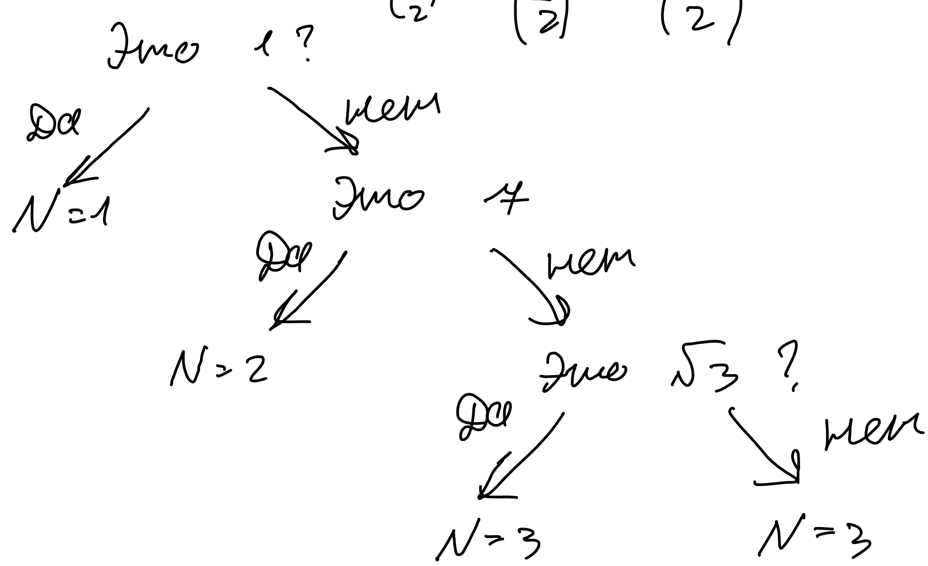
AIC, энтропия, KL

Неформальное определение:

Энтропия — сколько в среднем нужно бинарных вопросов, чтобы узнать значение реализации случайной величины при оптимальной стратегии.

	0	10	110	111
x	1	7	$\sqrt{3}$	Банан
$p(X=x)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$
		$\left(\frac{1}{2}\right)^2$	$\left(\frac{1}{2}\right)^3$	$\left(\frac{1}{2}\right)^3$

Закон распр. мы знаем



$$H = \frac{1}{2}\cdot 1 + \frac{1}{4}\cdot 2 + \frac{1}{8}\, 3 + \frac{1}{8}\cdot 3 = 1{,}5 \text{ [бит]}$$

В среднем 1,5 вопроса

$$H = \sum_x p(X=x) \log_{\frac{1}{2}} p(X=x) = -\sum_x p(X=x) \log_2 p(X=x)$$

$$H = -\sum_x p(X=x)\ln p(X=x) \quad \widetilde{(\text{natural bit})}$$

Неф. опред. г. Энтропия $H(X)$ – сколько в среднем бит уходит на передачу одного значения с. в. X при оптимальном кодировании

y	A	B
$p(Y=y)$	0,99	0,01

Разрешаем отвечать на пачку вопросов

$$H = -(0,99\log_2 0,99 + 0,01\log_2 0,01) \approx$$
$$\approx 0,024$$

$Y_1, \dots, Y_T$

$$H(Y_1, \dots, Y_T) = -\sum_{y} P(Y_1 = y_1, \dots, Y_T = y_T) \cdot$$
$$\cdot \ln P(Y_1 = y_1, \dots, Y_T = y_T)$$

Сколько в среднем бит информации нужно чтобы передать полученную выборку, если истинная модель известна обеим сторонам

$$H = -\int_{-\infty}^{\infty} f(x)\ln f(x)\,dx$$

f - ф-я плотности.

Относительная энтропия / кросс-энтропия

x	1	2	$\sqrt{3}$	Банан	
Истина	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	$P(X=x)$
Модель	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{2}$	$Q(X=x)$
	111	110	10	0	

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

CE – среднее количество вопросов чтобы узнать X, если структура вопросов (код) оптимальна под Q, а истинные вер-и – P.

$$CE(P\,\|\,Q) = -\sum P(X=x)\log_2 Q(X=x), \text{ бит}$$

$$-\int_{-\infty}^{\infty} f_P(x)\cdot \ln f_Q(x)\,dx, \text{ нат}$$

Теорема: $\underbrace{CE(P\,\|\,Q)}_{\text{неопт. код}} \geqslant H(P) = \underbrace{CE(P\,\|\,P)}_{\text{опт. код}}$

