

AGRUPAMENTO (“CLUSTERING”)

- Objectivo genérico: dado um conjunto de instâncias de treino, *sem informação fornecida sobre a classe ou categoria a que pertencem*, determinar um conjunto de classes que permita organizar de forma conveniente essas instâncias.

“... cluster analysis is the art of finding groups in data”

(Kaufmann and Rousseeuw, 1990)

- Forma de aprendizagem não supervisionada
- Partição do conjunto de instâncias num pequeno número de subconjuntos (grupos ou “clusters”), de modo a que instâncias similares pertençam ao mesmo grupo e instâncias muito distintas pertençam a grupos diferentes.

Abordagens

- Abordagens por optimização iterativa

Encontrar uma partição inicial razoável e ajustá-la progressivamente, movendo instâncias de um grupo para outro se isso melhorar a qualidade da partição. Normalmente requerem que se indique o número de grupos a considerar. Tentam obter a partição que optimiza uma certa função de avaliação.

- Abordagens hierárquicas:

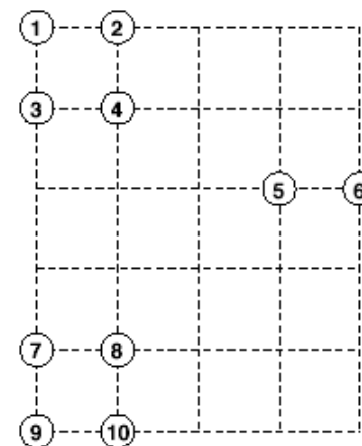
Constroem árvores de grupos ou partições (um nível da árvore indica uma partição possível). Cada grupo subdivide-se em vários subgrupos e assim sucessivamente. Após construída, a árvore de grupos pode ser analisada por níveis para determinar a partição que pareça melhor/mais útil.

- Métodos aglomerativos : construção ascendente da árvore
- Métodos divisivos: construção descendente da árvore

“K-means clustering”

- Método mais popular de agrupamento por optimização iterativa
- Requer que o utilizador indique o número k de grupos a considerar
- Método genérico:
 - Inicializar as “sementes” ou “centros iniciais” de cada grupo, r_j , $j=1,\dots,k$. Por exemplo, escolhem-se aleatoriamente k instâncias para centros iniciais.
 - para cada instância x_i :
 - calcular as distâncias euclidianas entre x_i e o centro de cada grupo r_j , d_{ij} , $j=1,\dots,k$
 - afectar x_i ao grupo cujo centro está a menor distância
 - recalcular o centro dos grupos r_j , $j=1,\dots,k$ (vectores-média das instâncias no grupo).
 - repetir passos 2-3 até que os novos centros sejam idênticos aos anteriores (ou, equivalentemente, que não haja alteração do valor da função de avaliação da partição).

Exemplo



Considere $K = 3$

Quais os grupos que se obtêm quando os centros iniciais escolhidos são:

- a) 1, 6 e 9
- b) 4, 9 e 10

Características do algoritmo k-means

- Complexidade do algoritmo “k-means”:
 - $O(knI)$ sendo I – nº de iterações, n – dimensão do conjunto de treino
- Pode não convergir para a melhor partição!
 - A escolha inicial de pontos pode determinar a obtenção de diferentes grupos; Os grupos finais representam apenas um óptimo local da função de avaliação;
 - Pode tornar-se este problema:
 - Fazendo múltiplas procura partindo de escolhas aleatórias distintas para os centros iniciais
 - Escolhendo criteriosamente os centros iniciais: 1º centro – aleatório (ou o ponto mais próximo do centro de todos os dados), 2º centro - à maior distância do 1º centro escolhido, 3º centro - que maximiza a distância ao centro já escolhido mais próximo, ...
 - Aplicando, para obter os próprios centros iniciais, outro método de agrupamento

Algumas funções de avaliação para agrupamento

Variação intra-grupos (“within cluster variation”):

- Procura-se *minimizar* a variação ou diferença dentro do mesmo grupo.
- O valor da variação para toda a partição é a soma de todas as variações intra-grupo

$$wc(\bigcup_j C_j) = \sum_{j=1}^k wc(C_j)$$

- **CrITÉrio da soma do quadrado dos erros** (“sum-of-squared errors”):
 - Este é o critério de avaliação intra-grupo mais frequentemente utilizado, quando se definem pontos centrais r_j para cada grupo C_j .

$$wc(C_j) = \sum_{x_i \in C_j} d(x_i, r_j)^2$$

- Pretende-se portanto escolher a partição que minimize $wc(\cup C)$
- Este critério é apropriado quando os grupos são relativamente compactos e separados uns dos outros.

