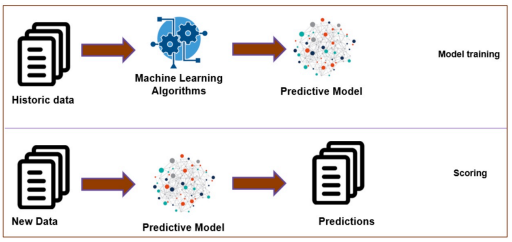
Reconhecimento de Padrões

1. Machine Learning — Aprendizagem de Máquina

É um tipo de Inteligência Artificial que consiste na execução de algoritmos que criam de modo automático modelos de representação de conhecimento a partir de um conjunto de dados.

A máquina deve ser treinada a partir dos dados históricos, considerando uma ou mais medidas de desempenho e permitindo que o algoritmo "aprenda" (i.e., ajuste, de modo iterativo, o modelo de representação do conhecimento de forma a que este melhore o seu desempenho). Após o treino, o modelo tem um potencial para efetuar previsões de qualidade em situações futuras e que estejam relacionadas com padrões históricos.



1.1. Tipos de Machine Learning

Os algoritmos podem ser classificados de acordo com a quantidade e tipo de supervisão durante a fase de treino:

I. Algoritmos Supervisionados — Criam modelos preditivos;
II. Algoritmos Não Supervisionados — Criam modelos descritivos.

O tipo de algoritmo a utilizar depende da tarefa de aprendizagem que se pretende abordar.

I. Aprendizagem Supervisionada

Este tipo de aprendizagem permite criar Modelos Preditivos. O Modelo Preditivo é usado para tarefas que envolvem a previsão de um determinado output (ou alvo) usando outras variáveis/características incluídas no conjunto de dados. O algoritmo de aprendizagem num Modelo Preditivo tenta descobrir e modelar as relações entre a variável alvo (i.e, a variável que está a ser prevista) e as outras características/variáveis (também, designadas por Variáveis Preditoras).

Exemplos de Modelação Preditiva:

- Com base nos atributos dos clientes, pretende-se prever a probabilidade do cliente abandonar nos próximos 6 meses;
- Com base nos atributos das casas, pretende-se prever o preço de venda.

Os exemplos de Modelação Preditiva apresentados, são exemplos de Aprendizagem Supervisionada. Uma vez que, os valores alvo fornecem um papel de supervisão que indica a tarefa que é preciso aprender. Especificamente, dado um conjunto de dados, o algoritmo de aprendizagem tenta otimizar uma função para determinar a combinação de valores dos atributos que corresponde ao valor estimado/ previsto que está tão próximo quanto possível do output atual alvo.

Na Aprendizagem Supervisionada, os dados de treino usados no algoritmo incluem os valores da variável alvo. Consequentemente, as soluções podem ser usadas para ajudar a supervisionar o processo de treino, de forma a determinar os parâmetros ideais do algoritmo.

A maioria dos Problemas de Aprendizagem Supervisionada pertencem a uma das seguintes categorias:

- **Problemas de Regressão** — São problemas em que o objetivo da aprendizagem supervisionada é prever um output/ alvo numérico. Ex: Prever o preço de venda de casas;
- **Problemas de Classificação** — São problemas em que o objetivo da aprendizagem supervisionada é prever um output/ alvo categórico (normalmente, binário ou multinomial). Ex: O cliente renovou o contrato? (codificação: sim/ não ou 1/ 0).
    - Quando se aplica modelos de Machine Learning a problemas de classificação, muitas vezes, em vez de se prever uma classe específica, por exemplo, "Sim" ou Não", pretende-se prever a probabilidade de uma determinada classe, por exemplo, "Sim": 0,65; Não": 0,35.
      Por defeito, a classe com a probabilidade prevista mais elevada é considerada a classe prevista. Neste caso, embora se trate de um problema de classificação é necessário prever um output numérico, a probabilidade. Apesar de tudo, devido à sua essência, o problema é considerado um problema de classificação.

II. Aprendizagem Não Supervisionada

Considerando um conjunto de ferramentas estatísticas, a Aprendizagem Não Supervisionada permite compreender e descrever os dados, mas a análise é realizada sem uma variável alvo. A Aprendizagem Não Supervisionada tem como objetivo identificar grupos num conjunto de dados.

Os grupos podem ser definidos pelas:

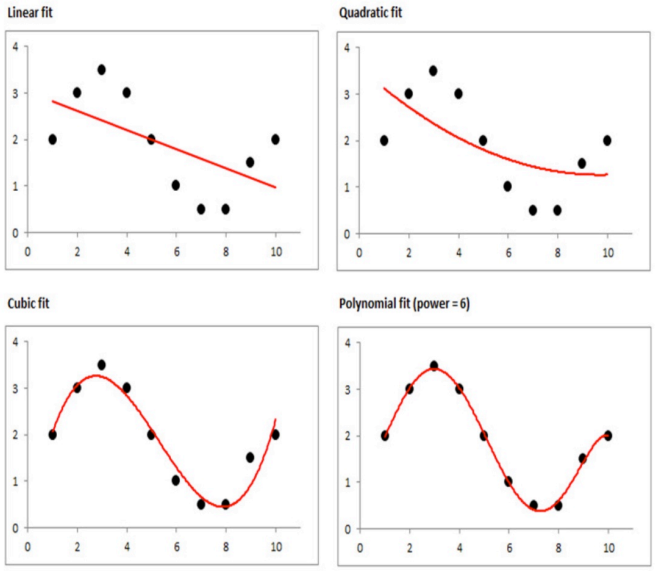
A. Linhas (ou seja, Agrupamento/ Clustering) — objetivo do Clustering é segmentar observações em grupos similares com base nas variáveis observadas;
B. Colunas (ou seja, Redução de Dimensão/ Dimension Reduction) — objetivo do Dimension Reduction é reduzir o número de variáveis no conjunto de dados.

A Aprendizagem Não Supervisionada é frequentemente realizada como parte de uma Análise Exploratória de Dados.

As técnicas de Aprendizagem Não Supervisionada são utilizadas nas organizações para:

- Dividir os consumidores em diferentes grupos homogéneos, clusters, para que sejam desenvolvidas e implementadas, em cada segmento, estratégias de marketing personalizadas.
- Identificar produtos com comportamento de compra semelhante para que os gerentes possam geri-los como grupo de produtos.

2. Ockham's Razor



Trade-off de bias-variance

A Teoria de Ockham diz que num debate entre 2 teorias, a que tem uma explicação mais simples deve ser a preferível. Tal como é feita a escolha de métodos estatísticos a partir de um Trade-off entre bias-variance. Ou seja, não deve ser demasiado complexo, deve ter um equilíbrio entre overfit e underfit.

Os gráficos anteriores, mostram um trade-off importante entre bias-variance que governa a escolha de métodos estatísticos.

- Bias: indica quão diferente é o modelo dos conjuntos de treino, do verdadeiro modelo;
- Variance: indica qual é a diferença entre os modelos baseados em conjuntos de treino.

- Modelos com poucos parâmetros não são precisos devido a um grande bias, underfit;
- Modelos com demasiados parâmetros não são precisos devido a uma grande variance, overfit.

