Видно? слышно?

AIC, энтропия, KL

нестр. опр. Энтропия — сколько в среднем нужно бинарных вопросов, чтобы узнать значение случ-ой величины при оптим-ой стратегии.

"0" "10" "110" "111"

упр. x	1	7	$\sqrt{3}$	банан
$h(x) = P(X=x)$	$\frac{1}{2}$	$\frac{1}{4} = \left(\frac{1}{2}\right)^2$	$\frac{1}{8} = \left(\frac{1}{2}\right)^3$	$\frac{1}{8} = \left(\frac{1}{2}\right)^3$

закон распределения знаем.

Q1. Это „1"? $\xrightarrow{\text{да}} N=1$

$\xrightarrow{\text{нет}}$ это „7"? $\xrightarrow{\text{да}} N=2$

$\xrightarrow{\text{нет}}$ это банан $\xrightarrow{\text{да}} N=3$, $\xrightarrow{\text{нет}} N=3$.

энтропия [бит]

$$H = \left[\frac{1}{2}\cdot 1 + \frac{1}{4}\cdot 2 + \frac{1}{8}\cdot 3 + \frac{1}{8}\cdot 3\right] =$$

сколько вопросов в среднем

$$= \frac{4+2+6}{8} = \frac{12}{8} = \frac{3}{2} = 1.5 \text{ [бит]}$$

1.5 вопроса — в среднем.

def
$$\boxed{H = \sum_x P(X=x)\cdot \log_{1/2} P(X=x)} =$$

$$= -\sum_x P(X=x)\cdot \log_2 P(X=x) \leftarrow \text{[бит]}$$

$$= k\cdot\left(-\sum_x P(X=x)\cdot \ln P(X=x)\right) \leftarrow \text{[нат]} = \text{[natural bit]}$$

Упр. Хочу измерять энтропию в килобайтах $= 1024 \times 8$ бит

$$H = -\sum_x P(X=x)\cdot \log_{?} P(X=x)$$

→ $\ln$ [нат]
→ $\log_2$ [бит]

неф. опр. Энтропия $H(X)$ сколько в среднем бит уходит на передачу одного значения случ-ой величины X при оптим-м кодировании.

Q. как интерпретировать $H < 1$? (бит)
A.

y	A	B
$P(Y_i=y)$	0,99	0.01

(Y_i): AA...AAAABAAAA... .... AAABA..

$$H(Y_i) = 0{,}99\cdot\log_{1/2} 0.99 + 0.01\cdot\log_{1/2} 0.01 = \approx \underline{0{,}024}...$$

Да! можно в среднем на угадывание одного Y_i тратить меньше вопроса, если вопросы задавать про серию (Y_i)

Вопрос: "Правда ли, что сл-ие десять игреков подряд равны A?"

"да"

$$0{,}99^k \approx \frac{1}{2}$$

$Y_1 \ldots \ldots Y_T$

(наз) $\underbrace{H(Y_1,\ldots,Y_T)} = -\sum_y \underbrace{P(Y_1=y_1,\ldots Y_T=y_T)}\cdot \ln P(Y_1=y_1,\ldots Y_T=y_T)$

неф. опр. сколько в среднем бит/вопросов бинарных нужно, чтобы передать полученную выборку, если истинные вер-сти известны и приемнику и передающему.

форм. опр. для величин с плотностью

$$H = -\int_{-\infty}^{\infty} f(y)\cdot \ln f(y)\, dy, \text{ где}$$

$f(y)$ − ф-ция плотности

* интерпр. немного хуже

Относительная энтропия / кросс-энтропия.

RE/CE

x	1	2	$\sqrt{7}$	Банан	
истинная вер-сть	1/2	1/4	1/8	1/8	$P(X=x)$
модельная вер-сть.	$\frac{1}{8}$	$\frac{1}{8}$	1/4	1/2	$Q(X=x)$
	„111"	„110"	„10"	„0"	

оптимальные коды с т. зр. модельной вер-сти.

$RE(P\|Q) = CE(P\|Q)$

неф. опр. среднее кол-во вопросов чтобы узнать X, если структура вопросов (код) оптимален под Q, а истинные вер-сти — это P.

$$CE(P\|Q) = -\sum_x P(X=x)\cdot \ln Q(X=x) \quad \text{(в дискр. вел.)}$$

$$= -\int_{-\infty}^{\infty} f_P(x)\cdot \ln f_Q(x)\, dx \quad \text{(в непр. вел.)}$$

Теорема: $CE(P\|Q) \geq H(P) = CE(P\|P)$

$\underbrace{CE(P\|Q)}_{\text{исп-м неопт-ую кодировку}}$; $\underbrace{CE(P\|P)}_{\text{исп-м оптим-ую кодировку.}}$

Опр. Дивергенция Кульбака-Лейблера $D_{KL}(P\|Q) = KL(P\|Q)$ — потери от неоптимальной кодировки по сравнению с оптимальной (бит/нат..)

$$KL(P\|Q) = CE(P\|Q) - H(P)$$

Теорема $KL(P\|Q) \geq 0.$

! чтобы посчитать $KL(P\|Q)$, надо знать P и Q.