– **CrITÉrio intra-grupos dos vizinhos mais próximos:**

- Uma outra possível forma de avaliar a variação intra-grupos, nomeadamente quando não se definem centros dos grupos
- Para cada grupo, calcula o máximo das distâncias entre um ponto e o seu vizinho mais próximo

$$wc(C_j) = \max_{x_i \in C_j} \min_{y \in C_j} \{d(x_i, y) \mid x_i \neq y\}$$

- A avaliação de uma partição pode depender de critérios de variação intra-grupos ou inter-grupos ou ambos combinados, mas os critérios simples são mais frequentemente utilizados.

K-means e agrupamento por minimização do quadrado dos erros

- O algoritmo k-means é basicamente um algoritmo de procura por “hill-climbing” da partição que minimiza o critério da soma do quadrado dos erros.
- O k-means espera até afectar todos os exemplos a grupos, antes de recalculer os centros dos grupos.
- Existem variantes sequenciais do k-means, que actualizam o valor da função de avaliação e o centro de cada grupo após afectar cada exemplo, e que podem ser utilizadas em problemas em que os dados são obtidos incrementalmente.

Métodos hierárquicos

- Produzem uma partição dos dados em grupos, que por sua vez podem ser divididos em subgrupos, ...
- Constroem uma sequência de partições. Cada partição corresponde a um nível distinto, numa árvore de grupos.
- A árvore de grupos pode ser representada com informação quantitativa sobre a semelhança entre grupos, por exemplo através de estruturas chamadas dendrogramas, ou sem informação quantitativa, por exemplo através de conjuntos de conjuntos, ...

Método hierárquico aglomerativo

- As instâncias são agrupadas em grupos gradualmente mais vastos, começando por grupos só com uma instância, até se obter um grupo contendo todas as instâncias (raiz da hierarquia)
- Métodos normalmente não incrementais.
- Pressupõe-se a capacidade para calcular distâncias entre duas instâncias (frequentemente distância euclidiana) e distâncias entre dois grupos (como calcular ?)
- Os dois grupos a menor distância são seleccionados, constituindo-se um novo grupo resultante da fusão dos dois
 - O uso de diferentes medidas de distância pode produzir hierarquias e grupos muito diferentes, a partir do mesmo conjunto de dados iniciais

Esboço do método

- **Entrada:**
 - conjunto de nós iniciais (contendo uma única instância): C
 - Matriz indicando a distância entre cada par de nós: MatDist
- **Proc($C, \text{MatDist}$)**

Sejam N_i e N_j elementos de C tais que $\text{MatDist}(N_i, N_j)$ é mínimo

Criar um novo nó N , pai de N_i e N_j e gerar a descrição associada a N

Seja $C' = C \setminus \{N_i, N_j\}$

Se $C' = \emptyset$

então retornar a árvore com raiz N




senão

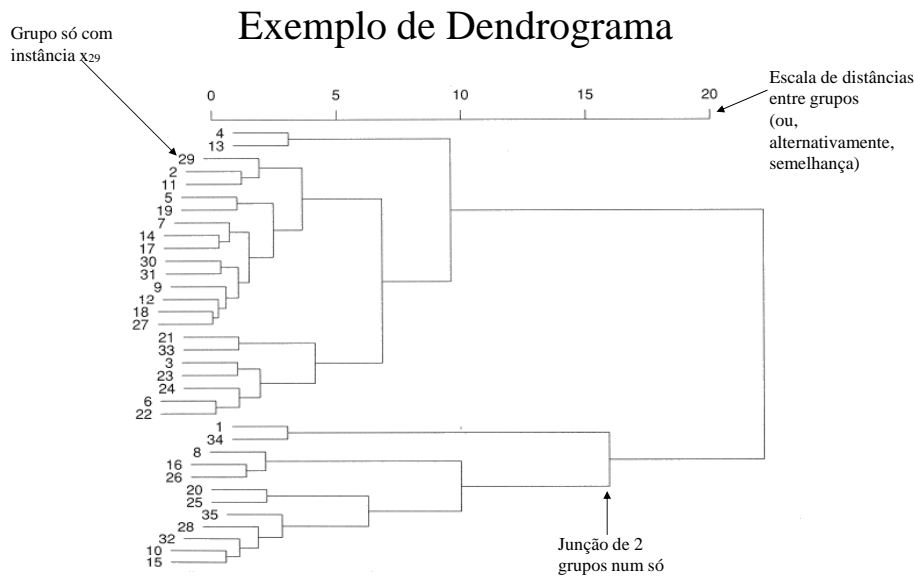
seja $C'' = C' \cup \{N\}$

calcular nova MatDist relativa a C''

proc($C'', \text{MatDist}''$)

