

PROBABILIDADE E DISTÂNCIAS

Definições Básicas

Variável aleatória

2

- Em Estatística, é muito comum ver o termo **variável aleatória**. Mas qual o seu significado?
 - ▣ Existem várias definições para o termo variável aleatória, todas equivalentes:
- **Definição 1:** Uma variável aleatória **X** é um tipo de variável que pode assumir diferentes valores numéricos, definidos para cada evento de um espaço amostral
 - ▣ Confuso?

Variável aleatória

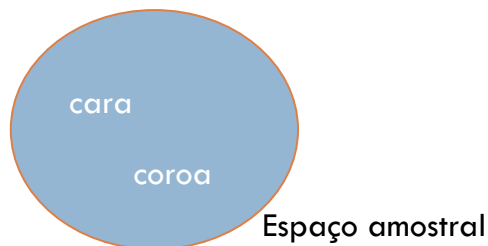
3

- Simplificando...
 - ▣ Uma variável aleatória pode ser entendida como uma variável quantitativa
 - ▣ Seu resultado (ou valor) depende de fatores aleatórios
 - Exemplo: lançamento de um dado ou moeda
- E o espaço amostral? O que é?
 - ▣ Conjunto de todos os resultados possíveis para uma variável aleatória

Variável aleatória

4

- Exemplo: lançamento de uma moeda
 - ▣ Espaço amostral: **cara** e **coroa**;
 - ▣ Variável aleatória: resultado obtido no lançamento de uma moeda (**cara** ou **coroa**)



Variável aleatória

5

- De modo geral, uma variável aleatória pode ser de classificada em dois tipos básicos
 - ▣ Variável aleatória discreta
 - ▣ Variável aleatória contínua

Variável aleatória

6

- Variável aleatória discreta
 - ▣ Trata-se da variável cujos valores podem ser contados ou listados
 - Valor de um dado: $\{1, 2, 3, 4, 5, 6\}$
 - Lançamento de uma moeda: cara ou coroa

Variável aleatória

7

- Variável aleatória discreta
 - ▣ Os valores desse tipo de variável pertencem a um conjunto finito ou infinito (desde que numerável)
 - ▣ Conjunto finito
 - Valor de um dado: $\{1, 2, 3, 4, 5, 6\}$
 - ▣ Conjunto infinito numerável
 - Número de pessoas de numa fila: $\{0, 1, 2, 3, 4, \dots \infty\}$.
 - Conjunto dos inteiros!

Variável aleatória

8

- Variável aleatória contínua
 - ▣ Trata-se da variável que pode assumir qualquer valor em um determinado intervalo ou coleção de intervalos
 - ▣ Todos os seus valores possíveis não podem ser listados como no caso das variáveis discretas

Variável aleatória

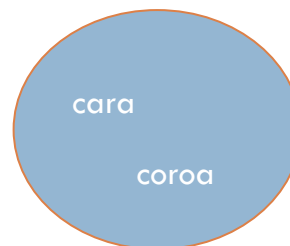
9

- Variável aleatória contínua
 - ▣ Trata-se de uma variável que assume valores dentro de intervalos de números reais
 - Peso das pessoas em uma sala
 - Altura das pessoas em uma sala
 - Distância entre cidades
 - etc

Variável aleatória

10

- Voltando ao exemplo do lançamento de uma moeda.
- Probabilidades são calculadas a partir da realizações da variável aleatória X
 - ▣ $P(X = \text{cara}) = 0,5 = 50\%$
 - ▣ $P(X = \text{coroa}) = 0,5 = 50\%$

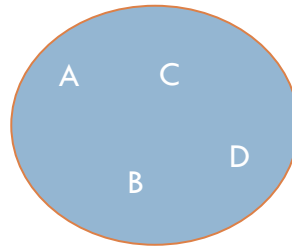


Espaço amostral

Variável aleatória

11

- Exemplo: lançamento de duas moedas
 - ▣ Espaço amostral
 - A: cara-cara
 - B: cara-coroa
 - C: coroa-cara
 - D: coroa-coroa

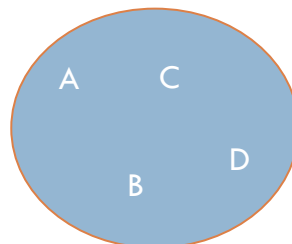


Espaço amostral

Variável aleatória

12

- Probabilidades são calculadas a partir da realizações da variável aleatória X
 - ▣ $P(X = A) = 0,25 = 25\%$
 - ▣ $P(X = B) = 0,25 = 25\%$
 - ▣ $P(X = C) = 0,25 = 25\%$
 - ▣ $P(X = D) = 0,25 = 25\%$

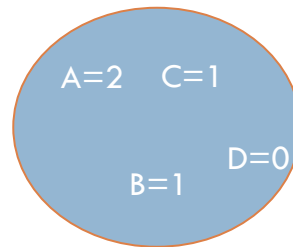


Espaço amostral

Variável aleatória

13

- E se considerássemos o evento “número de caras”?
- Espaço amostral
 - A: cara-cara = 2
 - B: cara-coroa = 1
 - C: coroa-cara = 1
 - D: coroa-coroa = 0

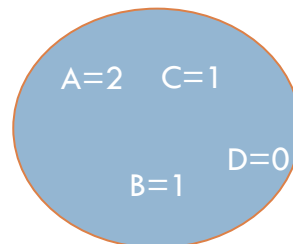


Espaço amostral

Variável aleatória

14

- Probabilidades são calculadas a partir da realizações da variável aleatória X
 - $P(X = 2) = 0,25 = 25\%$
 - $P(X = 1) = 0,50 = 50\%$
 - $P(X = 0) = 0,25 = 25\%$



Espaço amostral

Probabilidade

15

- É uma medida ou estimativa de quão provável é de que algo vai acontecer ou de que uma declaração é verdadeira
 - Se eu jogar uma moeda, qual a probabilidade do valor ser “cara”?
 - Se eu jogar um dado, qual a probabilidade do valor ser “2”?

