**MBA USP – Ciência de Dados**

**Anotações disciplina: Introdução a Ciência de Dados**

**Aula 1:**

* **Variância e Desvio padrão**

A variância e o desvio padrão calculam o quanto os dados estão dispersos. No caso, o ponto de referência no conjunto é a média de seus valores e assim calcula-se o quanto seus valores estão distantes desta média.

Fórmula:

Variância ( S2) = Soma dos quadrados do desvios / N-1

var(x) = onde = média da amostra.

Desvio Padrão (S ) = Raíz Quadrada da Variância ( √S2 )

dp(x) = ouou , onde = média da amostra.

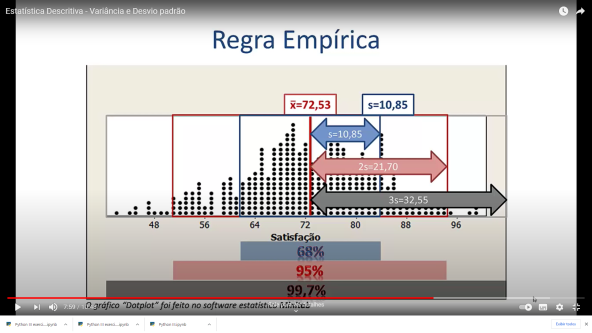
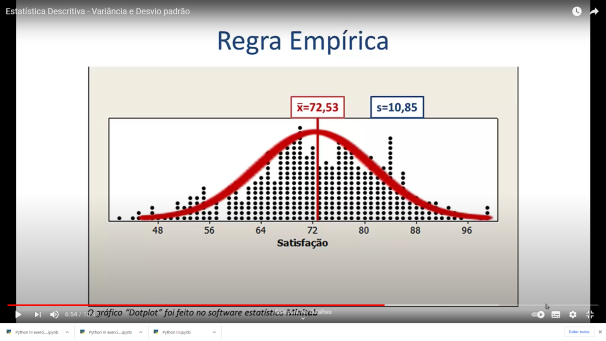
Exemplos:

<https://www.youtube.com/watch?v=4ABBnMUOSb4>

O desvio padrão é utilizado em problemas que envolvam incerteza, neste caso o DP nos ajuda a quantificar esta incerteza.

**Regra empírica da estatística:**

* Se a distribuição de um conjunto de dados for simétrica, ou seja, uma curva normal:
  + Aproximadamente 68% dos dados estão distantes, no máximo, a 1 desvio padrão da média.
  + E aproximadamente 95% dos dados estão distantes, no máximo, a 2 desvios padrão da média.
  + E aproximadamente 99,7% dos dados estão distantes, no máximo, a 3 desvios padrão da média.



**Aplicação exemplo:**

Uma padaria quer saber quantos pães produzir por dia de maneira que não falte pães para vender e evitar ao máximo produzir além do demandado, ou seja, sobrar pães.

No caso, a média de demanda é 1556 pães e o desvio padrão é de 114 pães, conforme ilustração abaixo.

Assim, para ter uma garantia de que em 95% dos casos não faltarão pães, o que é razoável ao dono da padaria, podemos calcular a média + 2 vezes o desvio padrão para saber a quantidade de pães que a padaria deve produzir, no caso:

1556 + (2\*114) = 1784 pães.

Apresentando:

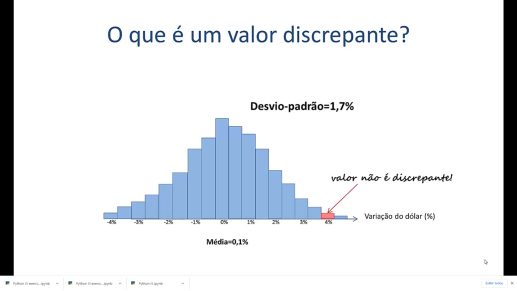
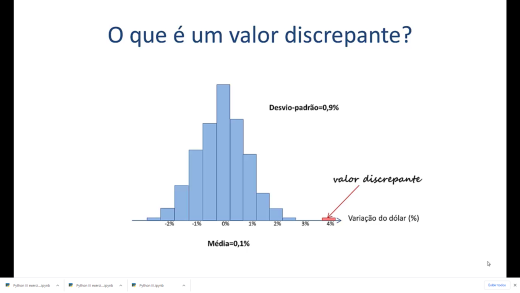


* **Outliers e Padronização**

**Valor Discrepante ou Outlier**

É quando um valor destoa do restante da amostra, fugindo bastante da média e desvios padrão da amostra.