Abordagens à definição de distância entre grupos

- **Vizinhos mais próximos:** Distância mínima entre 2 pontos, um de cada grupo
 
 - Se usada como critério a minimizar para junção de grupos, dá origem a grupos mais dispersos (sobre-generalização)
- **Vizinhos mais longínquos:** Distância máxima entre 2 pontos, um de cada grupo
 
 - Se usada como critério a minimizar para junção de grupos, dá origem a grupos mais compactos (sub-generalização)
- **Centróide:** Distância entre os centros dos 2 grupos
 
 - Potencial má representação de grupos com formas irregulares / não simétricas



Graça Gaspar - DI/FCUL

Agrupamento - 13

Dendrogramas

- A escala de variação pode ser utilizada para determinar que partição de grupos é “mais natural”, isto é, qual o nível da árvore de grupos que parece dividir melhor os dados:
 - Se os valores para as várias barras do dendrograma (pontos de junção de grupos) estão uniformemente distribuídos, não há argumento para considerar um certo número de grupos como mais natural que outro.
 - Se, em contrapartida, há um aumento significativo da distância entre os valores correspondentes a duas fusões consecutivas no dendrograma, pode-se argumentar que a última dessas fusões irá piorar significativamente a qualidade do agrupamento e seria preferível não a fazer.
- Por exemplo, se a distância no dendrograma entre $k=2$ ou $k=3$ grupos é nitidamente superior à distância entre outros valores consecutivos de k , pode-se argumentar que é mais natural considerar 3 grupos.

Graça Gaspar - DI/FCUL

Agrupamento - 14

Varição inter-grupos (“between cluster variation”):

- Procura-se *maximizar* a variação entre pares de grupos distintos.
- A função de avaliação da partição é a soma das variações inter-grupos, entre pares de grupos distintos

$$bc(\bigcup C) = \sum_{1 \leq i < j \leq k} bc(C_i, C_j)$$

CrITÉrio inter-grupos da distância entre pontos centrais:

- Supondo que se definem pontos centrais r_j

$$bc(C_j, C_i) = d(r_j, r_i)$$

CrITÉrio inter-grupos da média das distâncias:

$$bc(C_j, C_i) = \frac{1}{|C_j||C_i|} \sum_{x \in C_j} \sum_{x' \in C_i} d(x, x')$$

Graça Gaspar - DI/FCUL

Agrupamento - 15

CrITÉrio inter-grupos dos vizinhos mais próximos:

$$bc(C_j, C_i) = \min_{x \in C_j, x' \in C_i} d(x, x')$$

- Sensível a pequenas perturbações nos dados.
- Único critério inter-grupos tal que, no caso de pares de grupos equidistantes, torna irrelevante a ordem por que são considerados para aglomeração.
- O algoritmo hierárquico aglomerativo que usa este critério inter-grupos, e que pára quando a distância entre grupos mais próximos é superior a um dado limiar, é chamado **algoritmo de agrupamento de ligação simples** (“single link algorithm”).

CrITÉrio inter-grupos dos vizinhos mais longínquos:

$$bc(C_j, C_i) = \max_{x \in C_j, x' \in C_i} d(x, x')$$

- O algoritmo hierárquico aglomerativo que usa este critério inter-grupos e que pára quando a distância entre grupos mais próximos é superior a um dado limiar é chamado **algoritmo de agrupamento dos vizinhos mais longínquos** (“furthest-neighbor algorithm”) ou **de ligação completa** (“complete link algorithm”).