Probabilidade

16

- Lei de Laplace
 - A probabilidade de um acontecimento associado a uma certa experiência aleatória é dada pelo quociente entre o número de casos favoráveis ao acontecimento e o número de casos possíveis

Probabilidade

17

- A probabilidade é sempre representada por um valor entre 0 e 1
 - 0: 0% de possibilidade ou “não acontecerá”
 - 1: 100% de possibilidade ou “acontecerá”
- Quanto maior a probabilidade, mais provável será de acontecer
 - Ou, maior é o número de vezes que se espera que esse evento aconteça ao longo do tempo

Evento

18

- É o conjunto de resultados possíveis associado a um experimento \mathcal{E} e relativo a um determinado espaço amostral
 - Esse conjunto de resultados é um subconjunto do espaço amostral
 - A esse conjunto de resultados é associado um valor de probabilidade

Evento

19

- Exemplo: lançamento de dois dados
 - ▣ Espaço amostral S : todos os resultados possíveis
 - $S = \{(1,1), (1,2), (1,3), (1,4), \dots, (6,6)\}$
 - ▣ Evento A : a soma dos dados ser igual a 7
 - $A = \{(1,6), (2,5), (3, 4), (4,3), (5,2), (6,1)\}$
 - ▣ O evento A é um subconjunto do espaço amostral S
 - $A \subset S$

Evento

20

- A cada evento A , associa-se um número real representado por $P(A)$.
- Está é a probabilidade de A ocorrer no espaço amostral S , e ela deve respeitar as seguintes propriedades
 - ▣ $0 \leq P(A) \leq 1$
 - ▣ $P(S) = 1$

Combinação de Eventos

21

- Em probabilidade condicional, podemos estudar simultaneamente dois eventos
- Nessa situação, existem duas possibilidades quanto à relação entre as suas probabilidades
 - ▣ Elas serem **eventos independentes**
 - ▣ Elas serem **eventos dependentes**

Combinação de Eventos

22

- **Eventos independentes**
 - ▣ Dados dois eventos A e B, temos que a ocorrência do evento A em nada interfere na probabilidade de ocorrência do evento B
 - ▣ Nesse caso, a probabilidade de que ambos aconteçam ao mesmo tempo é igual ao produto de suas probabilidades
 - $P(A \text{ e } B) = P(A \cap B) = P(A) * P(B)$

Combinação de Eventos

23

- Exemplo de evento independente
 - ▣ A probabilidade de em uma família nascer um menino e ele ter olhos azuis
 - Nesse caso, a probabilidade do sexo da criança em nada interfere na probabilidade dela vir a ter olhos azuis
 - ▣ O lançamento de duas moedas

Combinação de Eventos

24

- Eventos dependentes
 - ▣ Dados dois eventos A e B, temos que a ocorrência do evento A exerce influência na probabilidade de ocorrência do outro evento, B

Combinação de Eventos

25

- Exemplo de evento dependente
 - ▣ A probabilidade de em uma família nascer um menino e ele ser daltônico
 - ▣ O gene do daltonismo na espécie humana está ligado ao sexo. Ele é provocado por genes recessivos localizados no cromossomo X (sem alelos no Y). Assim, o problema ocorre muito mais frequentemente nos homens que nas mulheres

Combinação de Eventos

26

- Eventos dependentes
 - ▣ Nesse caso, a probabilidade de ambos ocorrerem **ao mesmo tempo** assume um valor diferente dependendo da natureza da relação
 - ▣ Dados dois eventos A e B, a **probabilidade condicional de A dado B** é definida como o quociente entre a probabilidade conjunta de A e B, e a probabilidade de B:
 - $P(A|B) = \frac{P(A \cap B)}{P(B)}$
 - $P(B) > 0$

Combinação de Eventos

27

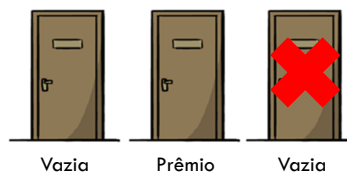
- Problema de Monty Hall
 - ▣ No Brasil, a porta dos desesperados (Sérgio Mallandro)
- Regras
 - ▣ O concorrente deve escolher uma de 3 portas para ganhar o prêmio
 - ▣ Uma das portas vazias é revelada
 - ▣ O concorrente pode escolher mudar de porta