Exemplo:<https://www.youtube.com/watch?v=dR549rr3RH4>



**Critérios para identificar um outlier:**

* Distância do valor até a média, em quantidade de desvios padrão (Valor Padronizado);
* Distâncias entre os quartis e que são analisados através dos gráficos de Box-Plot;
* Outros.

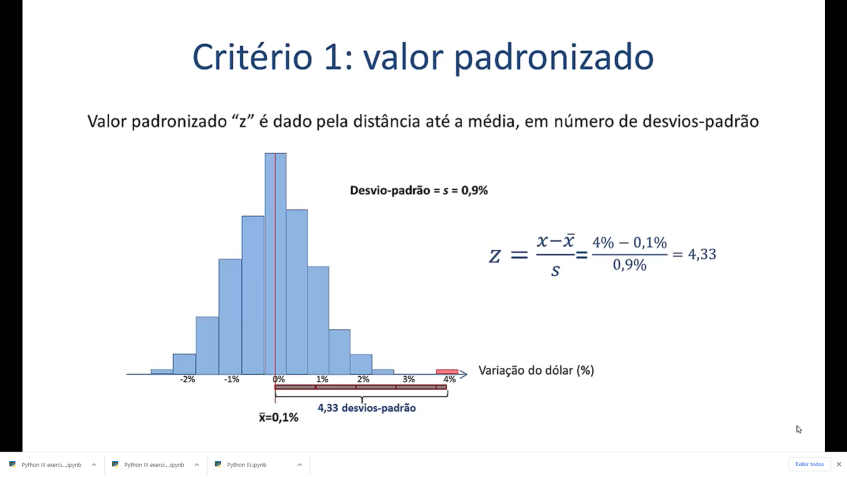
**Critério 1: Valor Padronizado.**

Valor Padronizado (Z) = , onde é a média e S o desvio padrão.

No exemplo abaixo, no caso da variação de dólar, o valor de 4%:

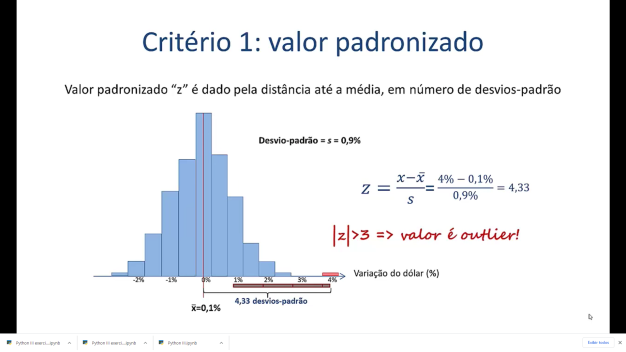
Valor Padronizado (Z) = (4% - 0,1%)/0,9% = 4,33

Ou seja, variação de 4% está a 4,33 desvios padrão da média de variação do dólar.



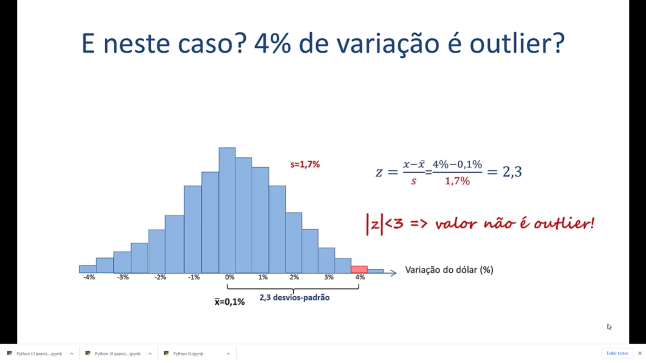
Portanto, levando-se em consideração a **regra empírica da estatística** (distribuição normal), se o valor analisado, no caso 4%, for maior que 3 vezes, em valor absoluto (sem considerar o sinal), o desvio padrão (99,7% de probabilidade), trata-se de um outlier.

No caso, como 4% está a 4,33 dp da média, **4% é considerado um outlier**.