Graça Gaspar - DI/FCUL

Agrupamento - 16

- Critério de avaliação de partições combinado:

- Combinação monótona da avaliação intra-grupos (*a minimizar*), $wc(\cup C)$, com a avaliação inter-grupos (*a maximizar*), $bc(\cup C)$.
- Por exemplo, escolher a partição, representada por $\cup C$, que minimiza o valor de

$$\frac{wc(\cup C)}{bc(\cup C)}$$

Método da silhueta (*silhouette*)

- Método que permite uma avaliação e interpretação gráfica da qualidade de uma partição

Sendo:

- i , uma instância do conjunto de dados
 - $g(i)$, o grupo a que i pertence
 - $m_g(i)$, a média das distâncias entre i e as instâncias do grupo g
1. Calcular $m_{g(i)}(i)$
 2. Para cada grupo $g' \neq g(i)$, calcular $m_{g'}(i)$ e determinar o grupo $g_{vizinho}(i)$ para o qual esse valor é mínimo
 3.
$$silhueta(i) = \frac{m_{g_{vizinho}(i)}(i) - m_{g(i)}(i)}{\max\{m_{g_{vizinho}(i)}(i), m_{g(i)}(i)\}}$$

$$-1 \leq silhueta(i) \leq 1$$

Método da silhueta (*silhouette*)

- Interpretação:

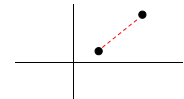
- $silhueta(i) \approx 1$: i está no grupo adequado
- $silhueta(i) \approx -1$: i estaria melhor no grupo vizinho
- $silhueta(i) \approx 0$: i está na fronteira entre o seu grupo e o vizinho
- A média de $silhueta(i)$ para todas as instâncias num grupo g é uma medida da qualidade geral desse grupo g
- A média de $silhueta(i)$ para todas as instâncias no conjunto de dados é uma medida da qualidade geral da partição obtida
- Se há demasiados ou poucos grupos na partição (má escolha de k), tipicamente alguns desses grupos terão valores de silhueta muito inferiores aos restantes
- Gráficos de valores de silhueta e a média global das silhetas podem ajudar a determinar o número adequado de grupos.

Medidas de distância entre instâncias

- Funções de distância ou métricas comuns, entre instâncias caracterizadas por d atributos *numéricos* ou dimensões:

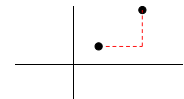
- Distância euclidiana:

$$D(a, b) = \sqrt{\sum_{i=1}^d (a_i - b_i)^2}$$



- Distância de Manhattan ou “city-block” ou de Hamming+:

$$D(a, b) = \sum_{i=1}^d |a_i - b_i|$$



- Família de distâncias de Minkowski, ou norma L_k :

$$D(a, b) = \left(\sum_{i=1}^d |a_i - b_i|^k \right)^{1/k}$$

A distância euclidiana corresponde a L_2 ;
A distância de Manhattan corresponde a L_1 .

Nota: Mudanças de escala das coordenadas podem afectar significativamente a distância;
Em caso de grandes disparidades dos intervalos de valores nas várias dimensões, é prática frequente normalizar todos os dados para uniformizar esses intervalos.

Medidas de distância entre instâncias

- Funções de distância ou métricas, entre instâncias caracterizadas por d atributos *nominativos*:

- Distância de Hamming:

$$D(a, b) = \#\{i: a_i \neq b_i\}$$

- Funções de distância ou métricas, entre instâncias correspondentes a *conjuntos de elementos* (ou caracterizadas por um número de atributos binários variável):

- Distância de Tanimoto:

$$D(A, B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} = \frac{|A| + |B| - 2|A \cap B|}{|A| + |B| - |A \cap B|}$$

Um elemento do conjunto (instância) pode também ser interpretado como indicando a validade, nesse conjunto, do atributo booleano que representa.

Ex: instâncias representando o conjunto de cadeiras em que um aluno se inscreveu.

Medidas de semelhança e dissemelhança

- O termo distância tem sido aqui usado informalmente para designar uma medida de dissemelhança.
- Uma medida de dissemelhança pode sempre ser facilmente transformada numa medida de semelhança, ou vice-versa, aplicando uma transformação adequada monótona decrescente.

Exemplos:

$$d(a, b) = 1 - s(a, b)$$

$$d(a, b) = (2^*(1 - s(a, b)))^{1/2}$$

- Uma métrica é uma medida de dissemelhança que satisfaz:

- $d(a, b) \geq 0$ e $d(a, b) = 0$ sse $a = b$
- $d(a, b) = d(b, a)$
- $d(a, b) \leq d(a, c) + d(c, b)$

Lidando com variáveis não comensuráveis

Exemplo:

< Peso , Altura >:

Devemos usar como unidades < kg, m> ou < kg, cm> ?

