Data Drift

O que é um drift?

- Mudança de distribuição da base de dados
- Mudança na relação dos dados com a classe
- Novos padrões presentes na base de dados com a evolução temporal dos dados

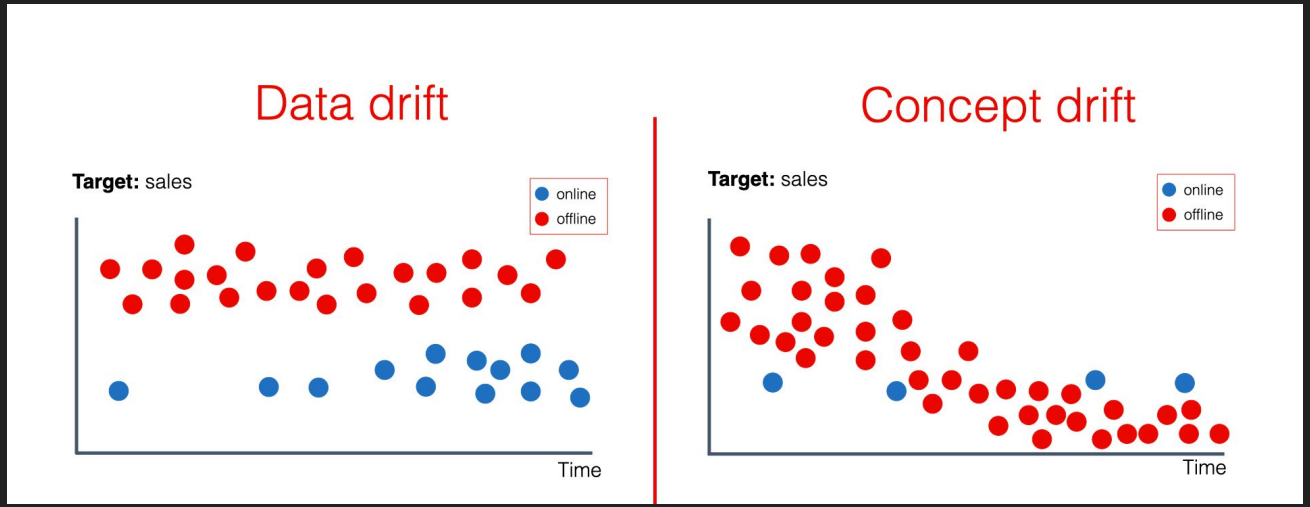
O que é um drift?

Data drift vs Concept drift

Data drift: representa mudanças nas propriedades estatísticas dos dados de entrada.

Concept drift: ocorre quando a relação entre o dado de entrada e a classe mudam

Data drift vs Concept drift



https://www.evidentlyai.com/ml-in-production/data-drift

Data drift vs Concept drift

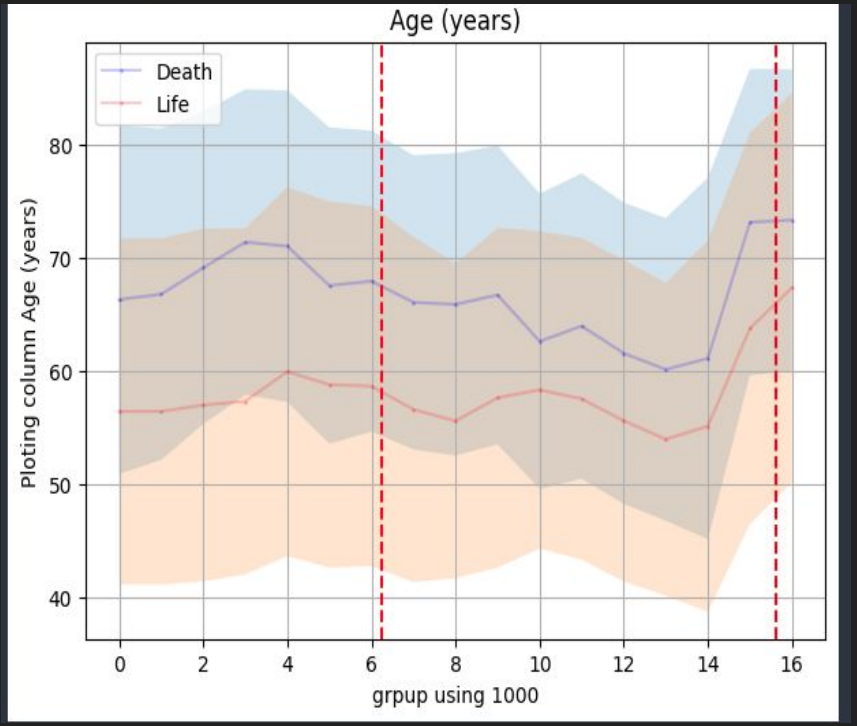
Data drift: representa mudanças nas propriedades estatísticas dos dados de entrada.

