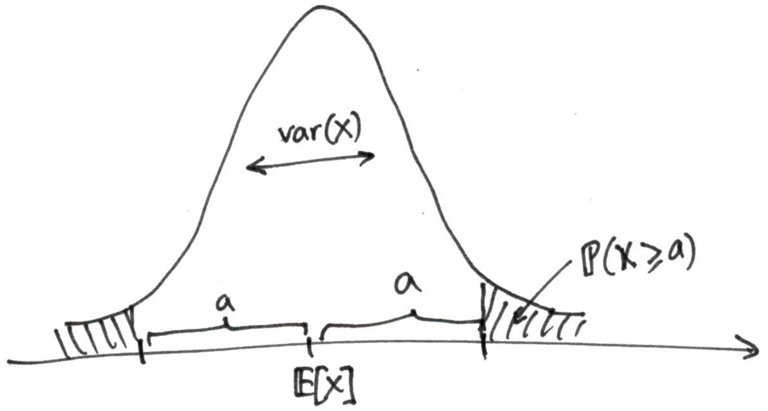
$$\mathbb{P}(Y \geq a^2) \leq \frac{\mathbb{E}[Y]}{a^2} \leftarrow \mathbb{E}[Y] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathrm{var}(X)$$

$$= \frac{\mathrm{var}(X)}{a^2}$$

$$\{Y \geq a^2\} = \{(X - \mathbb{E}[X])^2 \geq a^2\} \overset{\overset{\text{Argue this!}}{*}}{=} \{|X - \mathbb{E}[X]| \geq a\}$$

$$\Rightarrow \mathbb{P}\big(|X - \mathbb{E}[X]| \geq a\big) = \mathbb{P}(Y \geq a^2) \leq \frac{\mathrm{var}(X)}{a^2}.$$ ▨

Picture:



Sep 13, 2024
Anuran Makur

④ ∘ **Law of Large Numbers:** (weak)

For iid random variables $X_1, \ldots, X_n$ with mean $\mathbb{E}[X]=\mu$ and variance $\mathrm{var}(X)=\sigma$;

we have:
$$\forall \varepsilon > 0, \quad \lim_{n\to\infty} \mathbb{P}\Big(\Big|\underbrace{\frac{1}{n}\sum_{i=1}^{n} X_i}_{\bar{X}} - \underset{\substack{\uparrow \\ \mathbb{E}[\bar{X}]=\mathbb{E}[X_i]}}{\mu}\Big| \geq \varepsilon\Big) = 0.$$

Pf: By Chebyshev, for any $\varepsilon > 0$,

$$0 \leq \mathbb{P}\Big(\Big|\underbrace{\frac{1}{n}\sum_{i=1}^{n} X_i}_{\bar{X}} - \underbrace{\mu}_{\mathbb{E}[\bar{X}]}\Big| \geq \varepsilon\Big) \leq \frac{\mathrm{var}(\bar{X})}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}$$

$$\lim_{n\to\infty} \mathbb{P}(|\bar{X} - \mathbb{E}[\bar{X}]| \geq \varepsilon) = 0. \qquad \blacksquare$$

⑤ • **Central Limit Theorem:** (Lindeberg-Lévy)

For iid random variables $X_1, \ldots, X_n$ with $\mathbb{E}[X]=\mu$ and $\mathrm{var}(X)=\sigma^2$,

$$\forall x \in \mathbb{R}, \quad \lim_{n\to\infty} \mathbb{P}\Big(\underbrace{\frac{1}{\sqrt{n}\,\sigma}\sum_{i=1}^{n}(X_i-\mu)}_{\text{mean 0 and variance 1}} \leq x\Big) = \underbrace{\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt}_{\substack{\text{standard Normal CDF} \\ \text{with mean 0 and variance 1.}}}.$$

Slides ~~sta~~statement:

$$\frac{\sigma}{\sqrt{n}}\Big(\frac{1}{\sqrt{n}\sigma}\sum_{i=1}^{n}(X_i-\mu)\Big) \overset{n\to\infty}{\approx} N\Big(0, \frac{\sigma^2}{n}\Big)$$

$$\to \quad \mu + \Big(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)\Big) \approx N\Big(\mu, \frac{\sigma^2}{n}\Big)$$

$$\to \quad \boxed{\frac{1}{n}\sum_{i=1}^{n} X_i \approx N\Big(\mu, \frac{\sigma^2}{n}\Big)} \quad \text{(in slides)}$$

Sep 16, 2024
Anuran Makur

Induced Distributions:

① Change-of-Variables: (linear case)

Let X be a continuous random var. with PDF f_X and CDF F_X. For any $a \neq 0$ and $b \in \mathbb{R}$, let $Y = \underline{aX + b}$ be another continuous rand. var. with PDF f_Y and CDF F_Y.

Can we write f_Y in terms of f_X?

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(aX + b \leq y) = \begin{cases} \mathbb{P}\left(X \leq \frac{y-b}{a}\right), & a > 0 \\ \mathbb{P}\left(X \geq \frac{y-b}{a}\right), & a < 0 \end{cases}$$

$$= \begin{cases} F_X\left(\frac{y-b}{a}\right), & a > 0 \\ 1 - F_X\left(\frac{y-b}{a}\right), & a < 0 \end{cases}$$

$$\underset{\substack{\uparrow \\ \text{PDF}}}{f_Y(y)} = \frac{d}{dy} F_Y(y) \underset{\substack{\uparrow \\ \text{[Chain rule of calculus]}}}{=} \begin{cases} f_X\left(\frac{y-b}{a}\right) \frac{1}{a}, & a > 0 \\ -f_X\left(\frac{y-b}{a}\right) \frac{1}{a}, & a < 0 \end{cases} = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$$

$$\boxed{f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)} \quad \text{for all } y \in \mathbb{R}$$

Sep 16, 2024
Anuran Makur

② Convolution: (discrete)

Let X and Y be discrete rand. var.s with values in $\mathbb{Z}$ and PMFs f_X and f_Y, respectively. Assume X and Y are independent. Let $Z = X + Y$ be a disc. rand. var. with PMF f_Z. Can we write f_Z in terms of f_X and f_Y?