3. Covariância e correlação

Os objetos são representados como uma nuvem de n pontos num espaço multidimensional com um eixo para cada D variáveis (há tantos eixos como variáveis). Representa-se o conjunto de dados como uma matriz com n filas e D colunas. Por exemplo:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1D} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nD} \end{bmatrix}$$

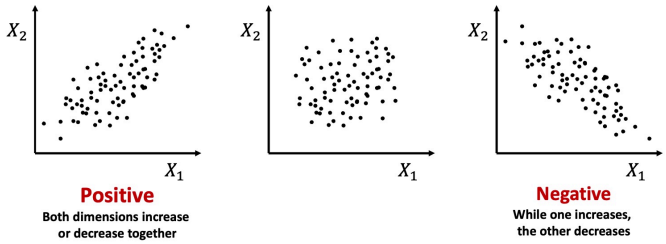
A média dos pontos é definida pela média de cada variável: $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$

A variância (apenas 1 variável) indica "o quão longe" em geral os seus valores se encontram do valor esperado (medida do desvio da média para pontos numa dimensão):

$$s_{jj'} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_{j'})(x_{ij'} - \bar{x}_{j'})$$

A covariância (2 ou mais variáveis) é a medida de quanto cada dimensão varia da média (determina se as relações são positivas ou negativas, mas não permite medir o grau de relação de variáveis): $s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$

O desvio padrão é a raiz quadrada da variância: $s_j = \sqrt{s_j^2}$



Matriz da Covariância

A matriz de covariância é uma matriz quadrada que contém as variâncias e covariâncias associadas a diversas variáveis. Os elementos diagonais da matriz contêm os desvios das variáveis, e os elementos fora da diagonal contêm as covariâncias entre todos os possíveis pares de variáveis.

- O conjunto original de X variáveis é caracterizado por DxD matriz variância-covariância, denotado por S.
- Os elementos diagonais de S, são as variâncias.
- Os elementos que não são diagonais de S, são as covariâncias.

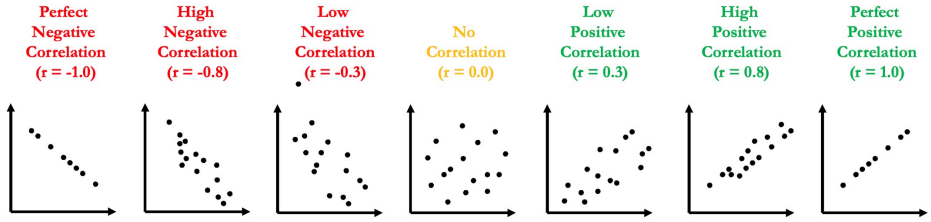
Matriz S: $\boldsymbol{S} = \begin{pmatrix} s_1^2 & \cdots & s_{1D} \\ \vdots & \ddots & \vdots \\ s_{D1} & \cdots & s_D^2 \end{pmatrix}$

É uma matriz simétrica, é uma matriz semi-definida positiva. É a soma de covariâncias entre n variáveis.

Coeficiente de correlação

O coeficiente de correlação de Pearson é dado por, $r_{jj'} = \frac{s_{jj'}}{s_j s_{j'}}$

A coeficiente de correlação de Pearson também permite identificar o grau de relação das variáveis,



Correlação e independência

Correlação significa qualquer relação estatística entre 2 variáveis, ou seja acaba por existir uma dependência. Independência quer dizer que não há relação entre variáveis, ou seja, não se pode aprender sobre uma variável apartir de outra.

4. Transformação de dados

Centralização de dados

Centralizar os dados significa subtrair a média aos valores das variáveis. Valores centrados são, $\tilde{x}_{ij} = x_{ij} - \bar{x}_j$

Nestes casos, os novos resultados têm média 0.

Os valores da variância e covariância não são afetados pelo valor da média. Subtrair a média faz com que o cálculo da variância e covariância seja mais fácil ao simplificar as suas equações.

A matriz centrada, $\tilde{\boldsymbol{X}} = \begin{bmatrix} x_{11} - \bar{x}_1 & \cdots & x_{1D} - \bar{x}_D \\ \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & \cdots & x_{nD} - \bar{x}_D \end{bmatrix}$

Standardização de dados

Usar covariâncias entre variáveis apenas faz sentido se forem medidos nas mesmas unidades. Se as variáveis são muitos dispersas, faz-se uma standardização.

As variáveis standardizáveis são, $z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$

Measure	Parameters (Population)	Statistics (Sample)
Mean	$\mu_X = E[X]$	$\bar{x}$
Mean vector	$\boldsymbol{\mu}_X = E[\boldsymbol{X}]$	$\bar{\boldsymbol{x}}$
Variance	$\sigma_X^2 = Var[X]$	s_x^2
Covariance	$\sigma_{XY} = \mathrm{Cov}(X,Y)$	s_{xy}
Covariance matrix	$\boldsymbol{\Sigma}_{\mathbf{X}} = Cov[\boldsymbol{X}]$	$\boldsymbol{S}$
Correlation	$\rho_{XY} = Cor(X,Y)$	r_{xy}
Correlation matrix	$\mathbf{P}$	$\boldsymbol{R}$

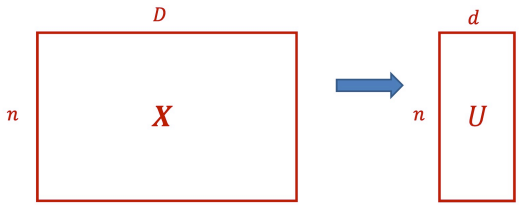
5. Principal Component Analysis (PCA)

5.1 Motivação

Big data está focada em como gerir grandes dimensões de dados, "big n", que são o número de filas num conjunto de dados. É habitual haver problemas com "big D", que são o número de variáveis. A informação pode ter milhares ou milhões de dimensões, o que pode causar problemas, como:

- Informação torna-se dispersa, algoritmos perdem significado;
- Variáveis são correlacionáveis, o que acaba por causar redundância;
- A complexidade de vários algoritmos depende da sua dimensão, o que os torna impraticáveis.

Com PCA pretende-se reduzir D (número de variáveis) para conseguir analisar um subconjunto pequeno de informação e não correlacionável. Ou seja, pretende-se transformar as variáveis D em conjuntos de informação pequenos com d variáveis derivadas:



A seguir, este pequeno conjunto pode ser usado em outros algoritmos como regressão linear, clustering, etc.

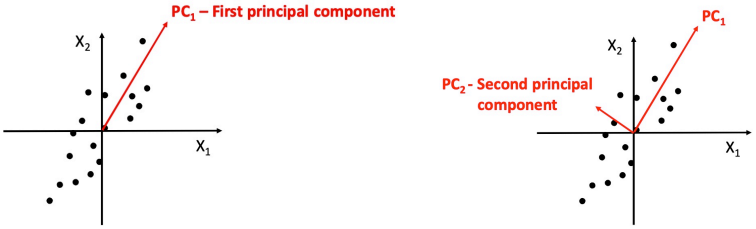
Tem como motivação descobrir e sintetizar padrões de intercorrelações entre variáveis. É útil para grandes bases de dados, com muitas variáveis e para resumir informação. A ideia por de trás destes métodos é evitar contagem a dobrar da mesma informação ao distinguir entre o conteúdo de informação individual de cada variável.

Ou seja:

PCA é um método usado para reduzir a dimensão de um conjunto de dados, ao transformar os mesmos em conjuntos de informação mais pequenos, mas que ainda assim contenham a maioria (conteúdo importante) da informação que estava no conjunto de dados grande.

PCA tem como ideia principal reduzir o número de variáveis de um conjunto de dados, enquanto preserva tanta quanta informação possível.