Ex: uso do sistema operacional. Com o passar do tempo mais pessoas estão usando SOs abertos.

Data drift vs Concept drift

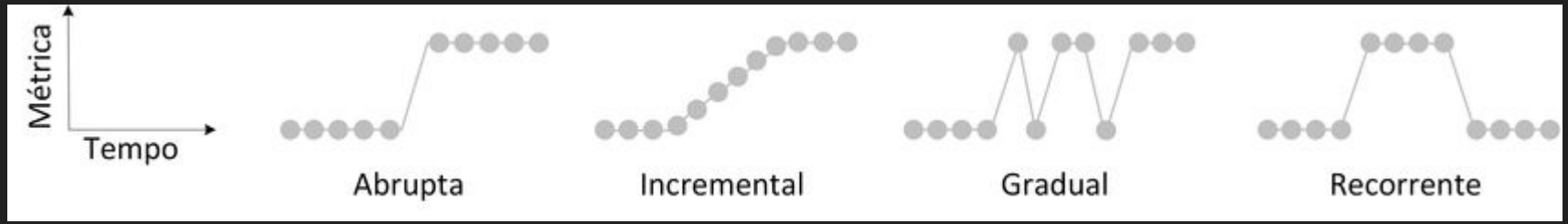
Concept drift: ocorre quando a relação entre o dado de entrada e a classe

Ex: Covid



Concept drift

Concept drift: ocorre quando a relação entre o dado de entrada e a classe mudam



Data Drift

Formas de medir: Verificando a estatística dos dados para observar mudanças de padrões:

- Chi-square test (Pearson (1900))
- Kolmogorov-Smirnov test (Massey Jr (1951))
- Jensen-Shannon distance (Lin (1991))

Concept Drift

Formas de medir: Verificando a saída do modelo:

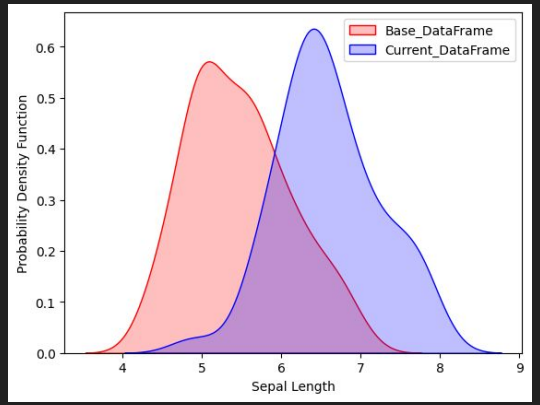
- Student - Teacher (Cerqueira, 2022)
- Adaptive Windowing (Bifet, 2007)
- ….

Vamos a um exemplo

Notebook:

Testes estatísticos

- O princípio básico deste teste é comparar proporções, ou seja, possíveis divergências entre as frequências observadas e esperadas para um certo evento.
- Verificar se a mudança é “real” e não simplesmente aleatória
- Alguns métodos são mais sensíveis que outros
- Fatores:
  - Tipo das features
  - Tamanho
  - Tipo de drift



TESTE DO QUI-QUADRADO

- Verificar se a frequência com que um determinado acontecimento observado em uma amostra se desvia significativamente ou não da frequência com que ele é esperado.

Karl Pearson (1857-1936)

TESTE DO QUI-QUADRADO: Condições

- Os grupos devem ser independentes,
- Valores devem ser discretos
- Os itens de cada grupo são selecionados aleatoriamente,
- As observações devem ser frequências ou contagens,
- A amostra deve ser relativamente grande (pelo menos 5 observações)

Definição

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Oi: frequência observada para cada classe

Ei: frequência esperada para aquela classe

Quando as frequências observadas são muito próximas às esperadas, o valor de X^2 é pequeno, e quando as divergências são grandes, consequentemente assume valores altos.

Definição

Frequência esperada para cada item para montar a tabela de valores esperados

$$E_{ij} = \frac{(\text{total da linha i}) \times (\text{total da coluna j})}{\text{total geral}}$$

Soma das frequências em cada linha e coluna, dividido sobre o valor total

Hipóteses

- **Hipótese nula** (H0) – frequências observadas = frequências esperadas.
  - Não detectamos drift
- **Hipótese alternativa** (H1) – as frequências observadas ≠ frequências esperadas.
  - Detectamos drift
- **Nível de significância** (α): significa o risco de se rejeitar uma hipótese verdadeira. Deverá ser estabelecido antes da análise de dados e é usualmente fixado em 5% (P=0,05).