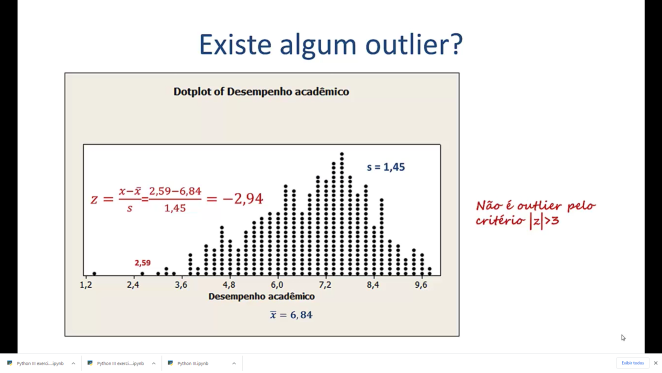
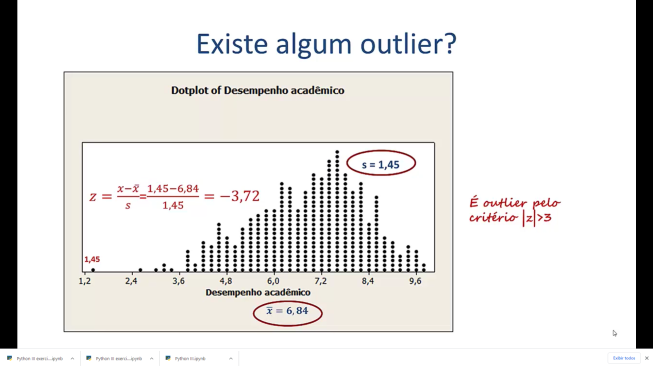


E no outro exemplo abaixo? 4% é um outlier? Resposta: Não.

Pois apesar da média ser a mesma, os dados estão mais dispersos, onde o DP = 1,7%



Outros exemplos:



**NOTE** que o 2º exemplo acima, os dados não estão tão simétricos, não é uma curva normal perfeita. Portanto a regra empírica da estatística pode não se aplicar ao exemplo e talvez não seja o melhor método para se identificar outliers.

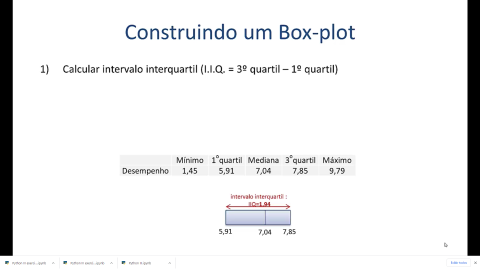
**Critério 2: BOX-PLOT.**

O Box Plot é um gráfico que mostra se existem outliers e o critério não depende da simetria ou distribuição dos dados.

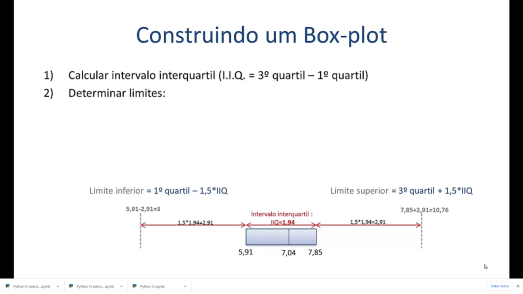
Construindo um Box Plot sobre uma amostra de desempenho acadêmico:

1. Montar os quartis.
   1. Ordenar crescentemente os valores da amostra;
   2. Valor Mínimo é o menor valor e o Valor Máximo o maior valor;
   3. Calcular a mediana, que será considerada o segundo quartil;
   4. Com o lado esquerdo e direito, calcular a mediana de cada lado, que serão o 1º quartil (lado esquerdo) e 3º quartil (lado direito).
2. Calcular Intervalo Interquartil (I.I.Q.) = ( valor do 3º quartil – valor do 1º quartil)

Intervalo Interquartil (IIQ) OU Variação Interquartil (IQR)



1. Determinar os limites:
   1. Limite Inferior = ( valor do 1º quartil – 1,5 IIQ )
   2. Limite Superior = ( valor do 3º quartil + 1,5 IIQ )



1. Desenhar o traço até o menor e maior valores existentes dentro dos limites calculados.

No exemplo, o primeiro elemento da amostra que está acima do Limite Inferior, de valor 3, é o valor 3,05, portanto a “caixa” do Box Plot começa nele.