5.2 Intuição



5.3 Os Básicos

A análise de componentes principais resume informação ao encontrar grandes correlações em combinações lineares de observações, logo:

- Pouca informação se perde no processo;
- Tem uma maior aplicação — variáveis correlacionáveis são transformadas em variáveis não correlacionáveis.

Pearson e Hotelling inventaram um método, provavelmente o mais usado e conhecido, método de standard multivariado.

A análise de componentes principais:

- Cria novas variáveis que são funções lineares das variáveis originais;
- Reduz o número de variáveis originais enquanto retém tanto quanto possível a variação presente em conjuntos de informação;
- Remove variáveis redundantes na base de dados, faz compressão de dados e remoção de "barulho".

5.4 Aplicações

- Deteção/redução de estrutura:
  - Descobre e resume padrões de intercorrelações entre variáveis;
  - Filtra informação e revela estrutura escondida;
  - Técnicas para lidar com multicolinearidade;
- Facilitação de interpretação de um grande número de variáveis;
- Desenvolvimento de escala é derivado de medidas de deriváveis diretamente observáveis;

- Definindo indicadores de construção e a avaliação de qualidades de medidas;
- A avaliação da dimensão do conjunto de variáveis;
- Component scores podem ser usados como a nova variável;

• PCA tens outros usos.

6. Teoria do PCA

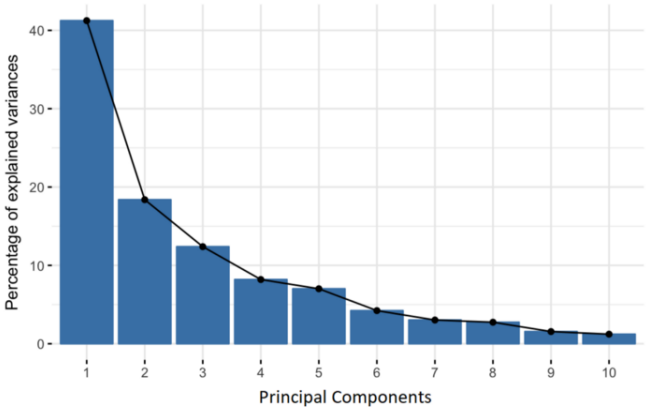
Os principal components são um novo sistema coordenado.

Dada a informação em D variáveis, espera-se que os pontos dos dados permaneçam principalmente num subespaço linear da dimensão mais baixa do que D. Por tanto se houver 5 variáveis, D, espera-se que depois de aplicado o PCA, haja 5 ou menos que 5 variáveis derivadas, d.
Na prática, a informação não vai ficar exatamente em algum subespaço dimensionalmente a baixo, mas pode-se tornar aproximado com o subespaço da dimensão reduzido d << D, o que retém a maioria da informação/variabilidade da informação.

Vão haver novos d, novas variáveis a definir o subespaço. As novas variáveis, que formam um novo sistema coordenado são chamadas de principal components, e são chamados de u1, … , uD — São transformações lineares ortogonais das variáveis originais.

Principal components são novas variáveis que são construídas como combinações lineares ou misturas das variáveis iniciais. Estas combinações são feitas de uma maneira a que as novas variáveis não sejam correlacionáveis e a maioria da informação dentro das variáveis iniciais seja comprimida para os primeiros components.

Ainda assim PCA tenta colocar o máximo de informação possível para o primeiro componente, e a seguir o máximo restante no segundo componente e por aí. Como é demonstrado no gráfico:



A variância desta componente é:

$$Var(u_j) = Var(a_{j1}x_1 + \cdots + a_{jD}x_D) = Var(\boldsymbol{a}_j^T \boldsymbol{x}) = \boldsymbol{a}_j^T \boldsymbol{S} \boldsymbol{a}_j$$

Onde S é uma amostra da covariância da matriz.

Escolhe-se a 1ª PC, (u1) para ter máxima variância. As PC's subsequentes vão ter sucessivamente partes mais pequenas do total de variabilidade. Assim o objetivo é encontrar a combinação linear (aj) que maximiza a variação de uj: Maximize $Var(u_j) = \boldsymbol{a}_j^T \boldsymbol{S} \boldsymbol{a}_j$

Onde S é uma amostra da covariância da matriz. Já que aj é arbitrário e pode aumentar ao escalar, normaliza-se o lambda (eigenvalue). Pode maximizar-se lambda (eigenvalue) dado por,

$$\lambda = \frac{\boldsymbol{a}_j^T \boldsymbol{S} \boldsymbol{a}_j}{\boldsymbol{a}_j^T \boldsymbol{a}_j}$$

Características de PC's:

- São transformações lineares ortogonais das variáveis originais.
- O objetivo é reduzir a dimensão, ou seja, que seja preciso d PC's para aproximar do espaço criado pelos valores de x1, … , xd.
- PC's são combinações lineares de variáveis observadas que sejam independentes de outros componentes.

6.1. Matriz de Covariância

O que é pretendido com isto é perceber como as variáveis do input do conjunto de dados variam da média em relação umas às outras. Ou seja, é verificar se há alguma relação entre elas. Isto porque, as variáveis são altamente correlacionáveis de tal maneira que contêm informação redundante. Então, para identificar essas correlações, computa-se a matriz de covariância.

A covariância da matriz é uma matriz simétrica p*p, onde p é o número de dimensões, que tem como entradas de covariâncias associadas com todos os pares possíveis.

Por exemplo, para um conjunto de dados de 3 dimensões, com 3 variáveis x, y, e z, a matriz de covariância é uma matriz 3*3 no seguinte formato:

[Cov(x, x) Cov(x, y) Cov(x, z)]
[Cov(y, x) Cov(y, y) Cov(y, z)]
[Cov(z, x) Cov(z, y) Cov(z, z)]

A covariância de matriz é uma matriz real simétrica positiva semi-definida.

- **Eigenvalues:**
  - São necessários para fazer a computação da matriz de covariância para determinar PC da informação;
  - Vem aos pares com um eigenvector correspondente;
  - Há tantos eigenvalues como dimensões, no conjunto de dados;
  - São os coeficientes agarrados a eigenvectors que dão a quantidade de variância levada em cada PC;
  - Um eigenvalue representa a variância de um componente:
  - Todos os eigenvalues são maiores ou iguais a zero;
  - Todos os eigenvalues têm de ser reais.

- **Eigenvectors:**
  - São necessários para fazer a computação da matriz de covariância para determinar PC da informação;
  - Vem aos pares com um eigenvalue correspondente;
  - Há tantos eigenvectors como dimensões no conjunto de dados;
  - Não muda de direção numa transformação. São a direção dos eixos onde há a maior variância, mais informação;
  - Eigenvectors correspondentes a eigenvalues diferentes, são ortogonais;
  - Há preocupações com informação desaparecida.

Ao ordenar os eigenvectors por ordem dos seus eigenvalues, do mais alto para o mais baixo, obtém-se os PC's por ordem de importância.

O input à análise pode ser tanto a covariância (S) ou a matriz de correlação (R).

Ao usar a matriz de covariância, variáveis com grande variância vão dominar os resultados do PCA como resultados dependentes nas unidades usadas para medir as variáveis originais e a amplitude de valores que assumem.

Deve-se usar a correlação de matriz, a não ser que as variáveis tenham sido standardizadas. Nesse caso:
- As variáveis são standardizadas;
- Variáveis originais podem ser em unidades diferentes;
- Todas as variáveis têm o mesmo impacto na análise.