Passos

É necessário obter duas estatísticas :

- X^2 calculado: obtido diretamente dos dados das amostras.
- X^2 tabelado: depende do número de graus de liberdade e do nível de significância adotado.

Se X^2 calculado ≥ X^2 tabelado: Rejeita-se Ho.

Se X^2 calculado < X^2 tabelado: Aceita-se Ho.

Passos

É necessário obter duas estatísticas :

- X^2 calculado: obtido diretamente dos dados das amostras.
- X^2 tabelado: depende do número de graus de liberdade e do nível de significância adotado.

Se X^2 calculado ≥ X^2 tabelado: Rejeita-se Ho. -> Drift

Se X^2 calculado < X^2 tabelado: Aceita-se Ho.

Será que homens e mulheres têm diferente enquadramento profissional?

Gênero	Exerce atividade profissional	Não exerce	Total
Homens	106	90	196
Mulheres	245	81	326
Total	351	171	522

Será que homens e mulheres têm diferente enquadramento profissional?

Passo 1: Estabelecer hipótese estatisticas:

H0: Homens e mulheres não diferem significativamente quanto ao enquadramento profissional

Se X^2 calculado < X^2 tabelado: Aceita-se Ho.

H1: Homens e mulheres diferem significativamente quanto ao enquadramento profissional

Se X^2 calculado ≥ X^2 tabelado: Rejeita-se Ho.

Passo 2: Nível de significância (a probabilidade máxima de aceitar um erro):

$\alpha = 0.05$ (erro tolerado de 5%)

Será que homens e mulheres têm diferente enquadramento profissional?

Passo 3: calcular o valor do teste:

Comparar as frequências observadas com as frequências esperadas

Gênero	Exerce atividade profissional	Não exerce	Total
Homens	106	90	196
Mulheres	245	81	326
Total	351	171	522

Gênero	Exerce atividade profissional	Não exerce
Homens	132	64
Mulheres	219	107

Será que homens e mulheres têm diferente enquadramento profissional?

Passo 3: calcular o valor do teste:

Comparar as frequências observadas com as frequências esperadas

Gênero	Exerce atividade profissional	Não exerce	Total
Homens	106	90	196
Mulheres	245	81	326
Total	351	171	522

Gênero	Exerce atividade profissional	Não exerce
Homens	132	64
Mulheres	219	107

X2≈ ((106-132)^2)/132=5.1

Será que homens e mulheres têm diferente enquadramento profissional?

Passo 3: calcular o valor do teste:

Comparar as frequências observadas com as frequências esperadas

Gênero	Exerce atividade profissional	Não exerce	Total
Homens	106	90	196
Mulheres	245	81	326
Total	351	171	522

Gênero	Exerce atividade profissional	Não exerce
Homens	132	64
Mulheres	219	107

$$\chi^2 \approx 5.1+10.24+2.99+6.24=24.52$$

Será que homens e mulheres têm diferente enquadramento profissional?

Graus de liberdade (tamanho da tabela):

df=(rows−1)×(columns−1)

df=(2−1)(2−1)=1

Valor crítico para α = 0.05: 3.841

Como χ 2 = 24.52 > 3.841 χ2 =24.52>3.841, rejeitamos a hipótese nula

Existe uma relação entre o gênero da pessoa e o fato de ela exercer uma atividade profissional