Decisão arbitrária cuja alteração pode, no entanto, afectar grandemente a importância da variável para o valor da distância obtida

Para ultrapassar os efeitos da arbitrariedade da escolha das unidades, uma técnica habitualmente usada consiste em normalizar as variáveis dividindo-as pelo seu desvio padrão

$$x'_i = x_i / \sigma_x$$

Desta forma elimina-se o efeito de escala, capturado pelo desvio padrão

Lidando com a dependência (linear) dos atributos

Exemplo:

< DiâmetroCintura, RaioCintura, Altura >:

As 2 primeiras variáveis, por estarem altamente correlacionadas, podem dominar o cálculo da distância

- Para eliminar os efeitos do uso de variáveis redundantes, uma técnica habitualmente usada consiste na normalização das variáveis não apenas numa direcção mas também considerando as covariâncias entre variáveis

$$\text{cov}(x, y) = \frac{1}{n} \sum_{i=1}^n (x(i) - \mu_x)(y(i) - \mu_y)$$

ou os coeficientes de correlação entre variáveis

$$\rho(x, y) = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$$

Distância de Mahalanobis

- Medida de distância entre instâncias, normalizada por incorporação da matriz de covariâncias

Considerando d atributos numéricos

e duas instâncias $a = \langle a_1, \dots, a_d \rangle$ e $b = \langle b_1, \dots, b_d \rangle$

e a matriz de covariâncias S entre os d atributos


$$d(a, b) = \sqrt{(a - b)^T S^{-1} (a - b)}$$

- Se a matriz de covariâncias for a matriz identidade corresponde à distância euclideana
- Se a matriz de covariâncias for uma matriz diagonal corresponde à distância euclideana normalizada

Exemplo de aplicação do método aglomerativo

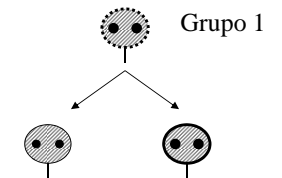
- Considerando instâncias de treino:

- Descrições de células usando os atributos nominais: nucleídeos, caudas, cor, espessura da membrana



	0			
	1	0		
	3	4	0	
	4	3	1	0

Distância entre instâncias:
distância de Hamming

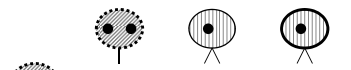


Graça Gaspar - DI/FCUL

Agrupamento - 26

Exemplo (cont.)

- Nova matriz de distâncias:



	0		
	3.5	0	
	3.5	1	0

- Descrições probabilísticas dos grupos:

- Grupo 1:

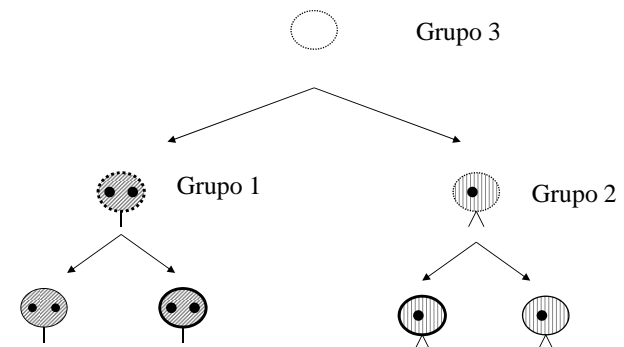
Caudas	Uma	1	Dois	0
Nucleídeos	Um	0	Dois	1
Cor	Clara	0	Escura	1
Membrana	Fina	1/2	Espessa	1/2

- Distância entre grupos:

- Média das distâncias entre instâncias

Exemplo (cont.)

- Hierarquia de grupos produzida:



Agrupamento probabilístico baseado em misturas (de modelos)

- Métodos de agrupamento que assumem um modelo probabilístico para cada grupo (*cluster*) e procuram determinar os valores dos parâmetros desses modelos
- Tipicamente assume-se o mesmo tipo de modelo para todos os grupos, mas com parâmetros possivelmente diferentes
Por exemplo: distribuição normal multivariada
- Mistura:
 - Uma mistura é um conjunto de k distribuições de probabilidade, representando k grupos, que governam os valores das instâncias que são membros desses grupos. Adicionalmente, os grupos podem não ser igualmente prováveis: há uma distribuição de probabilidade que reflecte as suas populações relativas.
 - Num caso básico, considerando instâncias com um único atributo, podemos por exemplo assumir um modelo de mistura de distribuições normais, no qual todas as instâncias seguem uma distribuição normal, mas com valor médio e variância diferentes, consoante o grupo a que pertencem.