$$
\begin{aligned}
f_Z(k) &= \mathbb{P}(Z=k) = \mathbb{P}(\exists j \in \mathbb{Z}, X=j \text{ and } Y=k-j) \\
&= \sum_{j=-\infty}^{\infty} \mathbb{P}(X=j \text{ and } Y=k-j) = \sum_{j=-\infty}^{\infty} \mathbb{P}(X=j)\,\mathbb{P}(Y=k-j) \\
&= \sum_{j=-\infty}^{\infty} f_X(j) f_Y(k-j)
\end{aligned}
$$

Define convolution: $$(f_X \overset{\text{"convolve"}}{\star} f_Y)(k) \triangleq \sum_{j=-\infty}^{\infty} f_X(j) f_Y(k-j)$$

$$\boxed{f_Z = f_X \star f_Y}$$

Note: If X and Y were cont. r.v.s with PDFs f_X, f_Y.
Let $Z = X+Y$ have PDF f_Z.
$$f_Z(t) = \int_{-\infty}^{\infty} f_X(u) f_Y(t-u)\,du$$

③ **Fact:** If X and Y are independent Gaussians with means μ_X and μ_Y, and variances σ_X^2 and σ_Y^2, respectively, then for any $a, b, c \in \mathbb{R}$, $aX + bY + c$ is Gaussian with mean $a\mu_X + b\mu_Y + c$ and variance $a^2\sigma_X^2 + b^2\sigma_Y^2$.

↙ can be generalized.

Two-Sample Test & Confidence Interval:

Sep 18, 2026
Anuran Makur

- $\underset{\uparrow \text{ Sample 1}}{\bar{x}_0} \sim N\left(\mu_0, \frac{\sigma_0^2}{n_0}\right)$ [CLT] , $\underset{\uparrow \text{ Sample 2}}{\bar{x}_1} \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right)$ [CLT] ($\bar{x}_0$ and $\bar{x}_1$ are independent)

$$\underset{\uparrow \text{ Statistic}}{\bar{x}} = \bar{x}_0 - \bar{x}_1 \sim \underline{\underline{N\Big(\overbrace{\mu_0 - \mu_1}^{=0}, \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}\Big)}}$$

$$\Rightarrow \boxed{\bar{x} \sim N\left(0, \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}\right)} \leftarrow$$ This is used in two-sample z-test.

(Recall: $\text{var}(aX + bY) = a^2 \text{var}(X) + b^2 \text{var}(Y)$)

- A confidence level $\overset{\in[0,1]}{c}$ refers to the following prob:

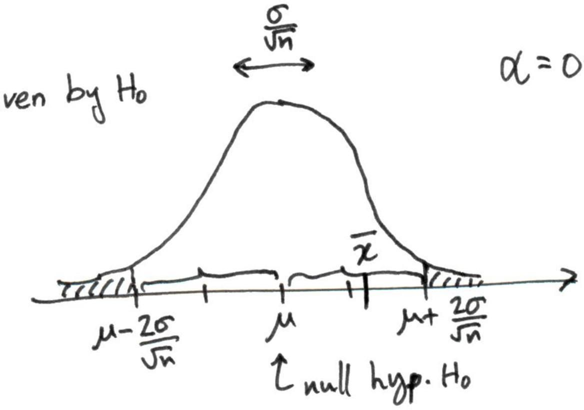
$$c = \mathbb{P}\Big(\mu \in \overbrace{(\bar{x} - \underset{\uparrow}{u}, \bar{x} + u)}^{\text{confidence interval}}\Big) = \mathbb{P}\big(|\bar{x} - \mu| < u\big)$$

what is this? $u = z_c \frac{\sigma}{\sqrt{n}}$ (slides)

Relation to z-test

Z-test: μ given by H_0

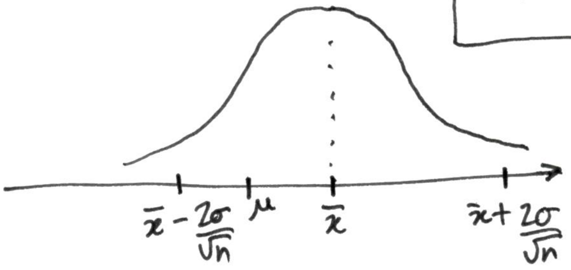
$\alpha = 0.05$



Does $\bar{x}$ live in $\left(\mu - \frac{2\sigma}{\sqrt{n}}, \mu + \frac{2\sigma}{\sqrt{n}}\right)$?

Note:
$\bar{x} \in \left(\mu - \frac{2\sigma}{\sqrt{n}}, \mu + \frac{2\sigma}{\sqrt{n}}\right)$
$\iff \mu \in \left(\bar{x} - \frac{2\sigma}{\sqrt{n}}, \bar{x} + \frac{2\sigma}{\sqrt{n}}\right)$

Confidence int: μ unknown



T-DISTRIBUTION:

Sep 20, 2024
Anuran Makur

- **Cauchy Distibution:** (Ex. t-dist with $\nu = 1$)

$X \sim$ Cauchy $\qquad \underset{\uparrow \text{PDF}}{f_X(x)} = \dfrac{1}{\underline{\underline{\pi}}(1+x^2)} \longrightarrow \mathbb{E}[X]$, $\text{var}(X)$ <u>undefined</u> !

$$\left\{ \begin{aligned} \left. \begin{aligned} &\int_{-\infty}^{\infty} \frac{1}{1+x^2}\,dx = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cancel{\frac{1}{1+\tan^2\theta}}\,\cancel{(1+\tan^2\theta)}\,d\theta = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} 1 \,d\theta = \underline{\underline{\pi}} \\ &\uparrow\; x = \tan\theta \\ &\qquad = \frac{\sin\theta}{\cos\theta} \\ &\frac{dx}{d\theta} = \frac{\cos\theta}{\cos\theta} + \frac{\sin\theta\,(\cancel{-}1)(\cancel{-}\sin\theta)}{\cos^2\theta} = 1 + \tan^2\theta \end{aligned} \right\} \text{normalization} \end{aligned} \right.$$