G.L. \ P	0,99	0,98	0,95	0,90	0,80	0,70	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	$0,0^{3}2$	$0,0^{3}6$	0,004	0,016	0,064	0,148	0,455	1,074	1,642	2,706	3,841	5,412	6,635	10,827
2	0,020	0,040	0,103	0,211	0,446	0,713	1,386	2,408	3,219	4,605	5,991	7,824	9,210	13,815
3	0,115	0,185	0,352	0,584	1,005	1,424	2,366	3,665	4,642	6,251	7,815	9,837	11,345	16,266
4	0,297	0,429	0,711	1,064	1,649	2,195	3,357	4,878	5,989	7,779	9,488	11,668	13,277	18,467
5	0,554	0,752	1,145	1,610	2,343	3,000	4,351	6,064	7,289	9,236	11,070	13,388	15,080	20,515
6	0,872	1,134	1,635	2,204	3,070	3,828	5,348	7,231	8,558	10,645	12,592	15,033	16,812	22,457
7	1,239	1,564	2,167	2,833	3,822	4,671	6,346	8,383	9,803	12,017	14,067	16,622	18,475	24,322
8	1,646	2,032	2,733	3,490	4,594	5,527	7,344	9,524	11,030	13,362	15,507	18,168	20,090	26,125
9	2,088	2,532	3,325	4,168	5,380	6,393	8,343	10,656	12,242	14,684	16,919	19,679	21,666	27,877
10	2,558	3,059	3,940	4,865	6,179	7,267	9,342	11,781	13,442	15,987	18,307	21,161	23,209	29,588
11	3,053	3,609	4,575	5,578	6,989	8,148	10,341	12,899	14,631	17,275	19,675	22,618	24,725	31,264
12	3,571	4,178	5,226	6,304	7,807	9,034	11,340	14,011	15,812	18,549	21,026	24,054	26,217	32,909
13	4,107	4,765	5,892	7,042	8,634	9,926	12,340	15,119	16,985	19,812	22,362	25,472	27,688	34,528
14	4,660	5,368	6,571	7,790	9,467	10,821	13,339	16,222	18,151	21,064	23,685	26,873	29,141	36,123
15	5,229	5,985	7,261	8,547	10,307	11,721	14,339	17,322	19,311	22,307	24,996	28,259	30,578	37,697
16	5,812	6,614	7,962	9,312	11,152	12,624	15,338	18,418	20,465	23,542	26,296	29,633	32,000	39,252
17	6,408	7,255	8,672	10,085	12,002	13,531	16,338	19,511	21,615	24,769	27,587	30,995	33,409	40,790
18	7,015	7,906	9,390	10,865	12,857	14,440	17,338	20,601	22,760	25,989	28,869	32,346	34,805	42,312
19	7,633	8,567	10,117	11,651	13,716	15,352	18,338	21,689	23,900	27,204	30,144	33,687	36,191	43,820
20	8,260	9,237	10,851	12,443	14,578	16,266	19,337	22,775	25,038	28,412	31,410	35,020	37,566	45,315
21	8,897	9,915	11,591	13,240	15,445	17,182	20,337	23,858	26,171	29,615	32,671	36,343	38,932	46,797
22	9,542	10,600	12,338	14,041	16,314	18,101	21,337	24,939	27,301	30,813	33,924	37,659	40,289	48,268
23	10,196	11,293	13,091	14,848	17,187	19,021	22,337	26,018	28,429	32,007	35,172	38,968	41,638	49,728
24	10,856	11,992	13,848	15,659	18,062	19,943	23,337	27,096	29,553	33,196	36,415	40,270	42,980	51,179
25	11,524	12,697	14,611	16,473	18,940	20,867	24,337	28,172	30,675	34,382	37,652	41,566	44,314	52,620

Será que homens e mulheres têm diferente enquadramento profissional?

$$\text{p-valor} = P(\chi^2 \geq \text{valor observado}) = 1 - F_{\chi^2}(\text{valor observado})$$

Fχ2(x) = função de distribuição acumulada da distribuição qui-quadrado com os **graus de liberdade apropriados**

No exemplo:

$df=(r-1)\times(c-1)=1$

P-value ~ 0,00000078 - >

0,00000078 < 0.05 -> fortemente rejeita H0

Vamos testar com novos valores

Será que homens e mulheres têm diferente enquadramento profissional?

Gênero	Exerce atividade profissional	Não exerce	
Homens	320	90	410
Mulheres	300	80	380
Total	620	170	790

Vamos testar com novos valores

$$E_{ij} = \frac{(\text{total da linha i}) \times (\text{total da coluna j})}{\text{total geral}}$$

Gênero	Exerce atividade profissional	Não exerce	
Homens	320	90	410
Mulheres	300	80	380
Total	620	170	790

Gênero	Exerce atividade profissional	Não exerce
Homens	321	88=
Mulheres		

Vamos testar com novos valores

Passos do teste Chi-square Pareado

Vendas em Janeiro

Gênero	Quantidade	Proporção
Homens	700	70%
Mulheres	300	30%
Total	1000	100%

Vendas em Maio

Gênero	Quantidade	Proporção
Homens	400	40%
Mulheres	600	60%
Total	1000	100%

Drift nos dados

$$E_{ij} = \frac{(\text{total da linha i}) \times (\text{total da coluna j})}{\text{total geral}}$$

Frequência Observada

Gênero	Observado (Produção)	Esperado (com base em Janeiro)
Homens	400	700
Mulheres	600	300

Frequência Esperada

Gênero	Observado (Produção)	Esperado (com base em Janeiro)
Homens	550.00	550.00
Mulheres	450.00	450.00

Drift nos dados

Frequência Observada

Gênero	Observado (Produção)	Esperado (com base em Janeiro)
Homens	400	700
Mulheres	600	300

Frequência Esperada

Gênero	Exerce atividade profissional	Não exerce
Homens	550.00	550.00
Mulheres	450.00	450.00

Gênero	Observado (Produção)	Esperado (com base em Janeiro)
Homens	$(400-550)^2/550$	$(700-550)^2/550$
Mulheres	$(600-450)^2/450$	$(300-450)^2/450$

