Lecture 3

HW posted
Thu pm
→ due next
Thu ≤ 12 noon

Prob vs Stat
Large discrete S

Repeated i.i.d sampling
Counts, Binomial, Multinomial

Lecture Notes I – (Discrete) Sample spaces and the Multinomial

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

January 2025

Sample space, outcome, event, probability, . . .

Discrete sample spaces

Repeated independent trials, Binomial, Multinomial

Reading: Ch. 2, 3

Question

Data RANDOM

Model

(Prior knowledge)

Before data

Inference algorithm → ANSWER

RANDOM !

But WHY need model?

PROBABILITY

STATISTICS

Data $= \{x_1, \dots, x_n\} \subset S$

w.p. $\mathbb{P}$ (iid)

Question 1 : Estimate $\mathbb{P}$

Question 2 : Model 1 $\rightarrow P_1$ Model selection

Model 2 $\rightarrow P_2$

...

+ same Data

Large discrete S

$S = \{$ all images B/W , $N_1 \times N_2$ size $\}$

$S = \{B, W\}^{N_1 \times N_2}$

N_1 pixels

N_2 pixels

$= x$

May 12 Women in statistics

$S_0 = \{B, W\}$

$\tilde{S}_0 = \{$ gray levels $\}$

pixel i,j

Assume $N_1 = 200$

$N_2 = 100$

$|S| = 2^{20000} \approx 10^{60}$

$2^{10} = 1024 \approx 10^3$

Q1 • describe P = assume a model P^* that can be represented

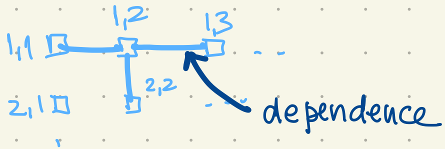
Q2 • inferences $\approx P^*[E]$ $E \subseteq \mathcal{P}(S)$ event

P^* 1. assume pixels independent

$P[\text{pixel } i,j = 1] = p_{ij}$

2. Ising

2. Ising



3. Neural Network, LLM, $= p^*$

$\boxed{p^*} \to x$ samples from $p \approx p^*$

Generative model

$x^{\boxed{t}}_{\boxed{j}} \to$ sample # t

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

$x = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} j$

Repeated independent sampling

$t = 1, 2, 3, \cdots n$

repeated sampling

S = basic sample space, discrete

sample $x^1, x^{\boxed{2}}, \dots x^n \sim P$ → 2nd sample

P probability on S

outcome $\tilde{x} = (x^1, \dots x^n) \in \tilde{S} = \underbrace{S \times S \times \dots S}_{n \text{ times}}$

a sequence

Repeated independent trials, Binomial, Multinomial

A coin is tossed 4 times, and the probability of 1 is $p > 0.5$. The outcomes, their probability and their counts are (in order of decreasing probability):

outcome $\tilde{x}$	$\boxed{n_0}$	$\boxed{n_1}$	$P(x)$	event
1111	0	4	p^4	$E_{0,4}$
1110	1	3	$p^3(1-p)^1$	$E_{1,3}$
1101	1	3	$p^3(1-p)^1$	
1011	1	3	$p^3(1-p)^1$	
0111	1	3	$p^3(1-p)^1$	
1100	2	2	$p^2(1-p)^2$	$E_{2,2}$
1010	2	2	$p^2(1-p)^2$	
1001	2	2	$p^2(1-p)^2$	
0110	2	2	$p^2(1-p)^2$	
0101	2	2	$p^2(1-p)^2$	
0011	2	2	$p^2(1-p)^2$	
0100	3	1	$p^1(1-p)^3$	$E_{3,1}$
1000	3	1	$p^1(1-p)^3$	
0010	3	1	$p^1(1-p)^3$	
0001	3	1	$p^1(1-p)^3$	
0000	4	0	$(1-p)^4$	$E_{4,0}$