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\,dx = \int_{-\infty}^{\infty} \frac{x}{\pi(1+x^2)}\,dx = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{2x}{1+x^2}\,dx = \frac{1}{2\pi}\Big[\log(1+x^2)\Big]_{-\infty}^{\infty} = \underline{\underline{\infty - \infty}}$$

- **t-confidence interval:**

$$\left(\underset{\text{sample mean}}{\bar{x}} - \underset{t_c\text{-value}}{t_c} \frac{\overset{\text{sample std. dev}}{s}}{\sqrt{n}},\ \bar{x} + t_c \frac{s}{\underset{\text{no. of samples}}{\sqrt{n}}}\right)$$

Sep 23, 2024
Anuran Makur

Binomial Coefficients:

- Recall $X \sim \text{Binomial}(\underset{\substack{\uparrow \\ \text{no. of trials} \\ \text{(pos. integer)}}}{n}, \underset{\substack{\uparrow \\ \text{success prob.} \in [0,1]}}{p})$

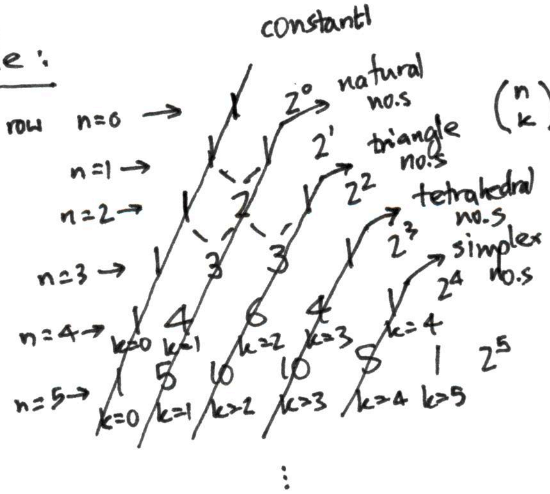
PMF: $\mathbb{P}(X = k) = \underset{\substack{\uparrow \\ \text{binomial coefficient}}}{\binom{n}{k}} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, \ldots, n-1, n$

$$\sum_{k=0}^{n} \mathbb{P}(X = k) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \underset{\substack{\uparrow \\ \text{binomial thm}}}{=} (p + 1 - p)^n = 1$$

- **Recall:** $\binom{n}{k} = \dfrac{n!}{\underset{\substack{\uparrow \\ \text{Factorial: } k! = k(k-1)(k-2)\cdots 3\cdot 2\cdot 1, \ 0! = 1}}{k!(n-k)!}} \leftarrow$ no. of ways to choose k elements from a set of n elements

[from hereon, not for exam]

- **Pascal's Triangle:**



$\binom{n}{k}$ = entry in row n and index k of Pascal's triangle

Pascal's identity:

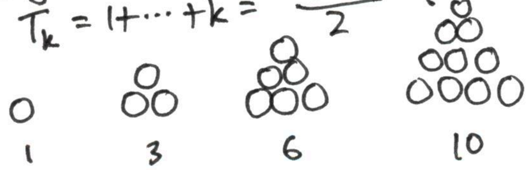
For $1 \leq k \leq n-1$, $n \geq 2$,

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}.$$

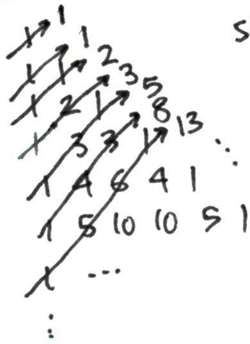
Powers of 2: $\sum_{k=0}^{n} \binom{n}{k} \underset{\substack{\uparrow \\ \text{binomial thm}}}{=} (1+1)^n = 2^n$

Triangle no.s:

$T_k = 1 + \cdots + k = \dfrac{k(k+1)}{2} = \binom{k+1}{2}$



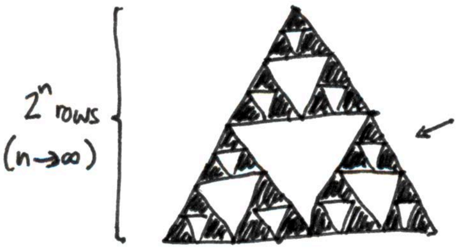
Shallow Diagonals:



Shallow diagonals sum to Fibonacci no.s.

Sierpiński Triangle:

- Color all odd numbers black and even numbers white.



Colored Pascal's triangle!

- Famous example of <u>fractal</u> !

LINEAR REGRESSION:

Oct 14, 2024
Anuran Makur

① Gradient:

Given a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, its gradient is the vector field $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$ given by:

$$\forall x \in \mathbb{R}^d, \quad \nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \frac{\partial f}{\partial x_2}(x) \\ \vdots \\ \frac{\partial f}{\partial x_d}(x) \end{bmatrix}.$$

② Quadratic Form:

Given fixed $A \in \mathbb{R}^{d \times d}$, define $f(x) = \overbrace{x^T A x}^{\text{real no.}}$ for $x \in \mathbb{R}^d$.

Prop: For $x \in \mathbb{R}^d$, $\nabla f(x) = (A + A^T)x$.

Pf: Observe $f(x) = \sum_{i=1}^d x_i [Ax]_i = \sum_{i=1}^d x_i \sum_{j=1}^d A_{ij} x_j = \sum_{i=1}^d \sum_{j=1}^d x_i x_j A_{ij}$.

$$\frac{\partial f}{\partial x_k}(x) = \frac{\partial}{\partial x_k}\left(x_k^2 A_{kk} + \sum_{j \neq k} x_k x_j A_{kj} + \sum_{i \neq k} x_i x_k A_{ik}\right)$$

$$= 2x_k A_{kk} + \sum_{j \neq k} x_j A_{kj} + \sum_{i \neq k} x_i A_{ik}$$

$$= \underbrace{\sum_{i=1}^d x_i \underbrace{A_{ik}}_{[A^T]_{ki}}}_{[A^T x]_k} + \underbrace{\sum_{j=1}^d x_j A_{kj}}_{[Ax]_k}$$

$$= [A^T x]_k + [Ax]_k$$

$$= \left[(A^T + A)x\right]_k .$$

▨

③ Linear Form:

Given fixed $b \in \mathbb{R}^d$, define $f(x) = b^T x$ for $x \in \mathbb{R}^d$.