Combinação de Eventos

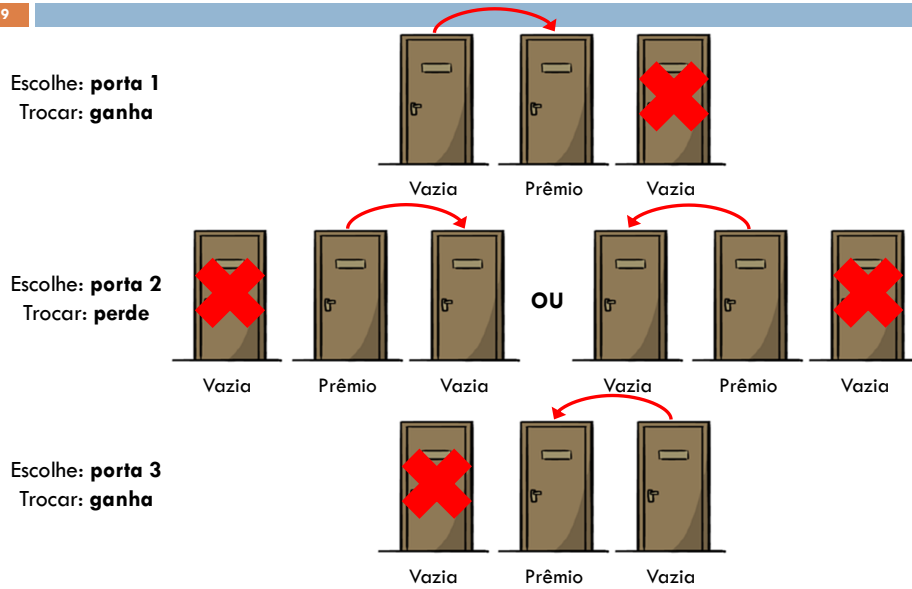
28

- Diante disso, devemos trocar de porta?
- Resposta intuitiva
 - ▣ Não. Com somente 2 portas, a probabilidade é de 50% para cada
 - ▣ Verdadeiro se fossem eventos independentes
 - A escolha do concorrente e a porta revelada são eventos que dependem um do outro



Combinação de Eventos

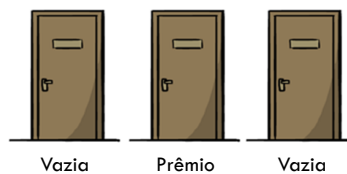
29



Combinação de Eventos

30

- Diante disso, devemos trocar de porta?
 - ▣ Sim.
 - ▣ Probabilidade de ganhar usando a porta inicial: $1/3$
 - ▣ Probabilidade de ganhar ao trocar de porta: $2/3$
 - Ver simulação



Combinação de Eventos

31

- Exemplo: cálculo da probabilidade condicional de um evento dependente
 - ▣ 250 alunos estão matriculados numa universidade
 - 100 homens e 150 mulheres
 - 110 no BCC e 140 no BSI

Sexo\Curso	BCC	BSI	Total
H	40	60	100
M	70	80	150
Total	110	140	250

Combinação de Eventos

32

- Exemplo: cálculo da probabilidade condicional de um evento dependente
 - ▣ Num sorteio, qual a probabilidade de sair alguém do BSI dado que o sorteada uma mulher?

$$□ P(BSI|M) = \frac{P(BSI \cap M)}{P(M)} = \frac{\frac{80}{250}}{\frac{150}{250}} = \frac{80}{150} = 0,53 = 53\%$$

Combinação de Eventos

33

- Por fim, temos também o complemento de uma probabilidade
 - ▣ $P(A^C) = 1 - P(A)$
- A probabilidade complementar de um evento A é a probabilidade de A não ocorrer
 - ▣ Ao lançarmos um dado, a probabilidade de sair um 6 será: $P(6) = 1/6$
 - ▣ A probabilidade de sair qualquer outro número será: $P(6^C) = 1 - 1/6 = 5/6$

Teorema de Bayes

34

- O Teorema de Bayes relaciona as probabilidades de A e B com suas respectivas probabilidades condicionadas

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}, \text{ para } P(B) > 0$$

- ▣ Onde
 - $P(A)$ e $P(B)$: probabilidades a **priori** de A e B
 - $P(B|A)$ e $P(A|B)$: probabilidades a **posteriori** de B condicional a A e de A condicional a B respectivamente.

Teorema de Bayes

35

- O Teorema de Bayes nos permite calcular a probabilidade a posteriori para um determinado padrão pertencente a uma determinada classe
 - Em resumo

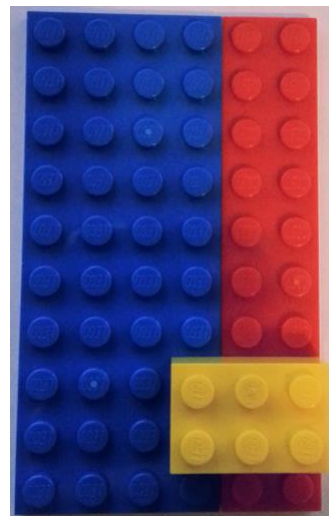
$$Prob\ Posteriori = \frac{Prob\ Priori * Distrib\ Prob}{Evidencia}$$

Teorema de Bayes

36

- Exemplo
 - Considere o conjunto de peças de Lego ao lado
 - Perceba que o Lego Amarelo sempre esconde uma das cores: **vermelho** ou **azul**
 - Qual a probabilidade de sair a cor **vermelha** dado selecionamos um ponto amarelo?