Agrupamento baseado em probabilidades e o algoritmo EM: Motivação

- Objectivo, para caso de mistura de duas normais:

Considerando duas classes c_1 e c_2 , isto é $k=2$,
pretende-se determinar os parâmetros da mistura de densidades normais:

$$\theta = (\mu_{c_1}, \mu_{c_2}, \sigma_{c_1}, \sigma_{c_2}, p_{c_1}) \quad (\text{com } p_{c_1} + p_{c_2} = 1)$$

tais que tenham verosimilhança máxima, isto é, tais que
para os dados de treino D

$$\theta = \arg\max_{\theta} p(D | \theta) \quad (\text{i.e. } \theta = \text{hipótese de máxima verosimilhança})$$

Se fosse conhecido o grupo (classe) de cada instância de treino, seria fácil
estimar estes parâmetros;

Se fossem conhecidos estes parâmetros, seria fácil determinar a
probabilidade de uma instância pertencer a um grupo, $p(c_i | x_j, \theta)$.

Agrupamento baseado em probabilidades e o algoritmo EM: Ideia base

- Abordagem iterativa: partindo de estimativas iniciais dos parâmetros, calcular as probabilidades de cada classe, usá-las para re-estimar os parâmetros, ...
 - Algoritmo EM:
Ciclo com dois passos:
 - Expectation**: cálculo do valor esperado para as classes das instâncias, dadas as estimativas correntes dos parâmetros $p(c_i | x_j, \theta)$
 - Maximization**: cálculo dos (novos valores dos) parâmetros de modo a maximizar a sua verosimilhança, dadas as instâncias de treino
- Condição de paragem:
- Idealmente, iterar até atingir a hipótese θ de máxima verosimilhança
 - Na prática, iterar até que $\log p(D | \theta)$ sofra um incremento desprezível.

Algoritmo EM

com mistura de 2 distribuições normais, atributo único

- Passo E: cálculo das probabilidades das classes das instâncias

$$p(c_i | x_j, \theta) = \frac{p(x_j | c_i, \theta) * p(c_i | \theta)}{p(x_j | \theta)} = \frac{f(x_j; \mu_{c_i}, \sigma_{c_i}) * p_{c_i}}{p(x_j | \theta)}$$

é calculado o numerador para cada classe c_i e os valores assim obtidos são normalizados dividindo pela sua soma

- Passo M: cálculo dos parâmetros sendo $w_{ij} = p(c_i | x_j, \theta)$

$$p_{c_i} = \frac{1}{m} \sum_{j=1}^m w_{ij} \quad \mu_{c_i} = \frac{1}{m p_{c_i}} \sum_{j=1}^m w_{ij} x_j \quad \sigma_{c_i}^2 = \frac{1}{m p_{c_i}} \sum_{j=1}^m w_{ij} (x_j - \mu_{c_i})^2$$

- Função usada na condição de paragem:

$$\begin{aligned} \log p(D | \theta) &= \log \prod_j p(x_j | \theta) = \sum_j \log \left(p(x_j | c_1, \theta) p_{c_1} + p(x_j | c_2, \theta) p_{c_2} \right) \\ &\approx \sum_j \log \left(f(x_j; \mu_{c_1}, \sigma_{c_1}) * p_{c_1} + f(x_j; \mu_{c_2}, \sigma_{c_2}) * p_{c_2} \right) \end{aligned}$$

Algoritmo EM: outras misturas

- A generalização para $k > 2$ é trivial.
 - Extensão para vários atributos:
 - Assumindo a independência entre atributos, a probabilidade conjunta de uma instância é o produto das probabilidades para cada atributo.
 - Existindo correlação entre 2 atributos, pode utilizar-se uma distribuição normal bivariada (cada classe tem o seu valor médio mas em lugar das 2 variâncias é utilizada uma matriz de 2×2 covariâncias)
 - Para n atributos covariantes pode utilizar-se uma distribuição multivariada tendo como parâmetros n valores médios e uma matriz simétrica de $n \times n$ covariâncias.
 - *O aumento do número de parâmetros aumenta a possibilidade de sobre-adaptação aos dados de treino*
 - Outras distribuições frequentemente utilizadas:
 - Contagem inteiras \rightarrow distribuição de Poisson
 - Atributos sem limite superior \rightarrow distribuição log-normal
-