1. P — не знаем
2. обычно $\hat{M1}$ и $\hat{M2}$ получены в процессе отбора модели из множества.

Истина P ?

пока даже оценивания нет

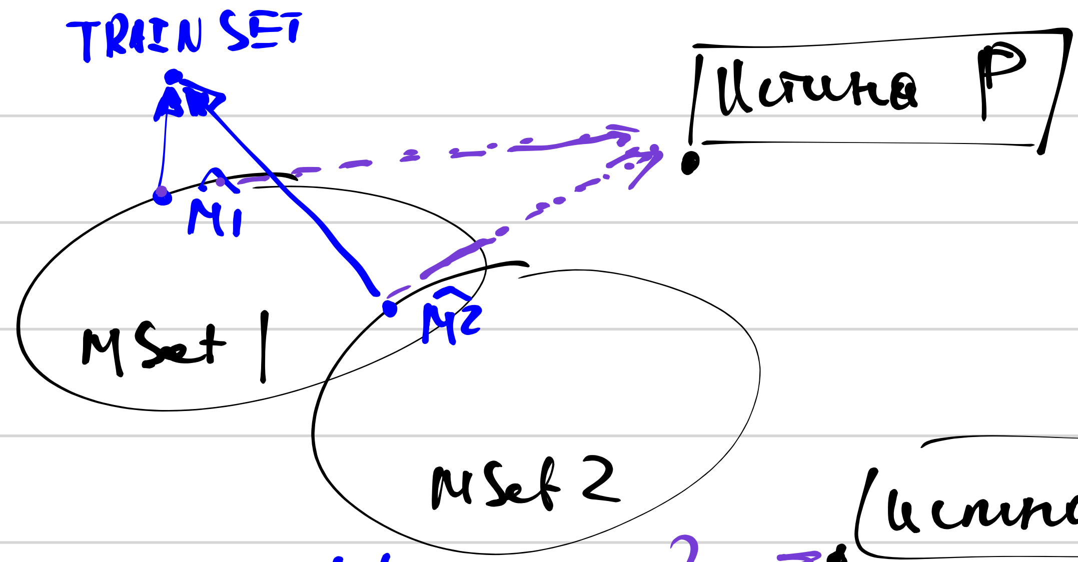
M1: $Y_1 \dots Y_T \sim N(5;10)$ нез.

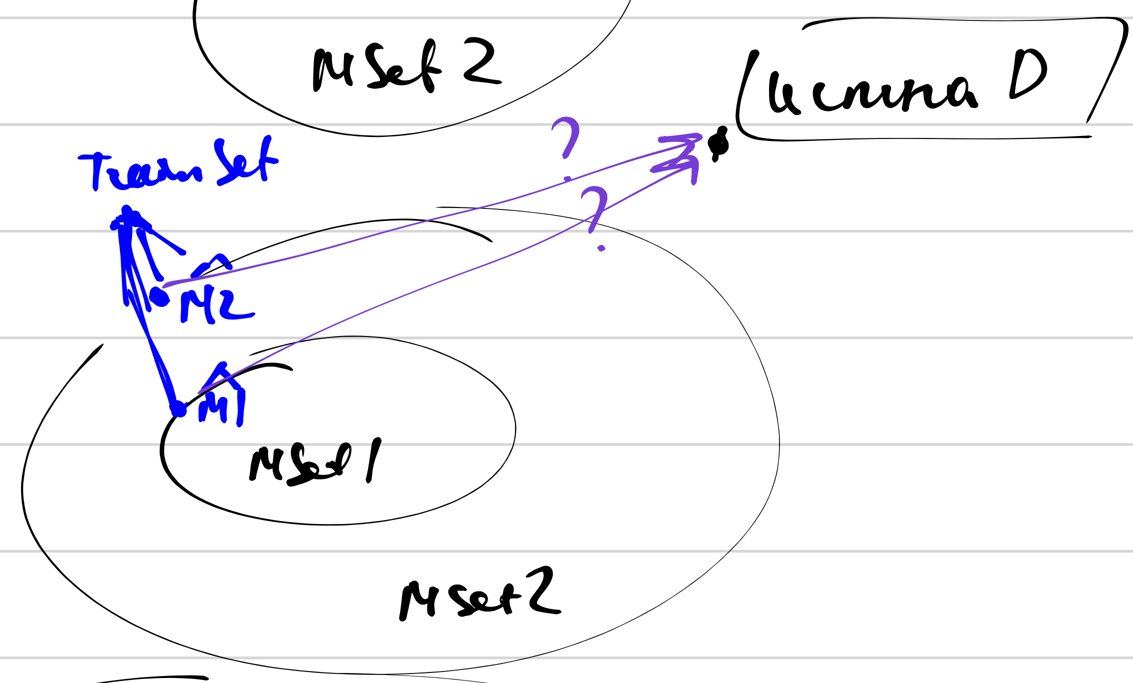
M2: $Y_t = u_t + u_{t-1} + 6$, $u_t \sim N(0;4)$. незав

(1) оценить $KL(P\|M1)$ невозможно!