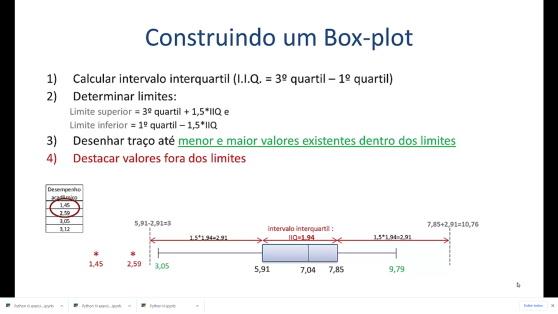
E o último elemento da amostra que está abaixo do Limite Inferior, de valor 10,76, é o valor 9,79, portanto a “caixa” do Box Plot termina nele.

**NOTE** que, neste exemplo, como o maior valor da amostra é o próprio valor 9,79, e ele é menor que o limite superior, não teremos outliers superiores.



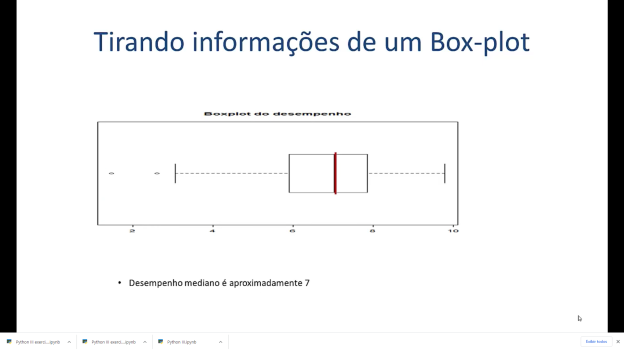
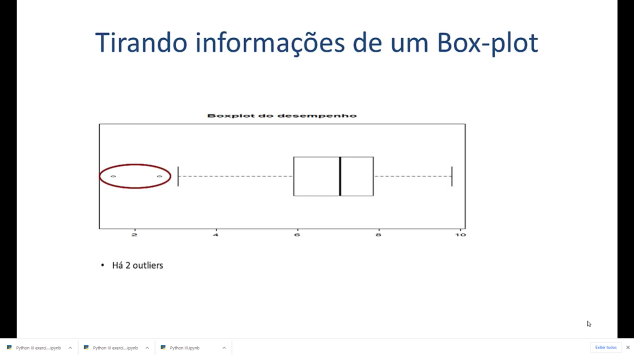
1. Destacar os valores fora dos limites (Outliers).

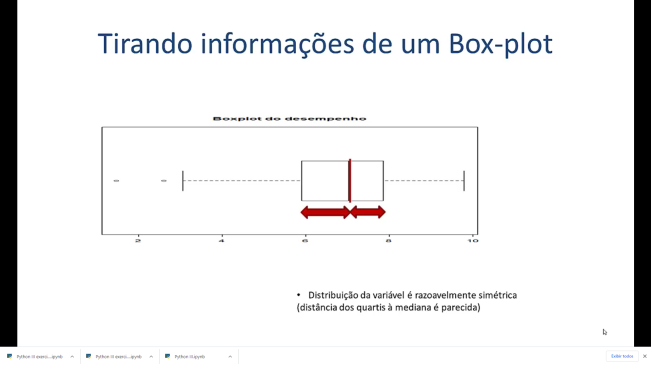
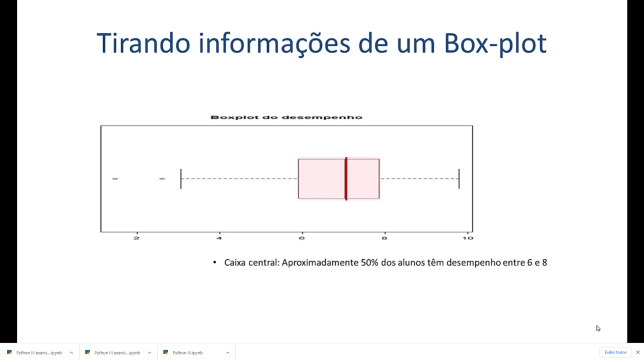
No caso, os valores 1,45 e 2,59 estão abaixo do Limite Inferior, de valor 3, portanto são os outliers desta amostra.



**Mas quais informações podemos tirar do Box Plot?**

1. Há 2 Outliers;
2. Que o desempenho acadêmico mediano é aproximadamente 7;
3. Que aproximadamente 50% dos alunos (caixa central do Box Plot) tem desempenho entre 6 e 8.
4. A distribuição da variável é razoavelmente simétricas, pois a distância do 1º e 3º quartis à Mediada são parecidas.





**Mas o que fazer com os outliers?**

* Excluir? Mas você pode estar perdendo uma informação importante...