$$P(\text{vermelho}|\text{amarelo}) = ?$$



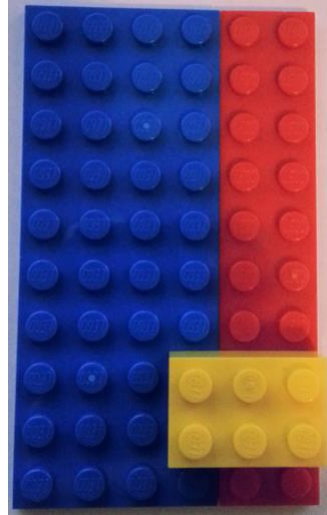
Teorema de Bayes

37

- Temos 60 pontos: calcular probabilidades
 - ▣ Probabilidade **vermelho**

$$P(\text{vermelho}) = \frac{20}{60} = \frac{1}{3}$$
 - ▣ Probabilidade **azul**

$$P(\text{azul}) = \frac{40}{60} = \frac{2}{3}$$
 - ▣ Soma das probabilidades dá 1

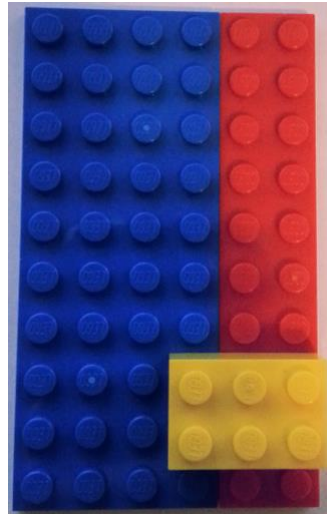


Teorema de Bayes

38

- Faltou calcular a probabilidade do amarelo
 - ▣ Probabilidade amarelo

$$P(\text{amarelo}) = \frac{6}{60} = \frac{1}{10}$$
 - ▣ Se somarmos as 3 probabilidades, o resultado é maior do que 1!
 - A peça amarela sempre vem com um outra cor
 - Probabilidade condicional!



Teorema de Bayes

39

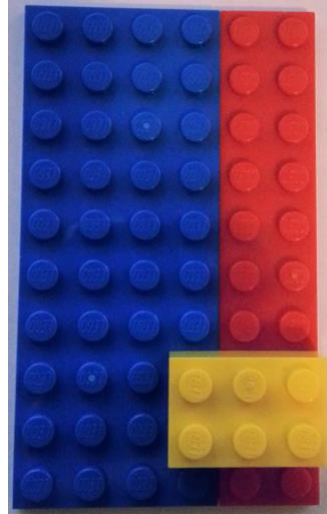
□ Probabilidades do amarelo

- Em relação ao **vermelho**

$$P(\text{amarelo}|\text{vermelho}) = \frac{4}{20} = \frac{1}{5}$$

- Em relação ao **azul**

$$P(\text{amarelo}|\text{azul}) = \frac{2}{40} = \frac{1}{20}$$



Teorema de Bayes

40

□ Voltando ao problema

- Qual a probabilidade de sair a cor **vermelha** dado selecionamos um ponto amarelo?

$$P(\text{vermelho}|\text{amarelo})$$

- Isso equivale a calcular

$$\frac{P(\text{vermelho})P(\text{amarelo}|\text{vermelho})}{P(\text{amarelo})}$$

Teorema de Bayes

41

- Voltando ao problema
 - Qual a probabilidade de sair a cor **vermelha** dado selecionamos um ponto amarelo?

$$\frac{P(\text{vermelho})P(\text{amarelo}|\text{vermelho})}{P(\text{amarelo})}$$

- Substituindo as probabilidades

$$P(\text{vermelho}|\text{amarelo}) = \frac{1/3 * 1/5}{1/10} = \frac{2}{3}$$

Função Distribuição de Probabilidades

42

- Suponha um conjunto de valores que uma variável aleatória **X** possa assumir:
 - x_1, x_2, \dots, x_n
- A partir dos seus resultados possíveis, podemos construir uma função de probabilidades **p(x_i)**
 - $p(x_i) \geq 0$, onde $p(x_i)$ é a probabilidade associada a $X = x_i$
 - A soma das probabilidades é sempre 1 (equivalente a 100%)

Função Distribuição de Probabilidades

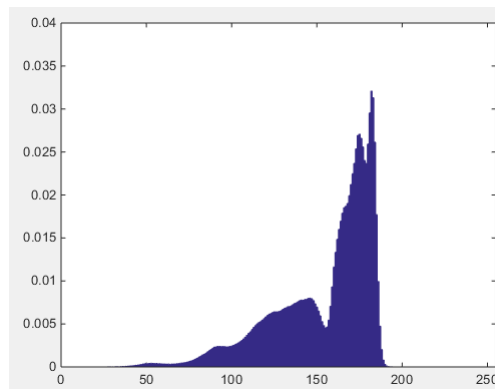
43

- A **Função Distribuição de Probabilidades** associa as probabilidades a cada valor individual de X , $f(x_i)$
 - Função Massa de Probabilidade (pmf), para variáveis aleatórias discretas
 - Função Densidade de Probabilidade (pdf), para variáveis aleatórias contínuas

Função Distribuição de Probabilidades

44

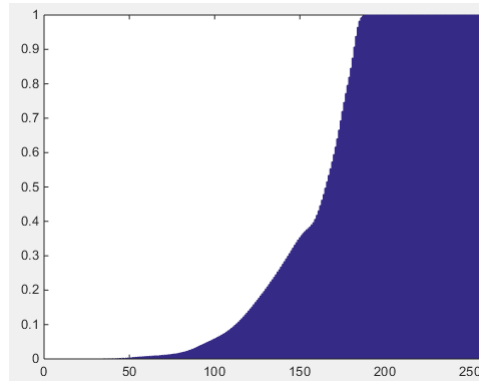
- Podemos construir um gráfico que relaciona o valor da variável X a sua probabilidades