(2) Но! оказывается при очень мягких предпосылках можно оценить

$$KL(P\|M1) - KL(P\|M2)$$





Akaike: $\dfrac{AIC(\hat{M1}) - AIC(\hat{M2})}{2} \approx KL(P \| \hat{M1}) - KL(P \| \hat{M2})$

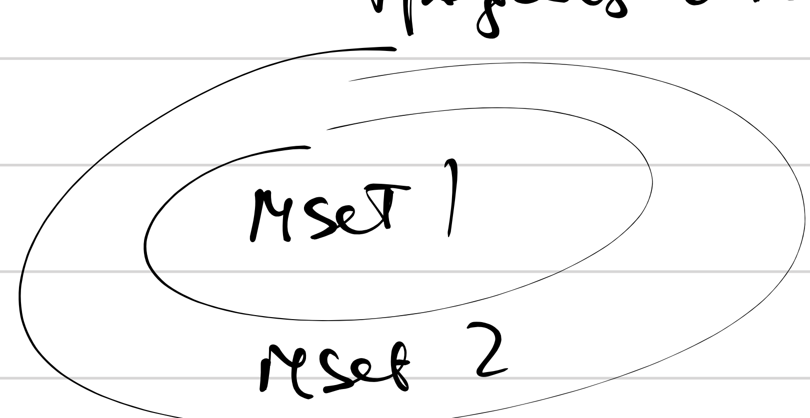
$$\boxed{AIC(M) = 2 \cdot k - 2 \cdot \ln \hat{L}}$$ канон

k – число параметров в модели M

$\hat{L}$ – максимум правдоподобия модели M.

Q. Зачем Акаике умножал и делил на 2?

A. Для вложенных моделей есть традиционный LR тест.



$H_0: P \in MSet\ 1$

$H_A: P \notin MSet\ 1,\ P \in MSet\ 2$

$$LR_{test} = 2 \cdot (\ln \hat{L}_2 - \ln \hat{L}_1)$$

тест отношения правдоподобия

$LR_{Test} \xrightarrow[n \to \infty,\ \text{распр.}]{H_0} \chi^2_{k_2 - k_1}$

Ловушка! (!) модель = функция правдоподобия выписана для тех же $y_1 \dots y_T$

W1

Mset 1 $y_i \sim \exp(\alpha \cdot u_i + \beta)$
$u_i \sim N(0;1)$ незав

можно

Mset 2 $y_i \sim N(\mu, 1)$ незав

$\hookrightarrow f(y_1 \dots y_T)$

пример табу

Mset 1 $y_1 \dots y_T \sim$ ETS(AAA) $m = 12$

$\theta = \begin{pmatrix} \alpha, \beta, \gamma, \sigma^2, \\ l_0, b_0, \\ s_0, s_{-1} \dots s_{-11} \end{pmatrix}$

$f(y_1 \dots y_T)$

Mset 2 y_1 — будем считать известным

$y_t = \underline{\underline{\alpha}} + \underline{\underline{\beta}} \cdot y_{t-1} + u_t$ $\quad u_t \sim N(0; \sigma^2)$ нез

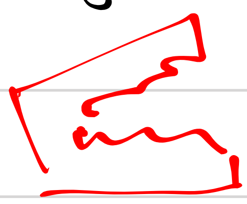
$t = 2, \dots, T$

$\theta = \begin{pmatrix} \alpha \\ \beta \\ \sigma^2 \end{pmatrix}$

$f(y_2, y_3 \dots y_T \mid y_1) \to \max\limits_{\theta}$

по AIC сравнить нельзя.

Ловушка 2



иногда авторы используют неканонич. опр. AIC
(убирают из него константу)

например я работаю с $\begin{pmatrix} y_1 \\ \vdots \\ y_T \end{pmatrix} \sim N(\mu; C)$

$$f(y_1 \dots y_T) = \frac{1}{\sqrt{(2\pi)^T \cdot \det C}} \exp\left(-\frac{1}{2}(y-\mu)^T \cdot C^{-1} (y-\mu)\right)$$

"$-\frac{T}{2}\ln(2\pi)$" — могу не включить в AIC

на разницу $AIC(\hat{M}_1) - AIC(\hat{M}_2)$.