Prop: For $x \in \mathbb{R}^d$, $\nabla f(x) = b$.

▨

Pf: $\frac{\partial f}{\partial x_k}(x) = \frac{\partial}{\partial x_k} \sum_{i=1}^d b_i x_i = b_k$.

Oct 14, 2024
Anuran Makur

④ Normal Equation:

Recall: $y \in \mathbb{R}^N$ ← no. of data samples

↖ target vector

$X \in \mathbb{R}^{N \times (M+1)}$ ↖ no. of explanatory variables

↖ feature matrix

} given dataset

Mean-squared Error: $E(\beta) = \frac{1}{N}\|y - X\beta\|_2^2$ for $\beta \in \mathbb{R}^{M+1}$

↙ minimize ($E(\beta)$); ↙ regression coefficients (β); ↑ ℓ^2 norm

$$\Rightarrow E(\beta) = \frac{1}{N}(y - X\beta)^T(y - X\beta)$$

$$= \frac{1}{N}(y^T - \beta^T X^T)(y - X\beta) = \frac{1}{N}\Big[y^Ty - \underbrace{y^TX\beta - \beta^TX^Ty}_{= -(X^Ty)^T\beta} + \beta^TX^TX\beta\Big]$$

$$= \frac{1}{N}\Big[y^Ty - 2(X^Ty)^T\beta + \beta^T(X^TX)\beta\Big]$$

$$\Rightarrow \nabla E(\beta) = -\frac{2}{N}X^Ty + \frac{1}{N}(X^TX + X^TX)\beta$$

$$= \frac{2}{N}(X^TX\beta - X^Ty)$$

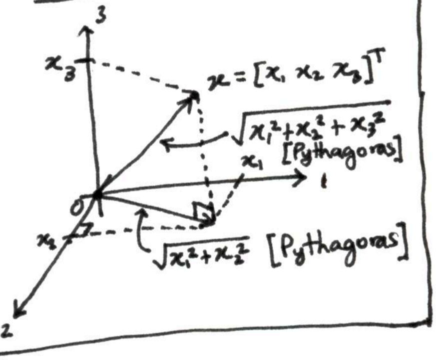
To min. $E(\beta)$, we use stationarity condition:

$$\nabla E(\beta) = 0 \leftarrow \text{zero vector}$$

$$\iff \frac{2}{N}(X^TX\beta - X^Ty) = 0$$

$$\iff \boxed{X^TX\beta = X^Ty} \Big\}\ \text{normal equations}$$

ℓ²-norm Intuition:



Note: $\nabla^2 E(\beta) = \frac{2}{N}X^TX$ is positive semidefinite
$\Rightarrow$ Stationary point is global minimum

⑤ Solutions of Normal Eq:

- Normal Equation always has solns because $\text{row}(X) = \text{col}(X^TX)$.

↙ rowspace ($\text{row}(X)$); ↙ col-space ($\text{col}(X^TX)$)

Let $y = X\gamma + v$, for $\gamma \in \mathbb{R}^{M+1}$ and $v \in \text{left-null}(X)$.

↙ left-nullspace (X); ↓ $(X^Tv = 0)$

↑ unique decomposition of $\mathbb{R}^N$ into col(X) and its ⊥-complement.

$$\Rightarrow X^Ty = X^T(X\gamma + v) = X^TX\gamma + X^Tv = X^TX\gamma$$

$\Rightarrow$ Normal eq. has solution!

Oct 16, 2024
Anuran Makur

- Normal equation $X^TX\beta = X^Ty$ has <u>unique</u> solution β^* when:

$$\text{equivalent}\begin{cases} -\ X^TX \text{ is invertible/non-singular} \\ -\ X \text{ has linearly indep. cols.} \\ -\ X \text{ full-rank} \end{cases}\Bigg\}\ \boxed{\beta^* = (X^TX)^{-1}X^Ty}$$

- What if X^TX is not invertible? Choose solution β^* with min ℓ^2-norm.
  ↳ prevents "overfitting"

⑥ <u>Moore-Penrose Pseudoinverse</u>: [not in exam]

- For any $A \in \mathbb{R}^{m\times n}$, its <u>singular value decomposition</u> ↙unique! is:

$$A = U D V^T \longleftarrow V \in \mathbb{R}^{n\times n},\ V^TV = I \text{ (orthogonal)}$$

$U \in \mathbb{R}^{m\times m}$
$U^TU = I$ (orthogonal)

$D \in \mathbb{R}^{m\times n}$ diagonal
$D_{ii} \geq 0$, $D_{ij} = 0$ for $i \neq j$

- For any $D \in \mathbb{R}^{m\times n}$ diagonal, its Moore-Penrose pseudoinverse is:

$$D^\dagger \in \mathbb{R}^{n\times m} \text{ diagonal}$$

$$[D^\dagger]_{ii} = \begin{cases} \frac{1}{D_{ii}}, & D_{ii} \neq 0 \\ 0, & \text{otherwise} \end{cases},\quad [D^\dagger]_{ij} = 0 \text{ for } i \neq j$$

- For any $A \in \mathbb{R}^{m\times n}$, its Moore-Penrose pseudoinverse is:

$$A^\dagger = V D^\dagger U^T \in \mathbb{R}^{n\times m}.$$

- $\beta^* = X^\dagger y$ ↙"dagger" is the <u>unique</u> min. ℓ^2-norm solution to normal equation $X^TX\beta = X^Ty$.