Drift nos dados

Frequência Observada

Gênero	Observado (Produção)	Esperado (com base em Janeiro)
Homens	400	700
Mulheres	600	300

Frequência Esperada

Gênero	Exerce atividade profissional	Não exerce
Homens	550.00	550.00
Mulheres	450.00	450.00

Gênero	Observado (Produção)	Esperado (com base em Janeiro)
Homens	$(400-550)^2/550$	$(700-550)^2/550$
Mulheres	$(600-450)^2/450$	$(300-450)^2/450$

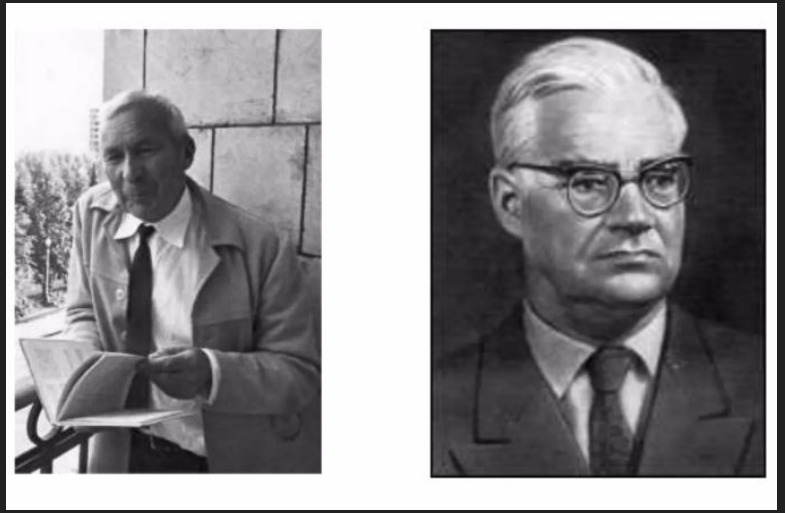
D value: 181.81 - > P - value menor que 0.05 Rejeitar H0 - > temos um drift

Código

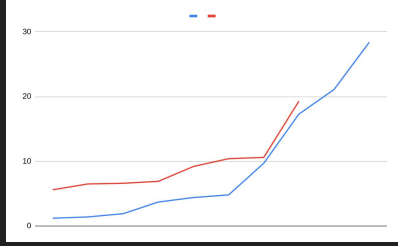
Neste Notebook:

Kolmogorov-Smirnov test (KS)

- Verifica se existe uma diferença significativa entre a frequência observada e esperada.
- Variáveis contínuas



KS test



- A ideia básica do teste é computar a distância entre as frequências acumuladas e comparar com o valor crítico.

- Dado duas hipóteses:
  - Hipótese nula: As distribuições de X e Y são idênticas (sem drift)
  - Hipótese alternativa: As distribuições diferem. (com drift)

- A ideia é que se o valor de distância for maior que o valor crítico

  ->rejeitar H0 **(drift detectado)**

- Se o valor for menor que o valor crítico -> Aceitar H0

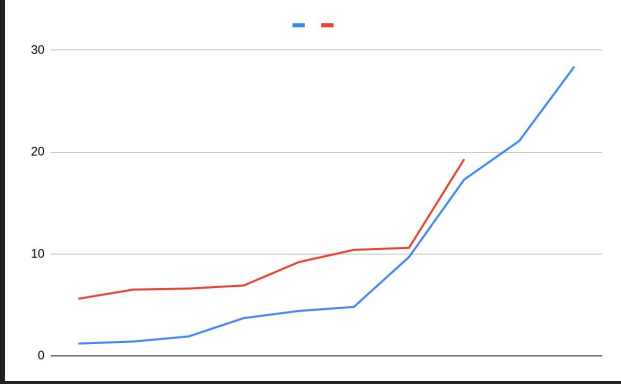
KS passos

1. Ordenar os dados
2. Converter os dados para distribuição cumulativa
3. Encontrar a diferença absoluta máxima (D) entre as distribuições
4. Se D é maior que o D crítico, então as distribuições são diferentes. Caso contrário, não temos evidências suficientes.
5. O P-Value também pode ser usado

Exemplo

X : 1.2, 1.4, 1.9, 3.7, 4.4, 4.8, 9.7, 17.3, 21.1, 28.4

Y : 5.6, 6.5, 6.6, 6.9, 9.2, 10.4, 10.6, 19.3.



Hipótese Nula: Não há diferença significativa entre as distribuições de X e Y

Hipótese Alternativa: Há diferença significativa entre as distribuições de X e Y (drift detectado).

Exemplo

X : 1.2, 1.4, 1.9, 3.7, 4.4, 4.8, 9.7, 17.3, 21.1, 28.4

Y : 5.6, 6.5, 6.6, 6.9, 9.2, 10.4, 10.6, 19.3.