6.2. Component Loadings
Os component loadings aj, são as correlações entre as variáveis (filas) e os componentes (colunas).
O loading component ao quadrado é a percentagem da variância na variável que é explicada pelo componente.

6.3. Comunalidades
A comunalidade da variável Xj é a soma do quadrado dos loadings para a variável nos componentes retidos.

A comunalidade representa a quantidade da variância da variável original que é somada pelos componentes retidos, enquanto 1-hˆ2j é a quantidade da variância da Xj descartada ao selecionar apenas os primeiros d componentes. Para cada variável, é a soma dos loadings ao quadrado ao longo dos componentes, Rˆ2 de cada variável.

6.4. Passos em PCA
1. Verificar adequação no PCA
A adequação vai depender do tamanho da amostra, do número de variáveis e o rácio do tamanho da amostra para o número de variáveis. Para isso, computa-se a matriz de correlação para todas as variáveis.

Quando as ligações entre as variáveis originais são fracas, não é possível fazer uma redução de informação relevante sem perder uma boa quantidade de informação, por tanto, se a correlação entre variáveis for pequena, é improvável que partilhem PCs comuns.

Os coeficientes de correlação maiores que 0.3 em valor absoluto são indicativos de correlação aceitáveis.

O teste de esfericidade de Bartlett pode ser usado para testar a hipótese nula em que as variáveis não são correlacionáveis na população. Por outras palavras, a população matriz de correlação é uma matriz de identidade. Se a hipótese não pode ser rejeitada, então a apropriação do PCA deve ser questionada.

2. Extração de PC e número de PC
O 1º PC é uma combinação que conta para a maior quantidade de variância na amostra.

O 2º PC é a 2ª maior quantidade de variância na amostra e não está correlacionada na amostra.

Os componentes sucessivos explicam progressivamente porções pequenas do total da variância da amostra e são todos não correlacionáveis com todos. Requer D componentes principais para reproduzir a matriz de covariância observada com D variáveis medidas.

Critério de Kaiser

O Critério de Kaiser retém componentes cuja variância eigenvalue (variância de 1 componente) é maior do que um eigenvalue médio. Os componentes com variância menor que 1 não são melhores que uma só variável, já que devido a standardização, cada variável tem a variância de 1. Daí que apenas factores com eigenvalues maiores que 1 são retidos.

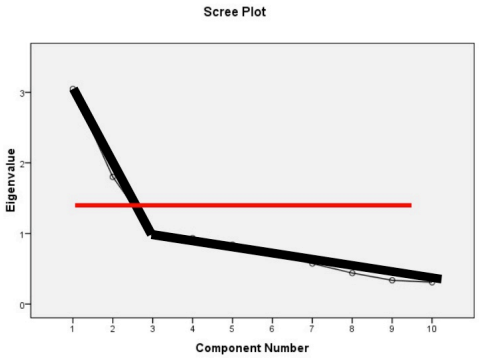
Variância explicada

Com esta abordagem o número de componentes extraído é determinado, para que a percentagem cumulativa da variância extraída pelos componentes atinja níveis satisfatórios. A variância explicada é dada por, $\frac{\sum_{j=1}^{d} \lambda_j}{\sum_{j=1}^{D} \lambda_j}$

É recomendado que os componentes extraídos devam ser pelo menos 60% da variância, retém 60% da informação em dados, e 40% é perdido.

Scree plot

Um scree plot representa os eigenvalues contra o número de components por ordem de extração. A examinação do scree plot fornece um representação da variância total associada com cada PC. O cotovelo é associado com eigenvalues menores que 1. Neste ponto a decisão sobre o número de PC's não é final.



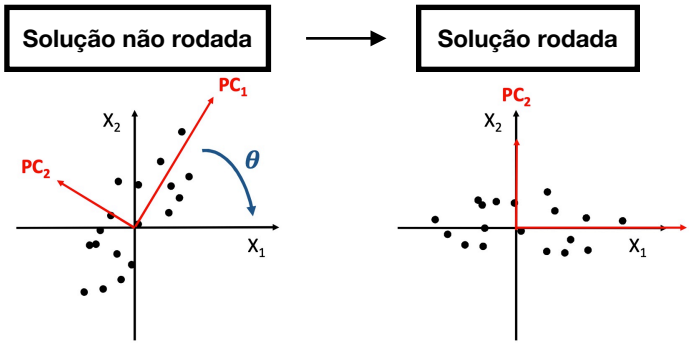
3. Rotação e interpretação de PC

3.1 Interpretação

Uma boa solução de componente é tanto simples como interpretável. Os loadings de components são correlações entre cada uma das variáveis originais e os componentes. Estão ligadas por mais 1 ou menos 1, onde valores altos (positivos ou negativos) indicam que a ligação entre os factores e as variáveis consideráveis é forte.

Daí que um componente possa ser interpretado em termos de variáveis que são loads high nela.

3.2 Rotação



O reposicionamento dos eixos muda os loadings nos componentes, mas mantém o posicionamento relativo dos pontos.

Os componentes foram atingidos ao usar o critério da máxima variância:

- É bom para previsões, usa o menor número de componentes possível;
- Não é muito interpretável.

Então, uma vez determinado o número de componentes desejados, roda-se para dar mais significado e para facilitar a interpretação, tem uma estrutura mais simples.

Ao rodar os factores, o ideal era que cada variável não tivesse zeros ou loadings significantes com poucos factores, se possível com apenas um. Diferentes métodos de rotação podem resultar de alguma maneira em diferentes PC's.

3.2.1 Varimax
É o método rotacional mais popular. Tem uma estrutura simples ao maximizar a variância dos loadings dentro de cada component ao longo das variáveis. Faz com que grandes loadings sejam ainda maiores e com que pequenos loadings fiquem mais pequenos dentro de cada component.

Usa rotações ortogonais, ou seja, os components são mantidos como não correlacionáveis. Espalha a variância do 1º factor para outros mais pequenos.

3.2.2 Oblique
Por vezes é melhor encaixar a informação com eixos que não são ortogonais, mas com o custo de poder ter components que são correlacionáveis uns com os outros. A rotação oblíqua permite aos factores correlacionar com uma imagem conceptualmente limpa, o que é mau para explicar.

Rotação oblíqua deve ser usada quando fatores na população são prováveis de ser fortemente correlacionáveis.

Decisão final
A decisão final sobre o número de PC's a escolher é o número de PC's para a rotação solucionada que é mais interpretável.

4. Realizar decisões finais no número de PC's
Geral
Quando uma solução de PC está boa, pode ser criado um conjunto de novas variáveis que representa os pontos de cada observação das PC's, isto com a solução não rodada. O ponto de observação i é o peso da soma da variável pontuações, onde o peso são os eigenvectors.

Para a componente j, o factor pontuações é, $u_{ij} = a_{1j}x_{i1} + \cdots + a_{Dj}x_{iD} = \boldsymbol{x}_i\boldsymbol{a}_j^T$

Uma vez que a matriz A tenha sido computada, o componente pontuação pode ser também computado ao aplicar, $\boldsymbol{U} = \boldsymbol{XA}^T$

Standardização
A variância da pontuação em cada eixo de PC é igual ao eigenvalue correspondente para esse eixo. Pode ser conveniente standardizar a componente pontuação para que tenha a variância igual a 1.

Usado como input
Tirando o seu uso direto, pontuações de PCA podem ser usadas para construir análises de dados mais efetivas em dimensões reduzidas de espaço, como:
- Deteção de outliers: PCA é útil em descobrir se anormalidades existem num conjunto de dados multivariado;
- Clustering: PCA é util em fazer clusters de variáveis altamente correlacionáveis;
- Regressão: pontuações de PCA podem ser usadas como previsões para abordar multicolinearidade de variáveis independentes.