Função Distribuição de Probabilidades

45

- Podemos construir também um gráfico que relaciona a probabilidade acumulada os valores de X



Modelos de Probabilidade

46

- Um modelo de distribuição de probabilidade atribui uma probabilidade para cada um dos possíveis resultados de uma experiência aleatória
 - ▣ Existem vários modelos, cada um adequado a um tipo de variável aleatória e a natureza dos dados

Modelos de Probabilidade para variáveis discretas

47

- Distribuição Uniforme Discreta
 - Seja uma variável aleatória X
 - X assume os valores $\{x_1, x_2, \dots, x_n\}$
 - Diz-se que X tem distribuição uniforme discreta se, e somente se
 - $P(X = x_i) = 1 / k$
 - Para todo $i = 1, 2, \dots, k$
 - Exemplo: lançamento de um dado

Modelos de Probabilidade para variáveis discretas

48

- Distribuição de Bernoulli
 - Seja uma variável aleatória X onde apenas dois resultados são possíveis
 - x : sucesso (1) ou insucesso (0).
 - Queremos observar sucessivos eventos, sendo que
 - A probabilidade de “sucesso” p é constante ao longo do experimento
 - Cada evento é independente.

Modelos de Probabilidade para variáveis discretas

49

- Distribuição de Bernoulli
 - ▣ O modelo de probabilidade será dado por
 - $P(X = x) = p^x (1 - p)^{1-x}$
 - ▣ Exemplo: uma urna tem 30 bolas brancas e 20 verdes
 - Brancas: 0
 - Verdes: 1
 - ▣ $P(X = x) = (30/50)^x (20/50)^{1-x}$

Modelos de Probabilidade para variáveis discretas

50

- Distribuição Binomial
 - ▣ Dada uma distribuição de Bernoulli, essa distribuição nos indica a probabilidade do número de “sucessos” numa sequência de n tentativas independentes
 - ▣ O modelo de probabilidade é dado por
 - $P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$

Modelos de Probabilidade para variáveis discretas

51

- Distribuição Binomial
 - ▣ Um sistema de segurança possui 4 alarmes com probabilidade de sucesso $p = 0,8$ cada
 - ▣ Qual a probabilidade de 3 alarmes soarem ao mesmo tempo?
 - $P(3) = 4 * (0,8)^3 * (1 - 0,8)^1 = 0,4096$

Modelos de Probabilidade para variáveis discretas

52

- Distribuição Geométrica
 - ▣ Dada uma distribuição de Bernoulli, essa distribuição nos indica o número de observações x até a ocorrência de um “sucesso”
 - ▣ O modelo de probabilidade é dado por
 - $P(X = x) = p(1 - p)^{x-1}$

Modelos de Probabilidade para variáveis discretas

53

- Distribuição Geométrica
 - ▣ Numa fábrica de peças, a proporção de defeitos é de 8%. Ao se inspecionar um lote, qual a probabilidade de se encontrar um defeito na quarta peça analisada?
 - $P(4) = 0,08 * (0,92)^3 = 0,0623$

Modelos de Probabilidade para variáveis contínuas

54

- Distribuição Normal ou Gaussiana
 - ▣ A distribuição normal é uma das mais utilizadas na estatística
 - ▣ Ela modela muitos acontecimentos da natureza
 - Fenômenos físicos e financeiros
 - Características morfológicas e sociológicas de uma determinada população.

Modelos de Probabilidade para variáveis contínuas

55

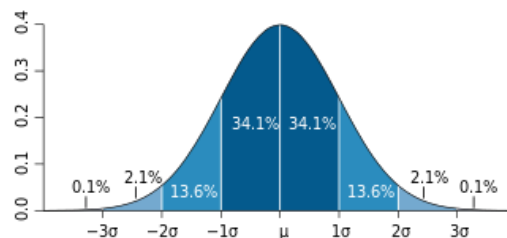
- Distribuição Normal ou Gaussiana
 - ▣ Sua função é inteiramente descrita por seus parâmetros de média e desvio padrão
 - ▣ Conhecendo-se estes consegue-se determinar qualquer probabilidade em uma distribuição Normal.

$$\square f(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}$$

Modelos de Probabilidade para variáveis contínuas

56

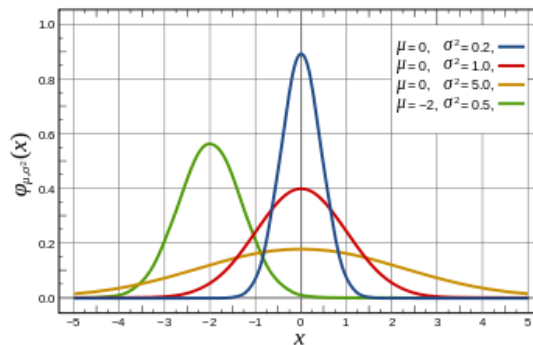
- Distribuição Normal ou Gaussiana
 - ▣ Seu gráfico é conhecido como curva de Gauss
 - ▣ O desvio padrão define a área sob a curva, e para cada valor de desvio padrão corresponde uma proporção de casos da população
 - ▣ $P(x_1 \leq x \leq x_2) = \int_{x_1}^{x_2} f(x, \mu, \sigma)$



Modelos de Probabilidade para variáveis contínuas

57

- Na curva de Gauss, a média refere-se ao centro da distribuição e o desvio padrão ao espalhamento (ou achatamento) da curva