CDF(x) = (número de elementos ≤ x) / n

Value	**1.2**	**1.4**	**1.9**	**3.7**	**4.4**	**4.8**	**5.6**	**6.5**	**6.6**	**6.9**	**9.2**	**9.7**	**10.4**	**10.6**	**17.3**	**19.3**	**21.1**	**28.4**

Exemplo

X : 1.2, 1.4, 1.9, 3.7, 4.4, 4.8, 9.7, 17.3, 21.1, 28.4

Y : 5.6, 6.5, 6.6, 6.9, 9.2, 10.4, 10.6, 19.3.

CDF(x) = (número de elementos ≤ x) / n

Value	**1.2**	**1.4**	**1.9**	**3.7**	**4.4**	**4.8**	**5.6**	**6.5**	**6.6**	**6.9**	**9.2**	**9.7**	**10.4**	**10.6**	**17.3**	**19.3**	**21.1**	**28.4**
Fx	0.1	0.2	0.3	0.4	0.5	0.6	0.6	0.6	0.6	0.6	0.6	0.7	0.7	0.7	0.8	0.8	0.9	1
Fy																		

Exemplo

X : 1.2, 1.4, 1.9, 3.7, 4.4, 4.8, 9.7, 17.3, 21.1, 28.4

Y : 5.6, 6.5, 6.6, 6.9, 9.2, 10.4, 10.6, 19.3.

CDF(x) = (número de elementos ≤ x) / n

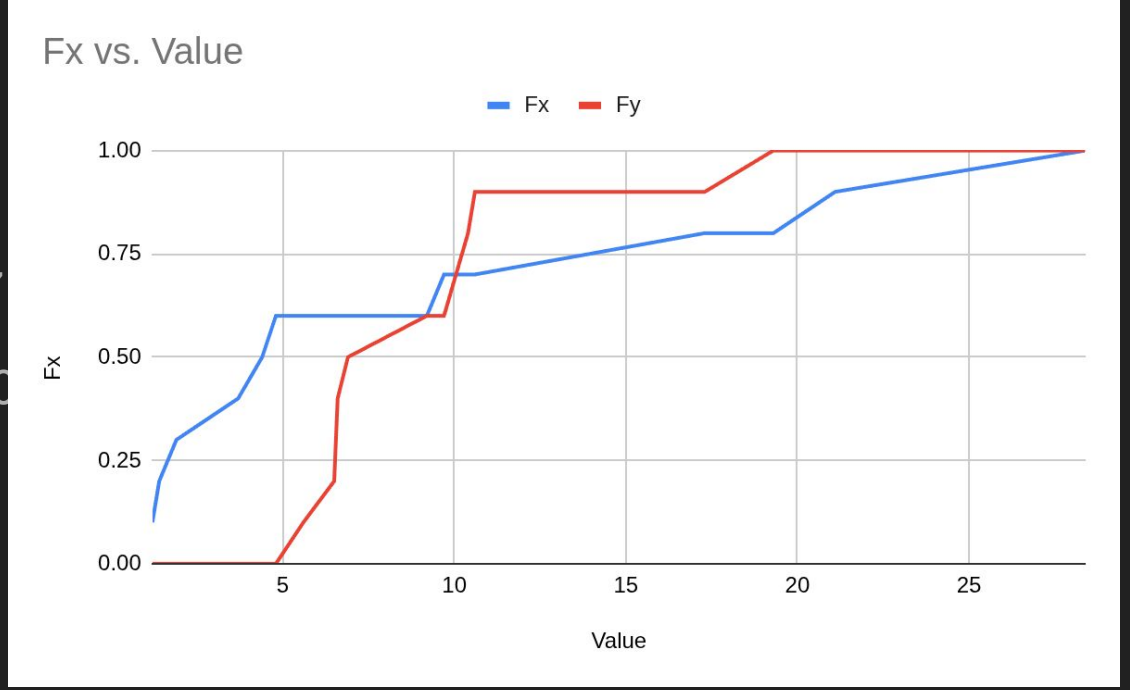
Value	**1.2**	**1.4**	**1.9**	**3.7**	**4.4**	**4.8**	**5.6**	**6.5**	**6.6**	**6.9**	**9.2**	**9.7**	**10.4**	**10.6**	**17.3**	**19.3**	**21.1**	**28.4**
Fx	0.1	0.2	0.3	0.4	0.5	0.6	0.6	0.6	0.6	0.6	0.6	0.7	0.7	0.7	0.8	0.8	0.9	1
Fy	0	0	0	0	0	0	0.1	0.2	0.4	0.5	0.6	0.6	0.8	0.9	0.9	1	1	1