5. Criar scores

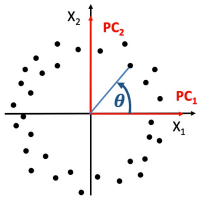
6.5. Limitações de PCA
- É preciso uma estrutura linear;
- Se as variáveis não são à partida correlacionáveis, então PCA não vai acrescentar nada;
- É preciso reter todos os componentes, apenas mudar as coordenadas não resulta.

Ortogonalidade
Muitas vezes conjuntos de dados contêm estruturas de dados especificas que não são capturadas pelo PCA.

Estatísticas de segunda ordem
Ao aplicar PCA, assume-se que médias de vectores e matriz de covariância são estatísticas suficientes para descrever as relações entre variáveis. Isto é verdade se a informação tender para ter uma distribuição normal. Caso contrário, médias de vetores e matrizes de covariâncias não são estatísticas suficientes e não resumem a informação.

Nesse caso há o ICA (Independent Componente Analysis):
- Parecido a PCA, mas assume características non-Guassian (de distribuição que não seja normal, curva de sino);
- Usa estatísticas de ordem alta.

Geometria do espaço
- A forma deste conjunto de dados não está bem descrita pelos seus componentes principais;
- O angulo 0 e o raio contêm a informação, mas é não-linear.



Não-linearidade
PCA assume que informação está numa dimensão linear diversa baixa. Se o padrão for não linear, os eixos dos PC's não são adequados a resumir a informação. Pode-se usar CCA (Curvilinear Component Analysis):

- É uma extensão não-linear de PCA;
- Preserva a proximidade entre os pontos do espaço input;
- Permite desdobrar a informação de input;
- Mantém a topologia local.

6.6. PCA VS clustering / classificação

PCA assume que grandes variâncias revelam as estruturas mais interessantes. A maior variância determina quais os componentes a ser usados, mas não garante pontos de vista interessantes para clustering de informação. A direção da variância máximo não é sempre boa para classificação.

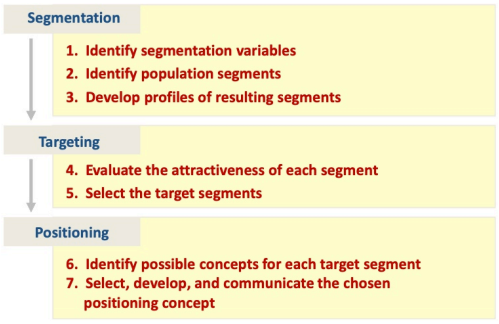
6.7. PCA e Análise de Factores (FA)

PCA é uma transformação matemática, enquanto FA é um modelo probabilístico. Os PC's operam em toda a variabilidade original, onde FA opera em variância comum. Os componentes são variáveis compostas pesadas e não se distinguem entre variâncias comuns e únicas.

7. Clusters

7.1 Segmentação de mercado

Segmentação de mercados envolve visualizar um mercado heterogéneo (diferente) como um número de mercados pequenos homogéneos (conjuntos de pequenos mercados iguais), em resposta a preferências diferentes, atribuindo a desejos de consumidores para uma maior satisfação das suas múltiplas vontades.



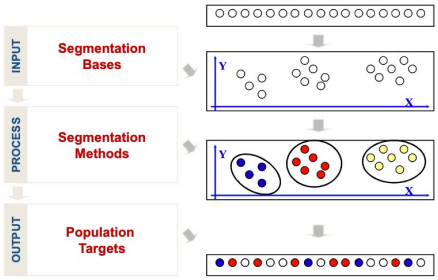
	General	Product-specific
Observable	Demographic, cultural, geographic and socio-economic variables	User status, usage frequency, situation
Unobservable	Psycographics, values, personality and lifestyle	Psychographics, benefits, perceptions, elasticities, attributes, preferences, intention

7.2 Benefícios da segmentação de mercados

Um único produto não é provável de agradar a todos os consumidores, então o objetivo é identificar os grupos de consumidores que encontram diferentes variações de produtos para ser atrativo. Desta maneira atinge-se um melhor conhecimento das necessidades de um grupo selecionado de clientes.

Há um melhor posicionamento e diferenciação, melhor uso de orçamentos e uma precisão maior em técnicas de média com o uso de segmentação de mercados.

7.3 Processo de Segmentação



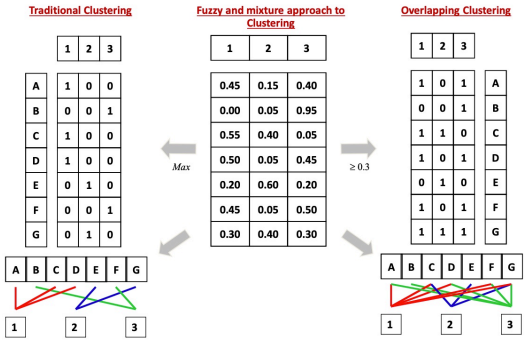
7.4 Abordagens com Clustering

❑ **Clustering**

Traditional Clustering	Fuzzy and mixture Clustering	Overlapping Clustering
- Hierarchical clustering - K-means algorithm	- Fuzzy clustering - GoM Model - Mixture models	- ADCLUS and its generalizations

❑ **Partition**

Hard partition	Fuzzy partition	Overlapped partition
$g_{ik} \in \{0,1\}$ $\sum_{k=1}^{K} g_{ik} = 1$	$g_{ik} \in [0,1]$ $\sum_{k=1}^{K} g_{ik} = 1$	$g_{ik} \in \{0,1\}$ $1 \le \sum_{k=1}^{K} g_{ik} \le K$



7.5 Problemas com informação

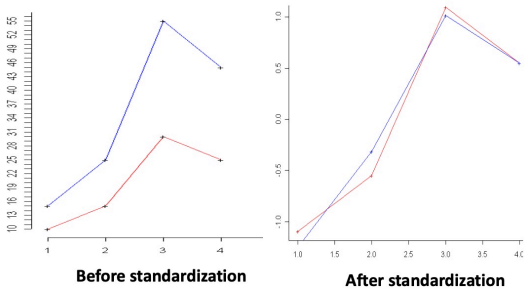
Existem problemas sobre os tipos de informação, sobre o que dita tipo de distância, sobre tipos de distribuição e eliminação de outliers.

7.6 Standardização de informação

O valor das medidas de distância está intimamente relacionado com a escala (unidade) na qual as medidas são feitas. Daí que, variáveis são standardizadas antes de medir as dissimilaridades inter-observáveis. Isto é recomendado quando as variáveis são medidas em escalas (unidades) diferentes. Caso contrário, a medida dissimilar obtida é gravemente afetada.

O objetivo é tornar as medidas comparáveis. As variáveis são standardizadas para ter um desvio padrão de 1 e uma média de 0.

$X' = [X_j - \text{mean}(X)]/\text{std}(X)$



7.7 Seleção de variáveis

A seleção de variáveis é baseada em razões teóricas e práticas. Depende também de serem variáveis ativas ou passivas. É possível reduzir o número de variáveis com PCA.

8. Diferença e distância

8.1 Semelhança, Diferenças e Distância

A classificação de observações em grupos requer alguns métodos para computar a distância e a (dis)similaridade entre os pares de observações. O resultado desta computação é conhecido como a matriz de dissimilaridade.