Modelos de Probabilidade para variáveis contínuas

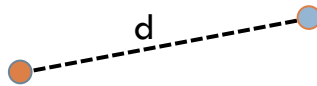
58

- Apesar da distribuição Normal ser a mais importante, existe outras que não serão aqui estudadas
 - ▣ Distribuição t de Student
 - ▣ Distribuição Exponencial
 - ▣ Distribuição Gamma
 - ▣ Distribuição de Poisson

Medidas de Distância

59

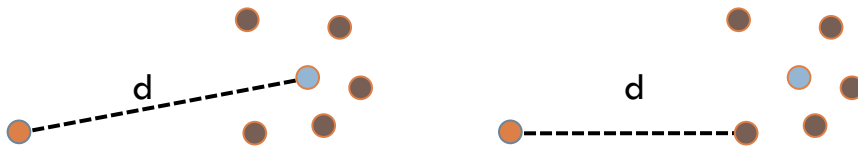
- Entende-se por distância a medida da separação de 2 objetos
 - ▣ comprimento do segmento de reta que os liga
- Em reconhecimento de padrões, a distância indica a dissimilaridade ou afastamento entre dois atributos ou vetores de atributos.



Medidas de Distância

60

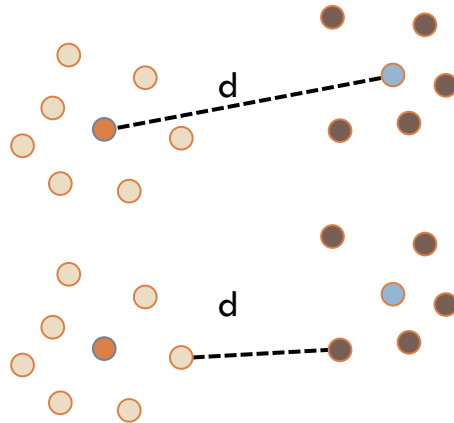
- Uma medida de distância também pode ser utilizada para indicar a dissimilaridade ou afastamento entre um vetor de atributos e uma classe (centroide ou elemento mais próximo)



Medidas de Distância

61

- Ou entre duas classes distintas de padrões
 - ▣ (centroide ou elementos mais próximos)



Métrica

62

- A métrica é uma formalização do conceito de distância.
 - ▣ Um espaço onde exista uma métrica definida é chamado de espaço métrico.
- Para uma função ser considerada uma distância, ou métrica, entre dois vetores de atributos, ela deve seguir alguns axiomas (consensos iniciais)

Métrica

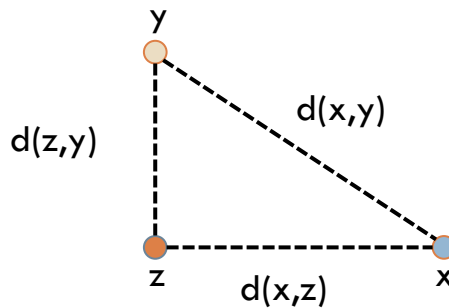
63

- Os axiomas ou propriedades que definem a métrica são 3
 - $d(x,y) = d(y,x)$, simetria
 - $d(x,y) \geq 0$
 - $d(x,x) = 0$
- Além dessas 3 propriedades, também valem
 - $d(x,y) = 0$, se e somente se $x = y$
 - $d(x,y) \leq d(x,z) + d(z,y)$, também conhecida como desigualdade do triângulo

Métrica

64

- Desigualdade do triângulo (ou triangular)



Distância de Minkowski de ordem s

65

- Trata-se de uma métrica para o espaço Euclidiano e que serve de generalização para outras distâncias, como a “city block” ($s = 1$) e a própria distância Euclidiana ($s = 2$)
- Dado dois vetores X e Y , a mesma é definida como sendo

$$d(X, Y) = \left(\sum_{i=1}^p |X_i - Y_i|^s \right)^{\frac{1}{s}} = \sqrt[s]{|X_1 - Y_1|^s + |X_2 - Y_2|^s + \cdots + |X_p - Y_p|^s}$$

Distância máxima, “city block” ou Manhattan

66

- Dado dois vetores X e Y , esta métrica é definida como o somatória dos módulos das diferenças, e possui a seguinte fórmula

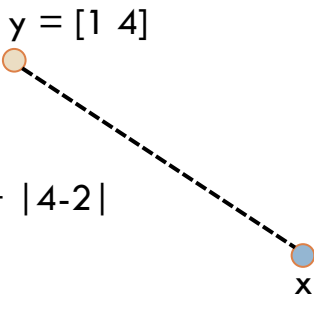
$$d(X, Y) = \left(\sum_{i=1}^p |X_i - Y_i|^1 \right)^1 = |X_1 - Y_1| + |X_2 - Y_2| + \cdots + |X_p - Y_p|$$

- Trata-se de uma distância que depende da rotação do sistema de coordenadas, mas não depende de sua reflexão em torno de um eixo ou suas translações