Mas se realmente for um valor absurdo como uma idade de pessoa = 180 anos?

Neste caso recomenda-se a exclusão, mas para fazer isto tem que conhecer bem os dados, pois em alguns casos há como corrigí-los.

* Buscar possíveis causas

O mais comum é que Outliers/Valores Discrepantes tenham uma causa especial.

Exemplo: Vendas muito altas, acima do normal, podem indicar que naquele dia ocorreu uma promoção de vendas.

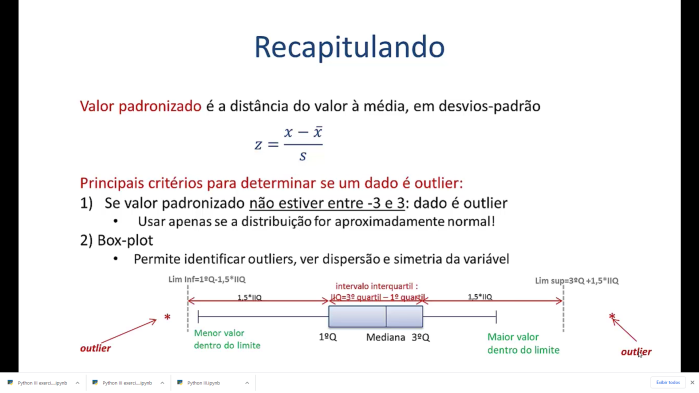
* Fazer análise com e sem o outlier e ver o impacto dele no resultado.
* Fazer uma transformação da variável.

Por exemplo categorizá-la, transformando uma variável numérica em faixas.

Exemplo: Valor Salário (50k) ->Faixa Salarial (Acima de 10k).

**NOTE** que nestes casos, a variável deixa de ser Quantitativa para se tornar Qualitativa. O que requer outros tipos de análises estatísticas.

* Outros.



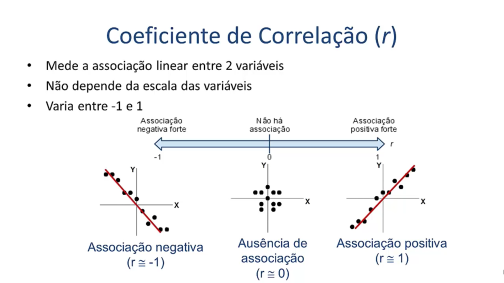
* **Correlação entre as variáveis**

Como medir que uma variável impacta, positivamente ou negativamente ou mesmo não impacta, uma outra variável? Como medir/calcular esta correlação entre variáveis ou identificar que elas são independentes?

**Coeficiente de correlação (r)**

* Mede a associação linear entre 2 variáveis quantitativas.
* Não depende da escala das variáveis.
* Varia entre -1 e 1, sendo que quando é igual a 0 (zero), indica que não há correlação entre as variáveis.

Exemplos: <https://www.youtube.com/watch?v=nW4yMYf8YDg&t=12s>



Cuidados com a correlação:

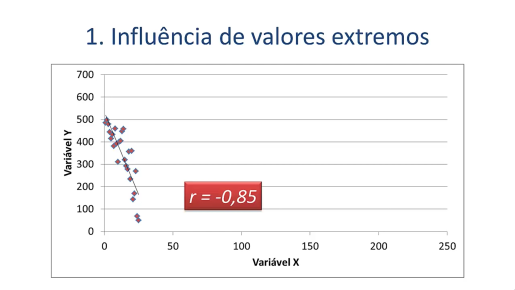
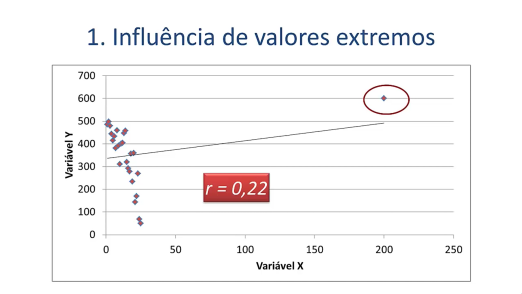
1. Influência de Valores extremos;

No exemplo abaixo, quadro à esquerda, um ponto ao extremo direito, puxa a correlação para positiva, no caso, r=0,22 indicando uma baixa correlação positiva.

Mas o que acontece se retirarmos este valor extremo?

No quadro à direita, após a remoção do outlier, podemos notar que o valor r mudou passou para -0,85, indicando uma forte correlação negativa.