Exemplo

X : 1.2, 1.4, 1.9, 3.7, 4.4, 4.8, 9.7

Y : 5.6, 6.5, 6.6, 6.9, 9.2, 10.4, 10

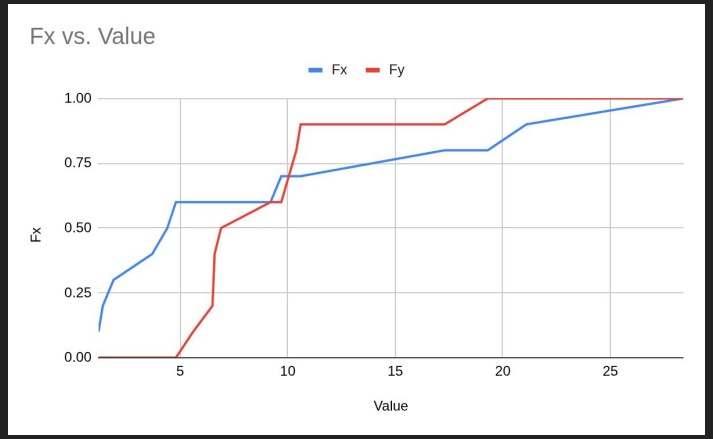
CDF(x) = (número de elementos



Value	**1.2**	**1.4**	**1.9**	**3.7**	**4.4**	**4.8**	**5.6**	**6.5**	**6.6**	**6.9**	**9.2**	**9.7**	**10.4**	**10.6**	**17.3**	**19.3**	**21.1**	**28.4**
Fx	0.1	0.2	0.3	0.4	0.5	0.6	0.6	0.6	0.6	0.6	0.6	0.7	0.7	0.7	0.8	0.8	0.9	1
Fy	0	0	0	0	0	0	0.1	0.2	0.4	0.5	0.6	0.6	0.8	0.9	0.9	1	1	1

Encontrando a distância Máxima

Value	1.2	1.4	1.9	3.7	4.4	**4.8**	5.6	6.5	6.6	6.9	9.2	9.7	10.4	10.6	17.3	19.3	21.1	28.4
Fx	0.1	0.2	0.3	0.4	0.5	**0.6**	0.6	0.6	0.6	0.6	0.6	0.7	0.7	0.7	0.8	0.8	0.9	1
Fy	0	0	0	0	0	**0**	0.1	0.2	0.4	0.5	0.6	0.6	0.8	0.9	0.9	1	1	1
|Diff|	0.1	0.2	0.3	0.4	0.5	**0.6**	0.5	0.4	0.2	0.1	0	0.1	-0.1	-0.2	-0.1	-0.2	-0.1	0



KS

D Calculado = 0.6

Para duas amostras, o valor do D crítico com 95% de confiança é definido pela fórmula:

D Cre´=1.36 * sqrt((|X| + |Y|)/(|X|*|y|))

D = 1.36 * raiz((8+10)/(8*10))

D = 0.645

Como D Máximo < D Crítico então

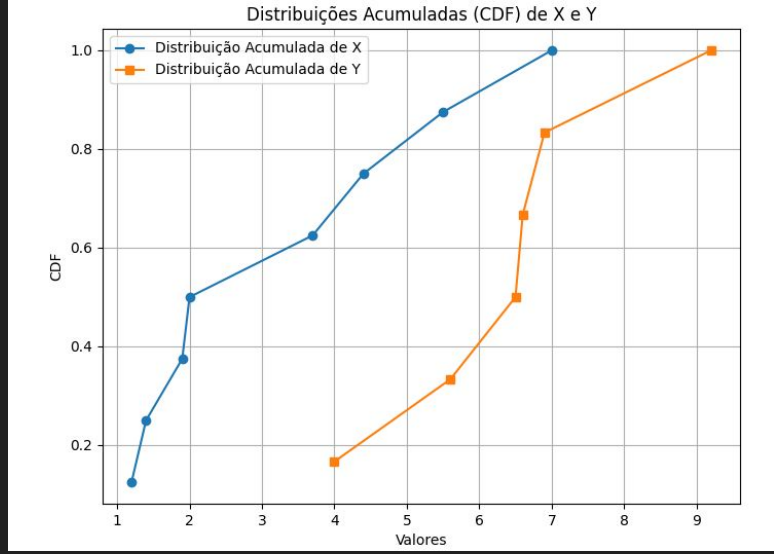
Não se rejeita a hipótese nula (não tem diferença estatística) - sem Drift

Ou P-value > 0.05 -> não rejeita a hipotese nula

Prática

X = np.array([1.2, 1.4, 1.9, 3.7, 2, 4.4, 5.5, 7])

Y = np.array([4, 5.6, 6.5, 6.6, 6.9, 9.2])



Atividade

Acessar o Colab

Fuente: F. J. Massey, Jr., *The Kolmogorov-Smirnov test for goodness of fit*, J. Amer Statistical Assoc. 46 (1951), 68 – 78.