<u>Facts:</u> 1) If X is invertible, $X^\dagger = X^{-1}$.
2) If X^TX is invertible, $X^\dagger = (X^TX)^{-1}X^T$. [see unique sol'n case]

- When $X^TX\beta = X^Ty$ has ∞ sol's, they are of the form:

$$\beta = X^\dagger y + v,\quad v \in \underset{\uparrow \text{ nullspace}}{\text{null}(X)}.$$

- For <u>ridge regression</u>, the stationary cond. is $(X^TX + \lambda I)\beta = X^Ty$, where $\overset{\text{regularization parameter}}{\lambda} > 0$.

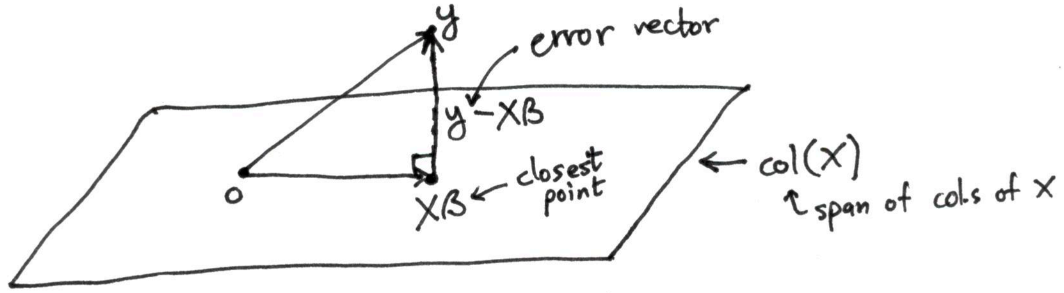
$\Rightarrow$ <u>Solution:</u> $\beta^* = (X^TX + \lambda I)^{-1}X^Ty$ ← always <u>exists</u> & <u>unique</u>!

Moore-Penrose pseudo inv. is limit of ridge soln! ⟶ $\boxed{\lim_{\lambda \to 0^+}(X^TX + \lambda I)^{-1}X^T = X^\dagger}$

Oct 16, 2024
Anuran Makar

⑦ Geometry of linear regression:

$$\min_{\beta} \frac{1}{N}\|y - X\beta\|_2^2$$ ← Given y, find a vector $X\beta \in \mathrm{col}(X)$ that is closest to y!



- error is ⊥ to every col. of X!

$$\rightarrow \quad X^T\underbrace{(y - X\beta)}_{\text{error vector}} = \underset{\text{zero vector}}{0} \iff X^Ty = X^TX\beta \leftarrow \text{"normal" eqn!}$$

Two worlds...

Linear algebra:

min distance to linear space ⟷ **Calculus:** min ℓ^2-norm between y & $X\beta$ } $\min_{\beta} \frac{1}{N}\|y - X\beta\|_2^2$

↕ Hilbert proj. thm — ↕ Differentiate — ↕

projection to space (orthogonality) ⟷ Stationarity condition } $X^TX\beta = X^Ty$

CLUSTERING:

Nov 1, 2024
Anuran Makur

- k-means clustering:

Given $\underset{\substack{\uparrow \\ \text{training data}}}{\{x_1, \dots, x_N\}} \subseteq \mathbb{R}^d$, find clusters $S_1, \dots, S_k$ such that:

$$\min_{S_1,\dots,S_k} \underbrace{\sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|_2^2}_{f(S_1,\dots,S_k,\mu_1,\dots,\mu_k)\ [\text{Objective Value}]}, \quad \text{where } \overset{\text{centroid}}{\mu_i} = \frac{1}{|S_i|} \sum_{x \in S_i} x .$$

- Assignment Step:

Fix μ_i's. Then, the best assignments to min. $f(S_1, \dots, S_k, \mu_1, \dots, \mu_k)$ are:

For x_j, choose cluster $\arg\min_i \|x_j - \mu_i\|_2^2$.

$\uparrow$ break ties deterministically

Assignment will reduce value of Objective func. or keep it the same.

Summary:
- k-means converges in finite time
- $O(k^n)$ iteration complexity in worst-case, but fast in practice
- Only converges to local optima
  - $\hookrightarrow$ multiple init.s
  - $\hookrightarrow$ kmeans ++

- Update Step:

Fix S_i's. Then, find best μ_i's:

$$\min_{\mu_1,\dots,\mu_k} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|_2^2 = \sum_{i=1}^{k} \underbrace{\min_{\mu_i} \sum_{x \in S_i} \|x - \mu_i\|_2^2}_{[*]}$$

$$[*] \quad \min_{\mu_i = (\mu_{i1},\dots,\mu_{id})} \underset{\substack{\uparrow \\ \text{constant}}}{\frac{1}{|S_i|}} \sum_{x \in S_i} \sum_{j=1}^{d} (x_j - \mu_{ij})^2$$

$$= \sum_{j=1}^{d} \min_{\mu_{ij}} \underbrace{\frac{1}{|S_i|} \sum_{x \in S_i}}_{\mathbb{E}[\cdot]} (x_j - \mu_{ij})^2$$

So, it suffices to solve the following problem. For rand. var. $X \in \mathbb{R}$, what is $\min_{c \in \mathbb{R}} \mathbb{E}[(X-c)^2]$?

Prop: $\min_{c \in \mathbb{R}} \mathbb{E}[(X-c)^2] = \operatorname{var}(X)$, $\arg\min_{c \in \mathbb{R}} \mathbb{E}[(X-c)^2] = \mathbb{E}[X]$

Pf: $$\mathbb{E}[(X-c)^2] = \mathbb{E}[(X - \mathbb{E}[X] + \mathbb{E}[X] - c)^2] = \overbrace{\mathbb{E}[(X - \mathbb{E}[X])^2]}^{\operatorname{var}(X)} + \overbrace{(\mathbb{E}[X] - c)^2}^{\geq 0} + 2\underbrace{\mathbb{E}[X - \mathbb{E}[X]]}_{=0}(\mathbb{E}[X] - c)$$

$\geq \operatorname{var}(X)$ with eq. iff $c = \mathbb{E}[X]$. ▨

Hence, Update with reduce value of Objective func. or keep it the same.