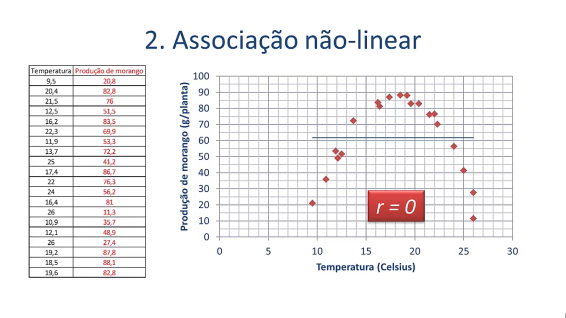
Portanto, devemos sempre fazer uma análise do gráfico de dispersão.



1. Associação não-linear entre as variáveis

No exemplo abaixo, podemos notar que não existe uma correlação linear entre as variáveis (r=0), mas sim uma relação quadrática e indica que a produção de morangos é alta quando as temperaturas variam entre 13 e 23º Celsius.

Somente fazendo o gráfico que identificamos que existe uma correlação não linear entre as variáveis.



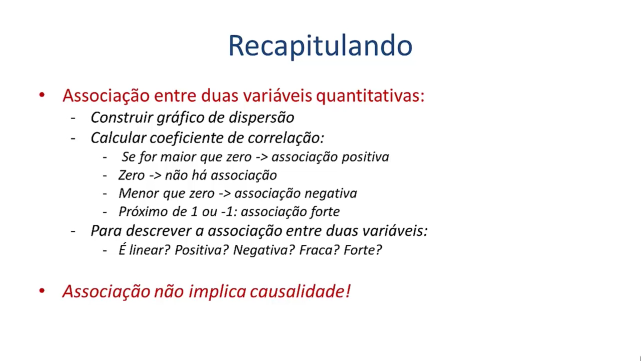
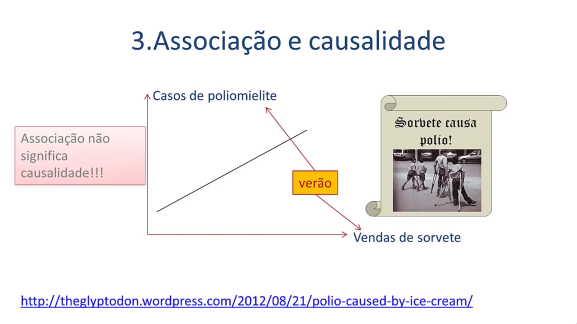
1. Associação e Causalidade

Analisando o exemplo abaixo, podemos notar que existe uma correlação linear entre as variáveis Quantidade de casos de poliomielite e Vendas de sorvete; mas será que existe mesmo esta correlação?

Por isto que devemos entender os dados.

Pois neste caso, não analisaram uma variável marcante que impacta diretamente estas 2 variáveis, a estação e a incidência do verão.

Portanto a associação entre estas 2 variáveis, não significa que uma cause a outra.



**Medidas de Correlação de Pearson e Spearman**

A medida mais popular para calcular a correlação de variáveis é o **Coeficiente de Pearson**.

Mas antes de entrar no detalhe destas correlações, vejamos o concito de **Esperança** ou Valor Esperado.

E(x) = x1.p(x1) + x2.p(x2) + ... + xN.p(xN) =>

Onde p(x) é a probabilidade da ocorrência do valor x.

Exemplo 1:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Resultado de X | 0 | 1 | 2 | 5 | 8 |
| Probabilidade de X | 1/8 | 1/4 | 1/2 | 1/4 | 1/4 |

Logo:

E(x) = 0\*1/8 + 1\*1/4 + 2\*1/2 + 5\*1/4 + 8\*1/4 =>E(x) = 0 + 1/4+1+ 5/4 + 2 => E(x)=4,5

Exemplo 2: Qual o valor esperado (esperança) da variável aleatória X, que representa o número de “coroas” obtido, quando uma moeda é jogada 3x?