Ex $S = \{0, 1\}$

$P[X=0] = 1-p$

$P[X=1] = p$

$E_{3,1} = \{\text{sequences with } n_0 = 3, n_1 = 1\}$

B on $S_B = \{(n_0, n_1)$

$n \geq n_0, n_1 \geq 0$

$n_0 + n_1 = n\}$

$\tilde{S} = S^4$

$n_0 = \#\,0's$

$n_1 = \#\,1s$

COUNTS

Repeated independent trials, Binomial, Multinomial

A coin is tossed 4 times, and the probability of 1 is $p > 0.5$. The outcomes, their probability and their counts are (in order of decreasing probability):

outcome x	n_0	n_1	$P(x)$	event
1111	0	4	p^4	$E_{0,4}$
1110	1	3	$p^3(1-p)^1$	$E_{1,3}$
1101	1	3	$p^3(1-p)^1$	
1011	1	3	$p^3(1-p)^1$	
0111	1	3	$p^3(1-p)^1$	
1100	2	2	$p^2(1-p)^2$	$E_{2,2}$
1010	2	2	$p^2(1-p)^2$	
1001	2	2	$p^2(1-p)^2$	
0110	2	2	$p^2(1-p)^2$	
0101	2	2	$p^2(1-p)^2$	
0011	2	2	$p^2(1-p)^2$	
0100	3	1	$p^1(1-p)^3$	$E_{3,1}$
1000	3	1	$p^1(1-p)^3$	
0010	3	1	$p^1(1-p)^3$	
0001	3	1	$p^1(1-p)^3$	
0000	4	0	$(1-p)^4$	$E_{4,0}$

B on $S_B = \{ (n_0, n_1) \; n \geq n_0, n_1 \geq 0 \; n_0 + n_1 = n \}$

distribution on S_B

$$B((n_0,n_1)) = \binom{n}{n_0, n_1} p^{n_1}(1-p)^{n_0}$$

$$\binom{n}{n_0,n_1} \equiv \binom{n}{n_1} = \binom{n}{n_0}$$

Binomial distribution

$B(n_1, \underset{\uparrow \; n_0+n_1}{n})$

Repeated independent trials, Binomial, Multinomial

$|S| = m > 2$ $\quad x \in S^n$ $\quad S = \{a_1, a_2, \dots a_m\}$

sequence $\quad a_1 \downarrow n_1 \quad \dots \quad a_m \downarrow n_m$

$E_{n_1, n_2 \dots n_m}$

$n_j = \#\{x_i = a_j\}$ = count of a_j in x

$$* \begin{cases} n_j \geq 0 \\ n_1 + n_2 + \dots + n_m = n \end{cases}$$

Events $E_{n_1, n_2 \dots n_m} = \{ x \text{ with counts } (n_1, \dots n_m) \}$

$S_M = \{ (n_1, \dots n_m) \text{ with } * \}$

$m = |S| = 3$, $S = \{\text{red}, \text{blue}, \text{green}\}$, $n = 4$

$x = brrg \in S^4$

$$\left.\begin{array}{l} n_g = 1 \\ n_b = 1 \\ n_r = 2 \end{array}\right\} *$$

Repeated independent trials, Binomial, Multinomial

Remark x with $(n_1, \dots n_m)$
x' with $(\text{—}''\text{—})$ $\Big\} \Rightarrow \tilde{P}(x) = \tilde{P}(x') = p_1^{n_1} p_2^{n_2} \cdots p_m^{n_m}$ depends only on counts

$p_j = P[a_j]$

$x, x' \in E_{(n_1, \dots n_m)}$

$\tilde{P}(x) = p_r^2 p_g p_b = \tilde{P}(x')$

$x' = rbrg$

M on S_M . $M(n_1, \dots n_m) = \binom{n}{n_1 \dots n_m} p_1^{n_1} p_2^{n_2} \cdots p_m^{n_m}$

$$\binom{n}{n_1 \cdots n_m} = \frac{n!}{n_1!\, n_2! \cdots n_m!}$$

Induction over m

Ex $\binom{4}{2,1,1} = \frac{4!}{2!1!1!} = 12$

Repeated independent trials, Binomial, Multinomial

1) $p_0 = p_1 = \frac{1}{2}$

$n = 4$

$$\arg\max_{n_0, n_1} B(n_0, n_1) = (2, 2)$$

2) $p_0 = 0.3 < p_1 = 0.7$ $\quad n = 10$

most probable sequence

$$x^{max} = (\overbrace{1, 1, 1 \ldots 1}^{\times n})$$

$$P(x^{max}) = 0.7^{10}$$

most probable (n_0, n_1) counts

$$(n_0^*, n_1^*) = (3, 7)$$