- Iteration: (Assignment + Update)
  - An iteration produces new clustering $\Leftrightarrow$ Objective func. strictly decreases.
  - As there are $\leq k^n$ cluster assignments, and each iter. produces new clusterings, k-means converges in $O(k^n)$ steps. [Stop when 2 iter.s have same Obj. Value]

EVALUATION METRICS:

Nov 8, 2024
Anuran Makur

- **F1-score / Dice-Sørensen coefficient:**

$$\frac{1}{F1} \triangleq \frac{1}{2}\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}\right)$$

$\Rightarrow$ $0 \leq F1 \leq 1$

$F1 = 0$ iff Precision $= 0$ or Recall $= 0$

$F1 = 1$ iff Precision $=$ Recall $= 1$

Ex: $F1 \approx 0.947\ldots$ for example in slides!

- **ROC Curve:**
  - Theory of Neyman-Pearson hypothesis testing [not to be confused with Fisher/significance testing]

$H = 0 : X \sim P$ ↕ distribution

$H = 1 : X \sim Q$

↑ hypothesis/class ↑ Observation

Goal: Classify $H=0/H=1$ based on X

  - When we do not know $\mathbb{P}(H=0)$, $\mathbb{P}(H=1)$, we find the best tradeoff between detection (sensitivity/recall) and false-alarm prob: (1-specificity/Type I error)

$$\forall \alpha \in [0,1],\ \underset{\substack{\uparrow \\ \text{ROC (receiver operating characteristic/} \\ \text{Neyman-Pearson func.)}}}{\beta(\alpha)} \triangleq \max_{\hat{H}(\cdot):\ \mathbb{P}(\hat{H}(X)=1|H=0) \leq \alpha} \mathbb{P}(\hat{H}(X)=1|H=1)$$

$\mathbb{P}(\hat{H}(X)=1|H=1)$ ← detection prob.

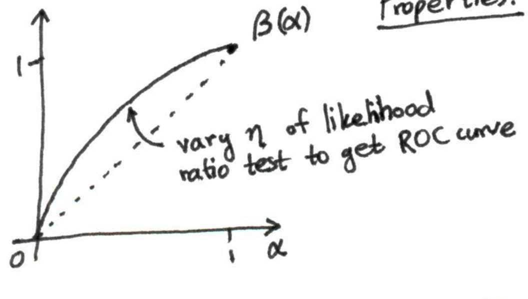
$\mathbb{P}(\hat{H}(X)=1|H=0)$ ← false-alarm prob.

$\hat{H}(\cdot)$ ← max. over all decision rules/classifiers based on X

AUROC: Area under ROC

$$\text{AUROC} = \int_0^1 \beta(\alpha)\, d\alpha \in [0,1]$$

$$= \frac{\text{Gini coeff.} + 1}{2}$$



Properties:

1) $\beta(0) = 0$, $\beta(1) = 1$
2) β is non-decreasing and concave
3) $\beta(\alpha) \geq \alpha$
4) Neyman-Pearson Lemma: (informal) likelihood ratio test:

For each α, $\beta(\alpha)$ achieved by likelihood ratio test:

$$\frac{P(x)}{Q(x)} \underset{\hat{H}(x)=1}{\overset{\hat{H}(x)=0}{\gtrless}} \eta \quad \leftarrow \eta \text{ depends on } \alpha$$

  - Note: In ML, we do not have (P, Q)
    $\Rightarrow$ Build ROC curve for classifier, eg, logistic regression, by varying „threshold on "soft" output.

Nov 11, 2024
Anuran Makur

NAIVE BAYES SUPPLEMENT:
(NB)

- Multinomial Distribution:

Fix a PMF (parameters) $(p_1, p_2, \dots, p_k)$ over $\{1, \dots, k\}$.
$\lfloor\ p_i \geq 0,\ \sum_{i=1}^{k} p_i = 1$

Sample n iid points from $(p_1, \dots, p_k)$. Let X_i = no. of samples equal to i.

The multinomial dist. is the (joint) PMF of $(X_1, \dots, X_k)$:

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k) = \underbrace{\binom{n}{x_1, x_2, \dots, x_k}}_{\text{multinomial coeff.}} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} \quad \text{for all } x_1, \dots, x_k \geq 0 \text{ and } \sum_{i=1}^{k} x_i = n$$

$$\binom{n}{x_1, \dots, x_k} \triangleq \frac{n!}{x_1!\, x_2! \cdots x_k!}$$

- Note: $k = 2$ is binomial.
- Multinomial Thm: $$(a_1 + a_2 + \cdots + a_k)^n = \sum_{\substack{x_1, \dots, x_k \geq 0 \\ x_1 + \cdots + x_k = n}} \binom{n}{x_1, \dots, x_k} a_1^{x_1} a_2^{x_2} \cdots a_k^{x_k}.$$

- To estimate p_i, take $\frac{X_i}{n}$.

- Three Types of NB:

In each case, the data looks like $(\overbrace{X_1, \dots, X_k}^{\text{k-dim feature}}, \overset{\text{label}}{Y})$.

[★ Naïve assumption: $X_1, \dots, X_k$ are indep. given Y ★]

① Gaussian NB: Each X_i is Gaussian given Y (diff. param's over i and $Y=y$)
$\rightarrow$ Learn $O(k)$ params (w/o NB, learn $O(k^2)$ param's)

② Bernoulli NB: Each X_i is Bernoulli given Y (diff. param's over i and $Y=y$)
$\rightarrow$ Learn $O(k)$ params (w/o NB, learn $O(2^k)$ param's)

③ Multinomial NB: $(X_1, \dots, X_k)$ is multinomial given Y (diff params over $Y=y$)
$\rightarrow$ Learn $O(k)$ params [no explicit cond. indep assumed]

Nov 11, 2024
Anuran Makur

LAPLACE SMOOTHING:

☆ **Sunrise Problem:** What is the prob. of Sun rising tomorrow?