Semelhança (Sij) indica a força de relações entre objetos i e j. Normalmente, Sij tem valores entre 0 e 1.

8.2 Algumas medidas de distância

- **Euclidean distance** $d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2}$ **or** d_{ij}^2
- **Manhattan distance (city-block distance)** $d_{ij} = \sum_{k=1}^{p}|x_{ik} - x_{jk}|$
- **Mahalanobis distance** $d_{ij} = (\mathbf{x}_i - \mathbf{x}_j)^T \Sigma^{-1} (\mathbf{x}_i - \mathbf{x}_j)$
- **Correlation distance** $d_{ij} = 1 - \frac{\sum_{k=1}^{p}(x_{ik} - \bar{x}_k)(x_{jk} - \bar{x}_k)}{\sqrt{\sum_{k=1}^{p}(x_{ik} - \bar{x}_k)^2 \sum_{k=1}^{p}(x_{jk} - \bar{x}_k)^2}}$

8.3 Seleção da medida de distância

A medida de distância define como a semelhança de 2 elementos (w, y) é calculada e como pode influenciar a forma dos clusters.

A escolha da medida de distância deve ser baseada na área de aplicação. Há 3 componentes: forma, dispersão e elevação. Medidas de distâncias têm um grande impacto no desempenho e nos resultados finais. Euclidean e Manhattan medem diferenças absolutas entre vetores, ou seja, levam em conta a magnitude da expressão.

A distância de correlação, isto é, a distância que mede a relação entre variáveis, mede tendências. A distância de correlação mede a distância angular, ou seja, insensibilidade à amplitude de expressão, leva em conta as tendências da mudança.

Standardização pode ser aplicada a variáveis, ou seja, subtrai-se as médias e divide-se pelo desvio padrão. Depois da standardização, a distância Euclidean e de correlação são iguais. A distância de Manhattan é mais robusta em relação a outliers.

8.4 Relação entre semelhança e distância

A semelhança é avaliada ao usar a noção de distância. Medidas de semelhança ou distância são componentes importantes usados em algoritmos de cluster, baseados em distância para pontos de informação parecidos, enquanto que pontos de informação distantes são colocados em clusters diferentes.

Distância é o contrário de semelhança.

9. Clustering

9.1 Classificação

Objetos são classificados em grupos. Esta organização torna-se necessária porque é conveniente, e com isto é possível prever e explicar melhor a informação. É importante notar que isto nem sempre leva à mesma classificação.

9.2 Classificação VS Clustering

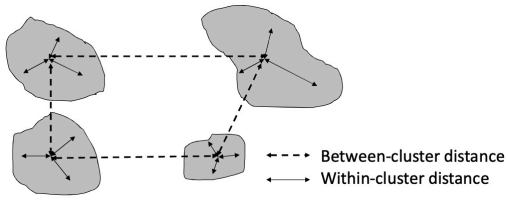
- Em Classificação conhece-se as labels e é de aprendizagem supervisionada
- Em Clustering não se conhece as labels e é de aprendizagem não supervisionada

Como não acontece com uma teoria, uma classificação resultante de um clustering deve ser julgada pela utilidade dos seus resultados. No entanto, uma classificação que resulta de um cluster pode ser útil para sugerir uma teoria, que pode mais tarde ser testada.

Clustering é encontrar grupos de objetos para objetos que sejam parecidos uns aos outros e diferentes aos objetos de outros grupos. Dada uma coleção de n objetos, cada um deles é descrito por um conjunto de p características ou variáveis que derivam uma divisão útil em um número de classes, clusters. Ambos os números e classes de propriedades são para ser determinadas. Isto é feito porque dá para organizar, perceber, explorar e prever com base em grupos.

Normalmente a análise em clusters é baseada em 2 coisas:

- Distance mesaure, ou seja, quantidade de semelhança nos objetos;
- Algoritmo de clustering, sendo este um processo para agrupar objetos.



Isto tem o objetivo de estabelecer distâncias dentro de pequenos clusters, e grandes distâncias entre clusters.

9.3 Dificuldades com clustering

- A estrutura de clusters pode-se manifestar de várias maneiras;
- Grandes conjuntos de dados (n) e grandes dimensionalidades (p) complicam os assuntos;
- Dados com grande dispersão;
- Outliers significativos.

9.4 Abordagens com clustering

- Model-free clustering: sem modelo probabilístico, hierarquia de clustering, media e SOM;

- Model-based clustering: com modelo probabilístico, mistura finita de modelos;

- Non-overlapping: clustering de partição e clustering hierárquico;

- Overlapping: em clustering não exclusivo, os pontos podem pertencer a múltiplos clusters;

- Fuzzy: fuzzy clustering indica que um ponto pertence a cada cluster com algum peso entre 0 e 1 com soma de pesos a dar 1. A atribuição de pontos de informação em qualquer um dos clusters não é decisivo. Aqui um ponto de informação pode pertencer a mais do que 1 cluster. O resultado disto é a probabilidade do ponto de informação de pertencer a cada um dos clusters;

- Clustering hierárquico: um conjunto de clusters agrupados organizado como uma arvore hierárquica, pode ser aglomerativo e divisivo. Este tipo agrupa clusters com base na distância métrica;

- Particionamento (divisão) de clustering: divisão de objetos de informação para non-overlapping. Os clusters são divididos com base em características dos pontos de informação. É necessário especificar o número de clusters a ser criado para este método. Este algoritmo de cluster segue um processo iterativo para retribuir os pontos de informação entre clusters com base na sua distância.

9.5 Clustering hierárquico

Produz um conjunto de clusters agrupados organizados como uma árvore hierárquica. Pode ser visto como um dendrograma, uma árvore como um diagrama que guarda as sequências de junções ou divisões.
Algoritmos tradicionais hierárquicos usam a semelhança ou distância de matriz. Fazem a junção ou divisão de um cluster de cada vez.
As estratégias para clustering hierárquico tendem a estar divididas em 2 tipos:

- Aglomerativo — é uma abordagem “bottom-up” em que cada observação começa no seu próprio cluster e pares de clusters são conectados como um movimento para cima na hierarquia;

- Divisivo — é uma abordagem "top-down", em que todas as observações começam num único cluster e são realizadas divisões recursivamente quando um se movimenta para baixo da hierarquia.

9.5.1 Cluster aglomerativo — "bottom-up"

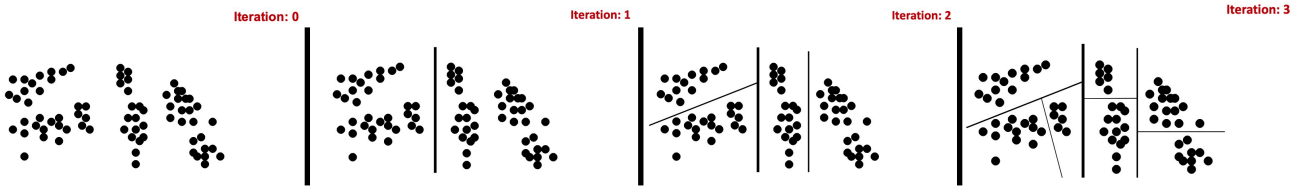
É o tipo mais comum de clustering hierárquicos usado para agrupar objetos em clusters com base na sua semelhança.

O algoritmo começa por tratar cada objeto como um cluster singular. A seguir, pares de clusters são sucessivamente misturados até que todos os clusters tenham sido agrupados para um grande cluster que contém todos os objetos. O resultado disto é uma representação de objetos que é baseada numa árvore, chamada de dendrograma.