Distância máxima, “city block” ou Manhattan

67

- Exemplo: distância “city block” ou Manhattan



$$y = [1 \ 4]$$

$$d(x,y) = |1-5| + |4-2|$$

$$d(x,y) = 4 + 2$$

$$d(x,y) = 6$$

$$x = [5 \ 2]$$

Distância Euclidiana

68

- Trata-se da distância mais comum entre dois pontos
 - ▣ Aquela distância medida com uma régua
- Dado dois vetores X e Y , a mesma é definida como sendo

$$d(X, Y) = \left(\sum_{i=1}^p |X_i - Y_i|^2 \right)^{\frac{1}{2}} =$$

$$\sqrt{|X_1 - Y_1|^2 + |X_2 - Y_2|^2 + \dots + |X_p - Y_p|^2}$$

Distância Euclidiana

69

- Exemplo: distância Euclidiana

$$y = [1 \ 4]$$

$$x = [5 \ 2]$$

$$d(x,y) = \sqrt{(1-5)^2 + (4-2)^2}$$

$$d(x,y) = \sqrt{4^2 + 2^2}$$

$$d(x,y) = \sqrt{16 + 4}$$

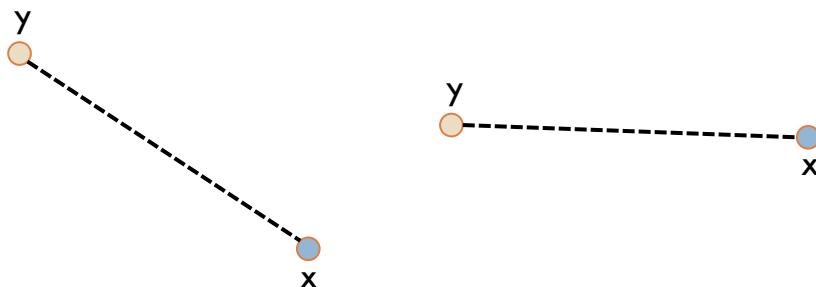
$$d(x,y) = \sqrt{20}$$

$$d(x,y) = 4,4721$$

Distância Euclidiana

70

- Trata-se de uma distância que é invariante a
 - ▣ rotação do sistema de coordenadas
 - ▣ a sua reflexão em torno de um eixo
 - ▣ translações



Distância Chebyshev ou chessboard

71

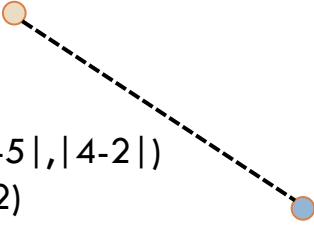
- A distância de Chebyshev se assemelha muito a city block. No caso, essa métrica considera o valor máximo dos módulos das diferenças dos pontos em respectivas posições
- Assim, dado dois vetores X e Y , a mesma é definida como sendo

$$d(X, Y) = \max(|X_1 - Y_1|, |X_2 - Y_2|, \dots, |X_p - Y_p|)$$

Distância Chebyshev ou chessboard

72

- Exemplo: distância Chebyshev ou chessboard

$$y = [1 \ 4]$$


$$d(x, y) = \max(|1 - 5|, |4 - 2|)$$

$$d(x, y) = \max(4, 2)$$

$$d(x, y) = 4$$

$$x = [5 \ 2]$$

Comparação entre as métricas

73

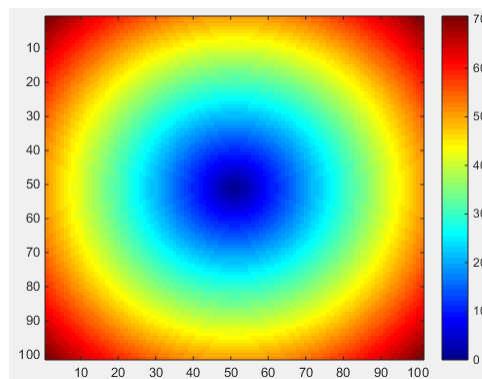
- Considere a matriz abaixo. Qual a distância entre o elemento central e os demais em cada métrica

0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0

Comparação entre as métricas

74

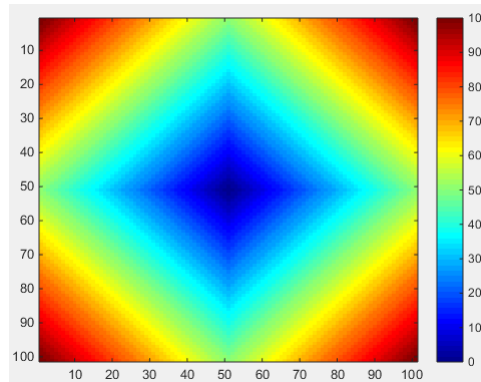
- Distância Euclidiana



Comparação entre as métricas

75

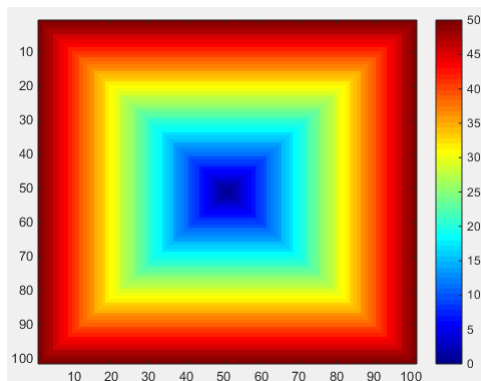
- Distância “city block” ou Manhattan



Comparação entre as métricas

76

- Distância Chebyshev ou chessboard



Distância de Mahalanobis

77

- Se baseia nas correlações entre atributos com as quais distintos padrões podem ser identificados e analisados
 - ▣ Leva em consideração as variações estatísticas dos atributos
 - ▣ Introduzida pelo matemático indiano Prasanta Chandra Mahalanobis em 1936.