Дивергенция Кульбака–Лейблера

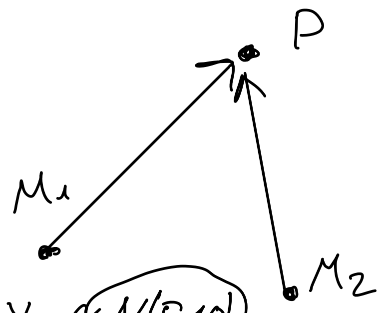
$$D_{KL}(P\,\|\,Q) = KL(P\,\|\,Q)$$

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

Теорема: $KL(P\,\|\,Q) \geqslant 0$

1) P – не знаем

2) обычно $\hat{M}_1$ и $\hat{M}_2$ получаются в процессе выбора модели



$Y_1, \dots, Y_T \sim N(5;10)$ iid

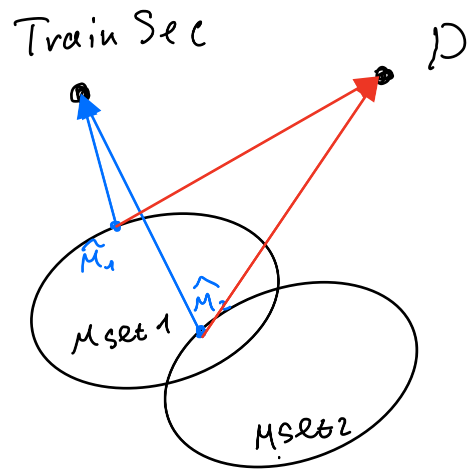
$Y_t = u_t + u_{t-1} + 6$

$u_t \sim N(0,4)$ iid

1) Оценить $KL(P \| M_1)$ невозможно

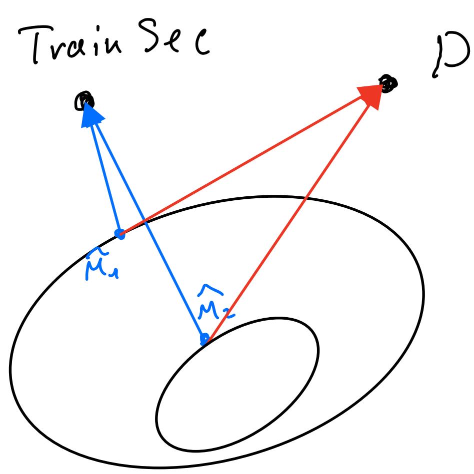
2) При очень мягких предпосылках можно оценить

$$KL(P \| M_1) - KL(P \| M_2)$$



Mset 1 – ETS

Mset 2 – ARIMA



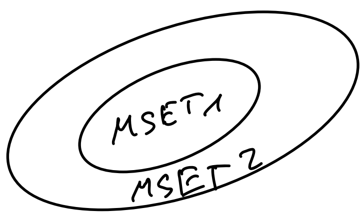
AIC – Akaike Information Criteria

$$\frac{AIC(\hat{M}_1) - AIC(\hat{M}_2)}{2} \approx \boxed{KL(P \| \hat{M}_1) - KL(P \| \hat{M}_2)}$$

$$AIC(\hat{M}) = 2 \cdot \underbrace{K}_{\text{кол-во параметров}} - 2 \ln \hat{L}$$

$\hat{L}$ – максимум правдоподобия

Для вложенных моделей существует «традиционный» LR-тест.



H_0: $P \in MSet\,1$

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

$$LR_{test} = 2(\ln \hat{L}_2 - \ln \hat{L}_1) \qquad LR \xrightarrow[n \to \infty]{H_0} \chi^2_{K_2 - K_1}$$

MSet 1 $\quad y_i \sim \exp(\alpha u_i + \beta)$, $u_i \sim iid\ N(0,1)$

MSet 2 $\quad y_i \sim N(\mu, 1)$ iid

$\qquad y_i \in \{y_1, \ldots, y_T\}$

1) MSet 1 $\quad y_1, \dots, y_T \sim ETS(A, A, A)$, $m = 12$

$$\Theta = \begin{pmatrix} \alpha, \beta, \gamma, \sigma^2 \\ \ell_0, b_0 \\ s_0, s_1, \dots, s_{11} \end{pmatrix}$$

$$f(y_1, \dots, y_T)$$

MSet 2 $\quad y_t = \alpha + \beta y_{t-1} + u_t \quad u_t \sim N(0, \sigma^2)$

$t = 2, \dots, T$

$\boxed{y_1}$ – известен

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

$$f(y_2, \dots, y_T \mid y_1)$$

Модели должны быть оценены на одинаковых данных

2) Авторы пакетов могут по-разному заводить критерии

AIC, AICc, BIC, HQIC