	$1-\alpha$				
n	0.80	0.85	0.90	0.95	0.99
1	0.900	0.925	0.950	0.975	0.995
2	0.684	0.726	0.776	0.842	0.929
3	0.565	0.597	0.642	0.708	0.828
4	0.494	0.525	0.564	0.624	0.733
5	0.446	0.474	0.510	0.565	0.669
6	0.410	0.436	0.470	0.521	0.618
7	0.381	0.405	0.438	0.486	0.577
8	0.358	0.381	0.411	0.457	0.543
9	0.339	0.360	0.388	0.432	0.514
10	0.322	0.342	0.368	0.410	0.490
11	0.307	0.326	0.352	0.391	0.468
12	0.295	0.313	0.338	0.375	0.450
13	0.284	0.302	0.325	0.361	0.433
14	0.274	0.292	0.314	0.349	0.418
15	0.266	0.283	0.304	0.338	0.404
16	0.258	0.274	0.295	0.328	0.392
17	0.250	0.266	0.286	0.318	0.381
18	0.244	0.259	0.278	0.309	0.371
19	0.237	0.252	0.272	0.301	0.363
20	0.231	0.246	0.264	0.294	0.356
25	0.210	0.220	0.240	0.270	0.320
30	0.190	0.200	0.220	0.240	0.290
35	0.180	0.190	0.210	0.230	0.270
Fórmula para una n mayor	$\frac{1.07}{\sqrt{n}}$	$\frac{1.14}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

Divergência de Jensen-Shannon

- Mede a similaridade entre duas distribuições de probabilidade
- Baseada na entropia de Shannon
- Fórmula:

$$JS(P \,||\, Q) = \tfrac{1}{2} * KL(P \,||\, M) + \tfrac{1}{2} * KL(Q \,||\, M), \text{ onde } M = \tfrac{1}{2}(P + Q)$$

Sempre entre 0 e 1 (0 = distribuições idênticas, 1 = completamente diferentes)

Divergência de Jensen-Shannon

- Dividir os dados em janelas temporais
- Calcular a distribuição de classes ou atributos
- Usar a divergência de Jensen-Shannon entre janelas consecutivas
- Aumento na divergência indica possível concept drift

Divergência de Jensen-Shannon

- Mede distância simétrica e finita entre distribuições
- Funciona com distribuições discretas e contínuas
- Exige discretização de dados contínuos
- Custo computacional em grandes volumes de dados

Jensen-Shannon: exemplo

- Janela 1: P = [0.6, 0.4]
- Janela 2: Q = [0.3, 0.7]
- Representam proporções de duas classes 'A' e 'B'

Jensen-Shannon: exemplo

Passo 1: Calcular a Média das Distribuições:

$M = \frac{1}{2}(P + Q)$

$M = \frac{1}{2}([0.6, 0.4] + [0.3, 0.7])$

$M = [0.45, 0.55]$

Jensen-Shannon: exemplo

Passo 2: KL Divergence de P para M:

$$KL(P \,||\, M) = 0.6 * \log_2(0.6 / 0.45) + 0.4 * \log_2(0.4 / 0.55)$$

$$\approx 0.6 * 0.4150 + 0.4 * (-0.4594)$$

$$\approx 0.249 - 0.1838 \approx 0.0652$$

Jensen-Shannon: exemplo

Passo 3: KL Divergence de Q para M:

$$KL(Q \,||\, M) = 0.3 * \log_2(0.3 / 0.45) + 0.7 * \log_2(0.7 / 0.55)$$

$$\approx 0.3 * (-0.5849) + 0.7 * 0.3474$$

$$\approx -0.1755 + 0.2432 \approx 0.0677$$

Jensen-Shannon: exemplo

Passo 4: Calcular a JS Divergence:

$$JS(P \,||\, Q) = \tfrac{1}{2} * KL(P \,||\, M) + \tfrac{1}{2} * KL(Q \,||\, M)$$

$$\approx \tfrac{1}{2} * (0.0652 + 0.0677)$$

$$\approx 0.06645$$

Jensen-Shannon: exemplo

- Divergência JS ≈ 0.066 → pequena mudança
- Limiares sugeridos:
  - JS < 0.1 → normal
  - 0.1 ≤ JS < 0.3 → possível drift
  - JS ≥ 0.3 → drift significativo

Referências

BEIGUELMAN, B. 1996.Curso de Bioestatística Básica. 4ed. Ribeirão Preto:

Sociedade Brasileira de Genética.