Distância de Mahalanobis

78

- Dado dois vetores X e Y , e a matriz de covariâncias Σ , a distância de Mahalanobis é definida como sendo

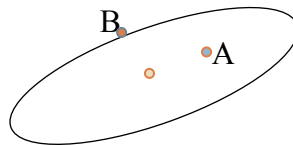
$$d(X, Y) = [(X - Y)^T \Sigma^{-1} (X - Y)]^{\frac{1}{2}}$$

- Se a matriz de covariâncias for uma matriz identidade, essa distância é igual a distância Euclidiana

Distância de Mahalanobis

79

- Os pontos A e B possuem as mesmas distâncias euclidianas da média (centro da elipse).
- No entanto, o ponto B é claramente "mais diferente" da população do que o ponto A.



Distância de Mahalanobis

80

- A distância de Mahalanobis leva em consideração a variância de cada atributo, assim como a covariância entre eles
 - ▣ Transforma os dados em dados normalizados não correlacionadas e calcula a distância euclidiana para os dados transformados
 - ▣ É invariante à escala (não depende da escala das medições)
 - ▣ Similar ao z-score

Distância Quadrática

81

- Basicamente, a distância quadrática é uma generalização da distância de Mahalanobis
 - ▣ Também leva consideração a relação entre os atributos

$$d(X, Y) = [(X - Y)^T A (X - Y)]^{\frac{1}{2}}$$

- No entanto, no lugar da matriz de covariâncias Σ , ela utiliza uma matriz A
 - ▣ A deve ser simétrica positiva definida
 - ▣ Isso significa que A é uma matriz válida $d(X, Y) \geq 0$

Distância Quadrática

82

- De modo geral, a matriz A deverá ser calculada de acordo com o problema.
- Por exemplo, na
 - ▣ distância de Mahalanobis: A é a matriz inversa da matriz de covariâncias
 - ▣ distância Euclidiana: A é a matriz identidade

Padronização dos dados

83

- Consiste do processo de conversão de um escore bruto de uma distribuição em um escore z
- Por escore bruto entende-se o valor individual observado na medição de um determinado atributo
- Isso nos ajuda a entender onde um determinado escore se localiza em relação aos outros em uma distribuição
 - ▣ Escore padrão, escore z ou z -score

Z-score

84

- Também conhecido como escore padronizado, ele indica o quanto acima ou abaixo da média um determinado escore está em termos de unidades padronizadas de desvio
 - ▣ Confuso??

Z-score

85

- Traduzindo: o score-z indica em quantas unidades de desvios padrão uma observação ou dado está acima ou abaixo da média
- Calculado a partir da média e desvio padrão dos dados

$$z = \frac{X - \mu}{\sigma}$$

Z-score

86

- O score z permite que se compare duas amostras obtidas em escalas diferentes de mensuração
- Isso ocorre por que ele retorna uma versão dos dados centralizada e com a escala ajustada
 - Centralizada: amostras com média 0
 - Ajuste de escala: amostras com desvio padrão 1

Z-score

87

Exemplo:

Dados brutos

Média Altura	Média Peso
1,73	64,22

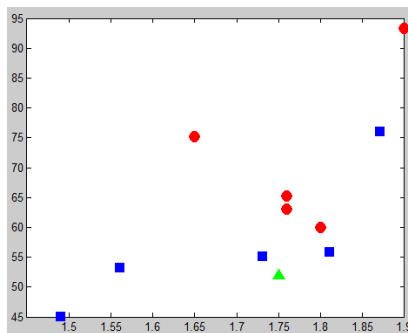
Desvio Altura	Desvio Peso
0,13	14,01

Altura	Peso	Sexo
1,87	76,1	0
1,65	75,2	1
1,80	60,0	1
1,81	55,9	0
1,90	93,3	1
1,74	65,2	1
1,49	45,1	0
1,56	53,2	0
1,73	55,1	0
1,76	63,1	1

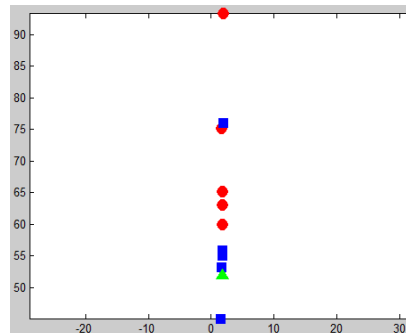
Z-score

88

Com ajuste de escala



Sem ajuste de escala



Z-score

89

Exemplo

Dados normalizados

Média Altura	Média Peso
0,00	0,00

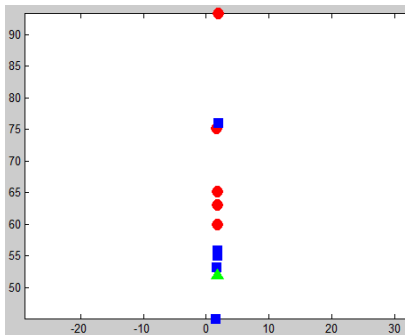
Desvio Altura	Desvio Peso
1,00	1,00

Altura	Peso	Sexo
1,06	0,84	0
-0,62	0,78	1
0,53	-0,30	1
0,60	-0,59	0
1,29	2,07	1
0,07	0,07	1
-1,84	-1,36	0
-1,31	-0,79	0
-0,01	-0,65	0
0,22	-0,08	1

Z-score

90

Dados brutos (sem ajuste de escala)



zscore