Montando as possibilidades:

Cara = C e Coroa = K

C-C-C C-C-K C-K-C C-K-K

K-K-K K-K-C K-C-K K-C-C Total de Possibilidades = 8

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Resultado de X (quantidade de coras) | 0 | 1 | 2 | 3 |
| Probabilidade de X | 1/8 | 3/8 | 3/8 | 1/8 |

Logo:

E(x) = 0\*1/8 + 1\*3/8 + 2\*3/8 + 3\*1/8 =>E(x) = 1,5

**Coeficiente de Correlação de PEARSON**

Professor: <https://www.youtube.com/watch?v=qqRUsY2Fu0A>

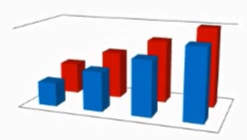
O coeficiente de correlação de Pearson (r) ou coeficiente de correlação produto-momento ou o r de Pearson mede o grau da correlação linear entre duas variáveis quantitativas. É um índice adimensional com valores situados ente -1,0 e 1.0 inclusive, que reflete a intensidade de uma relação linear entre dois conjuntos de dados.

Ela indica a força e direção do relacionamento entre estas 2 variáveis.

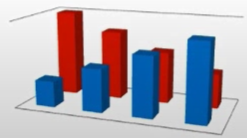
Este coeficiente, normalmente representado pela letra "r" ou “Ρ” assume apenas valores entre -1 e 1.

r= 1 Significa uma correlação perfeita positiva entre as duas variáveis.

Neste caso existe uma simetria ou covariância positiva, ou seja, quando uma variável cresce, a outra também cresce.

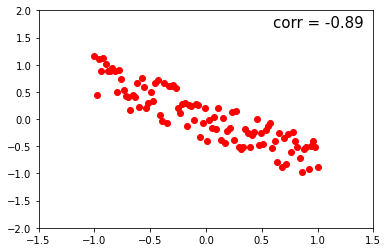
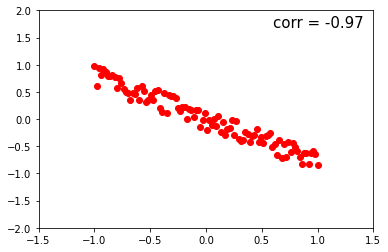
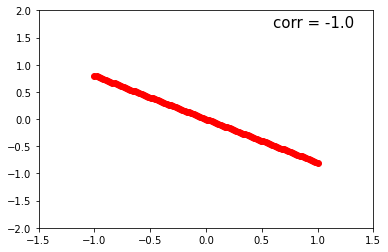


r= -1 Significa uma correlação negativa perfeita entre as duas variáveis - Isto é, se uma aumenta, a outra sempre diminui.

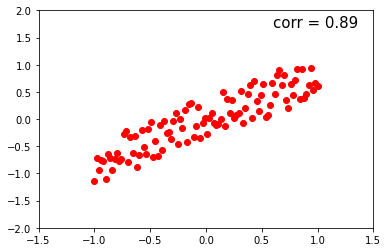
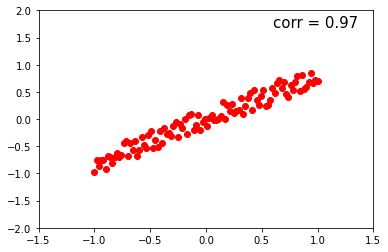


r= 0 Significa que as duas variáveis não dependem linearmente uma da outra. No entanto, pode existir uma outra dependência que seja "não linear". Assim, o resultado r=0 deve ser investigado por outros meios.

Negativo:



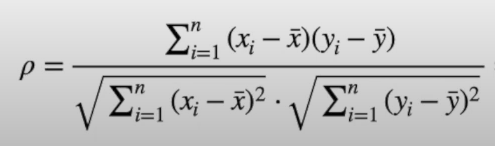
Positivo:



Neste caso, o coeficiente de Pearson é representado pela fórmula:



Ou seja, p = E(xy) – E(x)\*E(y) / DP(x)\*DP(y) OU p = cov(X,Y) / DP(x)\*DP(y)