☆ **Laplace's answer:** "Rule of Succession"

- Assume Sun rises indep.ly with prob. p each day.
- Assume p is unknown: $p \sim \text{Uniform}([0,1]) \leftarrow f(p) = 1$ for $p \in [0,1]$ (PDF)
- Known: Sun rose on days $1, \ldots, N$. $(N < +\infty)$

Want $\rightarrow \mathbb{P}(\text{Sun rises on day } N+1 \mid \text{Sun rose on days } 1,\ldots,N)$

$$\underset{\substack{\uparrow \\ \text{def. of cond. prob.}}}{=} \frac{\mathbb{P}(\text{Sun rises on days } 1,\ldots,N+1)}{\mathbb{P}(\text{Sun rises on days } 1,\ldots,N)} \overset{[*]}{=} \frac{(1/N+2)}{(1/N+1)} = \boxed{\frac{N+1}{N+2}} \; !!!$$

$$[*] \quad \mathbb{P}(\text{Sun rises on days } 1,\ldots,N) \underset{\substack{\uparrow \\ \text{Law of total probability}}}{=} \int_0^1 \underbrace{\mathbb{P}(\text{Sun rises on days } 1,\ldots,N \mid p)}_{= p^N} \underbrace{f(p)}_{=1} \, dp$$

$$= \int_0^1 p^N \, dp$$

$$= \frac{1}{N+1}$$

☆ **Concept:**

Naive answer	N Sun rises	0 Sun does not rise	$\rightarrow$	$\frac{N}{N+0} = 1$
Laplace's answer	$N+1$ Sun rises	$0+1$ Sun does not rise	$\rightarrow$	$\frac{N+1}{N+1+1} = \boxed{\frac{N+1}{N+2}}$ (Laplace smoothing)

$\swarrow$ estimates of prob.

☆ **Lidstone Smoothing:**

- Add $\alpha \in [0, \infty)$ to each bin!

$\uparrow$ not an integer

GRADIENT DESCENT:

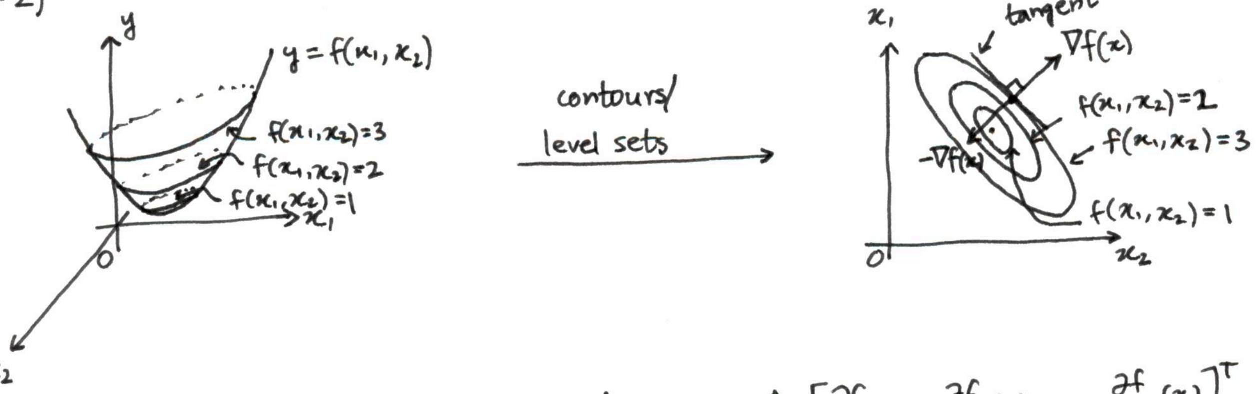
Nov 20, 2024
Anuran Makur

1) Greedy Algorithm:

Given a function $f: \mathbb{R}^d \to \mathbb{R}$, $f(x_1, \dots, x_d)$,

- **Problem:** $\min_{x \in \mathbb{R}^d} f(x)$.

Ex: $(d = 2)$



- The gradient of $f: \mathbb{R}^d \to \mathbb{R}$ is $\nabla f: \mathbb{R}^d \to \mathbb{R}^d$, $\nabla f(x) \triangleq \left[\frac{\partial f}{\partial x_1}(x) \ \ \frac{\partial f}{\partial x_2}(x) \cdots \ \frac{\partial f}{\partial x_d}(x)\right]^T$.

- **Facts:**
  - ∇f points in direction of steepest ascent.
  - $-\nabla f$ points in direction of steepest descent.

- **Recall:** (Directional derivative)

For a unit vector $u \in \mathbb{R}^d$, $\|u\|_2 = 1$, $(D_u f)(x) \triangleq \lim_{\delta \to 0} \frac{f(x+\delta u) - f(x)}{\delta} = \nabla f(x)^T u$

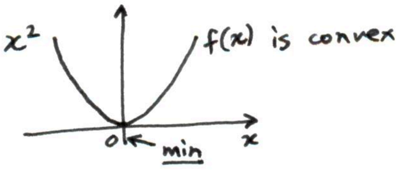
$$-\|\nabla f(x)\|_2 \underset{\substack{\uparrow \\ \text{eq. iff } u = \frac{-\nabla f(x)}{\|\nabla f(x)\|_2}}}{\leq} \overbrace{\nabla f(x)^T u}^{\text{directional deriv.}} \underset{\substack{\uparrow \\ \text{eq. iff } u = \frac{\nabla f(x)}{\|\nabla f(x)\|_2}}}{\leq} \|\nabla f(x)\|_2 \overbrace{\|u\|_2}^{=1} = \|\nabla f(x)\|_2 \quad \left[\text{Cauchy-Schwarz inequality}\right]$$

Hence, $u \propto -\nabla f(x)$ is direction of steepest descent at x.

- **Gradient Descent (GD):**
  1) Initialize at any arb. $x^0 \in \mathbb{R}^d$.
  2) $x^{t+1} = x^t - \underset{\substack{\uparrow \\ \eta > 0 \text{ step-size (small)}}}{\eta} \nabla f(x^t)$ [greedy!]