9.5.2 Cluster divisivo — "top-down"

Começa com todos os objetos num só cluster e vai-se dividindo esse cluster sequencialmente até ao ponto em que resta apenas um objeto por cluster.

- Tem a vantagem de obter a estrutura principal da informação;
- Tem a desvantagem de ter dificuldades computacionais quando se considera todas as divisões possíveis em 2 grupos.



O divisivo é o oposto de aglomerativo, começa com todos os pontos de informação e divide-os para criar mais clusters. Estes algoritmos criam a distância de matriz dos clusters existentes e realiza a ligação entre os clusters dependendo do critério de ligação.

Existem tipos diferentes de ligações:
- Ligação única — numa ligação única, a distância entre 2 clusters é a distância mais curta entre pontos nesses 2 clusters;
- Ligação completa — numa ligação completa, a distância entre 2 clusters é a maior distância entre pontos nesses 2 clusters;
- Ligação média — numa ligação média, a distância entre 2 clusters é a distância média de cada ponto no cluster com cada ponto em outro cluster.

O clustering destes pontos de informação é representado ao usar um dendrograma.

9.5.3 Vantagens

- A semelhança de objetos é representada numa estrutura em árvore, dendrograma, pode corresponder a técnicas de classificação importantes;
- Clusters agrupados podem ser representados;
- Não tem que se assumir um número particular de clusters: qualquer número dos mesmos pode ser obtido ao cortar o dendrograma no nível adequado.

9.5.4 Desvantagens

- Uma vez tomada a decisão de combinar 2 clusters, não pode ser desfeita;
- A função objetiva não é diretamente minimizada;
- Os resultados dependem da distância do método de update;
- Processo iterativo ganancioso;
- Não há medida inerente para identificar clusters estáveis.

A árvore pode ser construída em 2 maneiras diferentes:

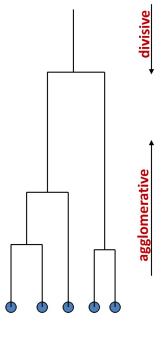
1. Métodos aglomerativos:
   - Fusão de n objetos em grupos;
   - n clusters para 1 cluster;
   - Bottom-up cluster.

2. Métodos divisivos:
   - Separa n objetos em grupos pequenos;
   - 1 cluster para n clusters;
   - Algoritmo top-down;
   - Métodos divisivos são menos comuns.

9.6 O dendrograma

Representa o resultado de algoritmos de cluster hierárquicos ao mostrar o quanto os clusters estão juntos ou separados de maneira hierárquica.

Fornece uma sequência de partições agrupada. O vertical mostra uma medida geral de semelhança entre clusters.



9.7 O processo

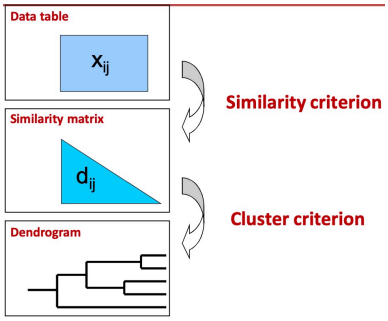
I. Obtém-se a informação sem ser tratada;

II. Aplica-se um critério de semelhança sobre a informação recolhida;

III. O passo anterior resulta numa matriz de semelhança;

IV. Aplica-se um critério de cluster sobre o resultado anterior, agrupando a informação em clusters;

V. Por fim, o passo anterior resulta num dendrograma.



9.8 Decisão no número de clusters

O clustering hierárquico, não dita a partição e o número de clusters. A altura de um nó no dendrograma representa a distância a que se encontra dos seus 2 filhos clusters. Normalmente uma altura é escolhida e é aí que o corte é feito, mas os cortes mais informativos costumam estar em alturas diferentes para ramos diferentes.

Para assentar sobre uma partição, tem de se cortar o dendrograma no nível desejado, então cada componente que esteja ligado forma um cluster.

10. Abordagem de partição

10.1 Métodos de partição

Os objetos são partidos num número de grupos K pré-especifico. Os objetos são iterativamente realocados em clusters até condizer com critérios.

10.2 O algoritmo de média K (K-means)

É um método de quantização de vetores, que tem como objetivo o particionamento de n observações em k clusters, em que cada observação pertence ao cluster com a média mais próxima, servindo como protótipo do cluster.

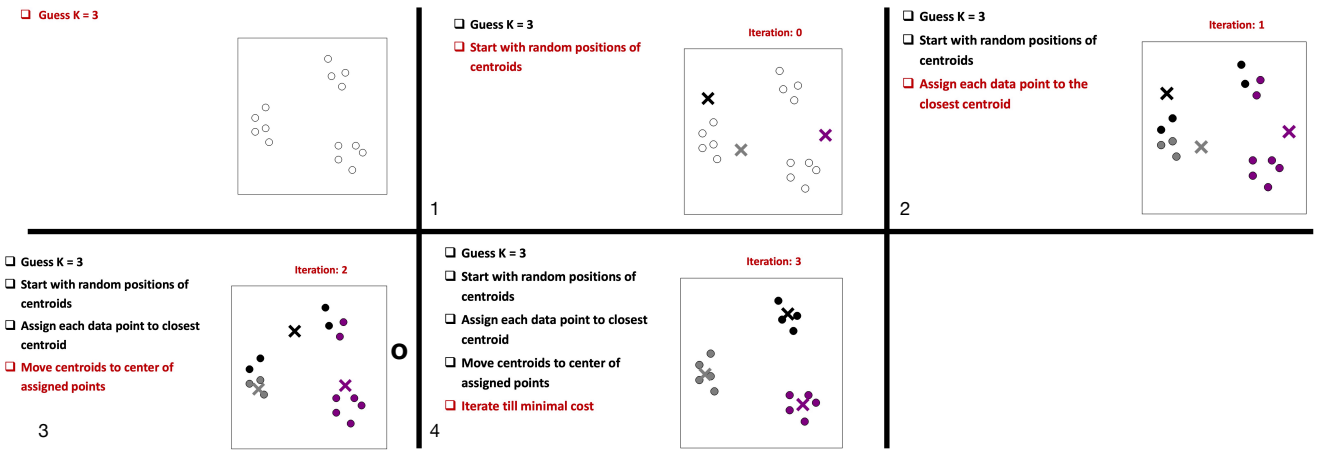
O algoritmo K-means é um algoritmo iterativo que tenta dividir o conjunto de dados em K clusters, onde cada ponto de informação pertence a um único grupo. Este algoritmo tenta fazer com que pontos de informação intra-clusters (dentro do mesmo cluster) sejam tão parecidos quanto possível, enquanto os clusters são mantidos o mais longe possível uns dos outros.

O algoritmo atribui pontos de informação a clusters, tal que a soma da distância ao quadrado entre os pontos de informação e os centróides dos clusters sejam o mínimo possível. Quanto menor for a variância entre os clusters, mais iguais vão ser os pontos de informação dentro do mesmo cluster.

A abordagem que este algoritmo segue para resolver problemas é chamada de Expectation-Maximization. O E-step é atribuído a pontos de informação do cluster mais próximo. O M-step é computar o centróide de cada cluster.

As variáveis devem ser quantitativas, usa-se a distância Euclidean ao quadrado. O algoritmo base é simples:

1. Seleciona-se k objetos como centróides iniciais;
2. Forma-se k clusters ao atribuir todos os objetos para o centróide mais próximo;
3. Recomputa-se os centróides de cada cluster;
4. Repetir os passos 2 e 3 até que os centróides não se mexam mais.



11.1 Modelos de mistura finitas

Um modelo mistura é um modelo probabilístico que tem o objetivo de representar a presença de sub-populações dentro de uma população geral, sem exigir que o conjunto de dados observados deva identificar as sub-populações que pertencem a uma observação individual. Um modelo mistura corresponde à distribuição mistura que representa a probabilidade de distribuição de observações na população em geral.

Algumas maneiras de implementar modelos de mistura envolvem etapas que atribuem determinados pedidos de identidades de sub-população para observações individuais (ou pesos para essas sub-populações), caso em que estas podem ser consideradas como um tipo de aprendizagem não supervisionada ou procedimentos em clusters.

Um modelo mistura típico de finitas dimensões é um modelo hierárquico consistindo dos seguintes componentes:

- N variáveis aleatórias correspondentes às observações, cada uma assumindo ser distribuída de acordo com uma mistura de K componentes, com cada componente pertencente à mesma família paramétrica de distribuições, mas com parâmetros diferentes;
- N correspondentes variáveis latentes aleatórias, especificando a identidade do componente mistura de cada observação, distribuídas de acordo a uma distribuição categórica K-dimensional;
- Um conjunto de K pesos mistura, cada um dos quais é uma probabilidade (um número real entre 0 e 1), todos os quais com soma igual a 1.

11.2 Algoritmo EM (Expectation-Maximization)

É um método iterativo para encontrar a likelihood (probabilidade conjunta de uma informação observada como uma função dos parâmetros do modelo estatístico escolhido) máxima, onde o modelo depende de variáveis latentes não observáveis.

A iteração EM vai alternando entre a realização do passo Expectation (E), que cria a função de expectativas de log-likelihood avaliadas ao usar a estimativa atual para os parâmetros, e a Maximization (M), que computa os parâmetros maximizando a log-likelihood expectável encontrada no passo E.

Estes parâmetros expectáveis são depois usados para determinar a distribuição de variáveis latentes no próximo passo E.

O algoritmo simplifica uma função complexa para um conjunto de funções que facilmente se resolvem ao introduzir um parâmetro/informação desaparecida.

O algoritmo itera entre 2 passos:

- E-step — estima a média condicional do parâmetro desaparecido dado uma estimativa prévia do parâmetros de modelo e de observações;
- M-step — estima de novo os parâmetros do modelo dadas as observações e o clustering feito pelo E-step.

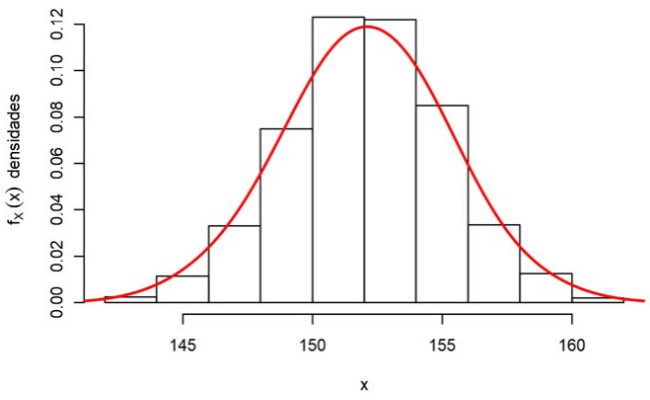
O algoritmo é estável e o valor log-likelihood não diminui a cada iteração.

12. Gaussian mixture model (GMM)

Dentro da área da aprendizagem não supervisionada, modelos de mistura são modelos probabilísticos que representam a probabilidade da presença de clusters dentro de uma população geral.

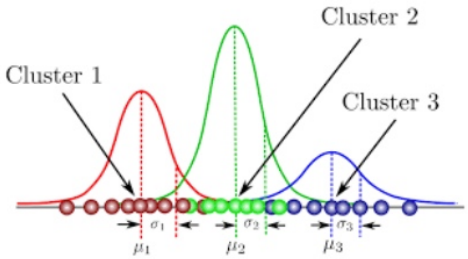
O modelo de Mistura de Gaussian deriva da distribuição normal, ou distribuição de Gaussian. Esta é uma distribuição probabilística que é simétrica em relação à media, mostrando que informação próxima da média acontece com mais frequência do que informação afastada da média.

A distribuição de Gaussian possui 2 parâmetros: a média e a variância, que descreve o grau de dispersão. Também se refere à dispersão como desvio padrão. O centro da curva de sino é a média de todos os pontos de informação, e a distribuição de todos os pontos de informação pode ser representada pelo desvio padrão.



O modelo de mistura de Gaussian é utilizado em situações de aprendizagem não-supervisionada, em que existem várias distribuições de Gaussian (normais). Este modelo calcula a probabilidade para todos os pontos de informação num conjunto de dados, de pertencerem a qualquer uma das distribuições existentes. Por isso é que este é considerado como uma distribuição probabilística.

A vantagem deste modelo é que não requer a especificação da relação entre pontos de informação e de clusters. O modelo aprende as especificações dos sub-clusters e ao usar o que aprendeu, cria clusters de pontos de informação.



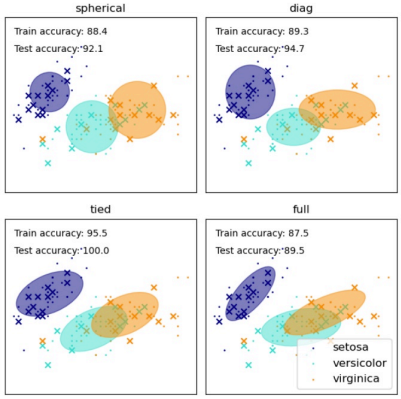
O modelo de mistura de Gaussian, é um modelo probabilístico que assume que todos os pontos de informação são gerados a partir de uma mistura finita de distribuições de Gaussian com parâmetros desconhecidos. Pode-se pensar que modelos de mistura são como k-means clustering para incorporar informação sobre a estrutura de covariância da informação como os centros latentes (variáveis que não podem ser diretamente observadas, mas podem ser inferidas, deduzidas a partir de outras variáveis que são observadas) de Gaussian.

12.1 Gaussian Mixture

A distribuição de mistura é uma distribuição probabilística de pontos de informação num espaço de informação a 3D. É um processo para informação espacial onde ao usar uma distribuição Gaussian se separa a população geral em clusters diferentes.

A mistura de Gaussian implementa o algoritmo de Expectation-Maximization. Isto pode mostrar confiança em elipsóides para modelos multivariados e computar o Critério de informação Bayesian para avaliar o número de clusters na informação. O método GaussianMixture.fit é fornecido para aprender um modelo de mistura de Gaussian da informação de treino.

A GaussianMixture vem com opções diferentes de covariância da diferença de classes estimada: esférica, diagonal, tied ou full covariance.



13. Latent Class Model

O modelo de classe latente relaciona um conjunto de variáveis observadas multivariadas com um conjunto de variáveis latentes. Estas são variáveis que não podem ser diretamente observadas, mas podem ser inferidas, deduzidas a partir de outras variáveis que são observadas.

Por tanto, as classes latentes são as variáveis não observáveis que são derivadas das variáveis observáveis.

É um tipo de modelo de variável latente. É chamado de modelo de classe latente porque a variável latente é discreta. Uma classe é caracterizada por um padrão de probabilidades condicionais que indicam a probabilidade que as variáveis levam para certos valores.