- **Facts:**
  - GD converges to local min. or saddle point (under cond.s). [Why? See next page.]
  - If f is (strongly) convex, then GD converges to global min.

↳ **Ex:** $f(\underset{\substack{\uparrow \\ d=1}}{x}) = \frac{1}{2}x^2$



Nov 20, 2024
Anuran Makur

2) Fixed Point: [not in exam]

- Banach's fixed point thm: Given any $g: \mathbb{R}^d \to \mathbb{R}^d$ such that g is contractive, i.e., $\forall x, y \in \mathbb{R}^d$, $\underbrace{\|g(x) - g(y)\|_2 \leq L \|x-y\|_2}_{\text{Lipschitz continuous}}$ with $L \in (0,1)$, there is a unique $x^* \in \mathbb{R}^d$ such that:

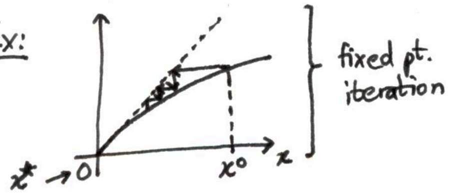
  1) $g(x^*) = x^*$, [fixed point]

  2) For any $y \in \mathbb{R}^d$, $x^{t+1} \overset{\text{iteration}}{=} g(x^t)$, $\overset{\text{initialization}}{x^0 = y}$,

  $\lim_{t\to\infty} x^t = x^*$. [fixed point iteration]

  ↑ conv. exp. fast!

Ex:



- Consider $g(x) = x - \eta \nabla f(x)$. Then, $\underbrace{x^* = g(x^*)}_{\text{fixed point of } g} = x^* - \eta \nabla f(x^*) \Leftrightarrow \underbrace{\nabla f(x^*) = \overset{\text{zero vector}}{0}}_{\text{stationary point of } f}$.

So, fixed pt iteration ⇔ GD algorithm:

$$x^{t+1} = g(x^t) = x^t - \eta \nabla f(x^t)$$

$$\Rightarrow \lim_{t\to\infty} x^t = x^*$$

Hence, GD converges to stat. pt. under cond.s.

3) Gradient Flow: [not in exam]

$$\overset{\text{discrete interval 1}}{x^{t+1} = x^t} - \eta \nabla f(xt) \quad \leftarrow \text{GD iteration}$$

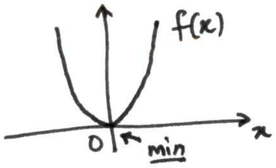
- Change discrete time interval to η: $x^{t+\eta} = x^t - \eta \nabla f(x^t)$

$$\Rightarrow \frac{dx}{dt} \approx \underbrace{\frac{x^{t+\eta} - x^t}{\eta}}_{\text{approx. of derivative!}} = -\nabla f(x^t)$$

ordinary differential equation

- ODE: $\frac{dx}{dt} = -\nabla f(x)$ [Gradient flow]

Ex: $f(x) = \frac{1}{2}x^2$ $(d=1)$



$x(0) = 1$ [Initialize]

$\frac{dx}{dt} = -f'(x) = -x$ ← linear ODE $\Rightarrow x(t) = e^{-t}$ ← exp. fast convergence!

$\lim_{t\to\infty} x(t) = 0$ ← This is the argmin!

- Fact: (informal) If ODE is "stable" (eg: Lyapunov stability), then its solution $x(t)$ converges to $\lim_{t\to\infty} x(t) = x^*$, where x^* is an equilibrium point (i.e., $\nabla f(x^*) = 0$). (Indeed, $x(t) = x^*$ is a steady-state solution to the ODE.)

RELATION BETWEEN SIGMOID & TANH:

Nov 22, 2024
Anuran Makur

- Sigmoid/Logistic function: $\sigma : \mathbb{R} \to [0,1]$,

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

- Hyberbolic Tangent function: $\tanh : \mathbb{R} \to [-1,1]$,

$$\tanh(x) \triangleq \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} = \frac{1-e^{-2x}}{1+e^{-2x}} = \frac{1}{1+e^{-2x}} - \left(1 - \frac{1}{1+e^{-2x}}\right) = \frac{2}{1+e^{-2x}} - 1$$

$\Rightarrow \boxed{\tanh(x) = 2\sigma(2x)-1}$ ← rescaled sigmoid

- Recall: $\tan(x) = \frac{\sin(x)}{\cos(x)}$

$$= \frac{1}{i}\left(\frac{e^{ix}-e^{-ix}}{e^{ix}+e^{-ix}}\right)$$

$\Rightarrow \boxed{\tanh(x) = -i\tan(ix)}$ ← relation to tan(x)

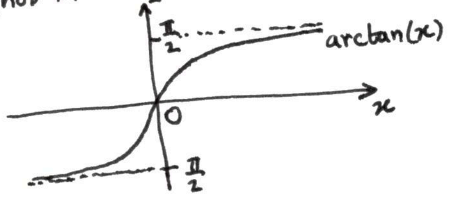
$i = \sqrt{-1}$ (imaginary unit)

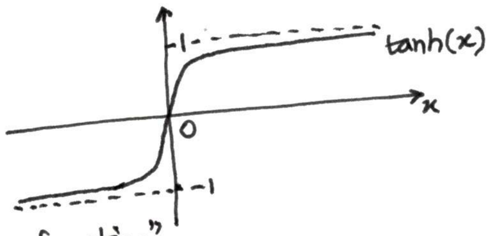
$$e^{ix} = \cos(x) + i\sin(x) \quad \text{[Euler's formula]}$$

$$e^{-ix} = \cos(x) - i\sin(x) \leftarrow \text{conjugate}$$

$$\begin{cases} \cos(x) = \dfrac{e^{ix}+e^{-ix}}{2} \\ \sin(x) = \dfrac{e^{ix}-e^{-ix}}{2i} \end{cases}$$

- Recall: [not in exam]





Relationship? See "Gudermannian function".
↳ relates hyperbolic & circular angles