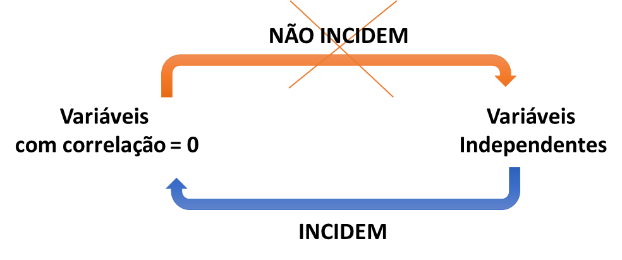


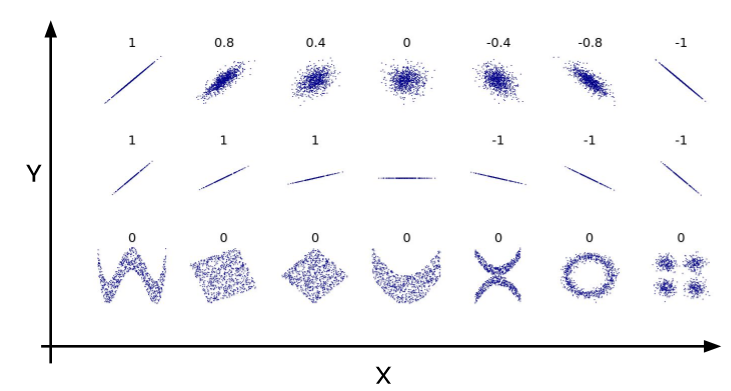
Onde a parte superior Numerador) da fórmula acima define se o valor do coeficiente é positivo ou negativo, e a parte debaixo (denominador) define a amplitude, no caso o quanto positivo ou negativo será este coeficiente de correlação.

Podemos reparar que o cálculo do numerador é afetado pela média, onde esta é altamente afetada pelos outliers. Portanto, o Coeficiente de Pearson também é afetado pelos outliers.

Portanto, podemos dizer quanto ao Coeficiente de Pearson:

* Possui as mesmas limitações da Média, ou seja, afetado pela existência dos outliers;
* Mede correlações lineares.
* Se tenho 2 variáveis que tem correlação igual à zero, isto não significa que elas são independentes, mas...
* ... se X e Y são independentes, então E(x,y) =E(x)\*E(y), logo a correlação é igual a zero





Repare que na 3ª linha do gráfico acima, como o coeficiente de Pearson trata apenas correlações lineares, ele não consegue encontrar a relação nos gráficos apresentados, neste caso, r=0.

Resumo: Quando usar o Coeficiente de Pearson:

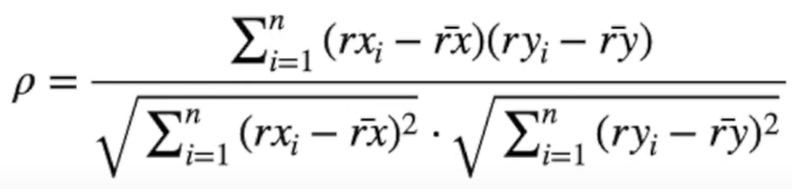
* Quando existe uma relação linear entre as variáveis estudadas;
* Quando X e Y forem distribuídas de forma simétrica, como em uma distribuição normal;
* Quando não existem outliers;
* Quando não for capaz de inferir independência.

Uma maneira de corrigir o problema dos outliers é o emprego do Coeficiente de Spearman.

**Coeficiente de Spearman**

O coeficiente de correlação de postos de Spearman, denominado pela letra grega ρ (rho), é uma medida de correlação não-paramétrica.

A correlação de Spearman é igual ao do coeficiente de Pearson, porém, ao invés de utilizar os valores das variáveis, usa-se a ordem das variáveis.



Por exemplo, se X={3,5,1,9,6} utilizamos rx={2,3,1,5,4} => que representa a posição de cada elemento de x na ordem crescente do conjunto de dados. Assim, ordenamos os valores de x e substituímos cada valor pela ordem e, com isto, deixamos de sofrer os impactos das ocorrências de outliers.

Observações quanto ao uso do Coeficiente de Spearman sobre o de Pearson::

* Vantagens:
* Avalia a relação monotônica (sempre crescente ou sempre decrescente ) entre 2 variáveis continuas ou ordinais; => \*Em uma relação monotônica, as variáveis tendem a mudar juntas mas não necessariamente a uma taxa constante .
* Não sofrem impactos de outliers e assimetrias na distribuição, já que consideramos a ordem e não os valores das variáveis;
* Enquanto a correlação de Pearson mede relações lineares, a de Spearman mede apenas relações monotônicas (lineares ou não-lineares), sendo mais geral que a de Pearson.
* Não requer que as variáveis sejam quantitativas; pode ser usado para as variáveis medidas no nível ordinal;
* Limitação:
* Quando temos muitas observações com a mesma ordem.

Ex: x={1,1,1,2} => rx={1,2,3,4} Assim, quando temos muitos valores iguais na amostra, o cálculo do coeficiente de Spearman pode ser prejudicado.

* **NOTE que Correlação não implica em Causalidade.** Os coeficientes de Pearson e Spearman só quantifica o quanto 2 variáveis estão correlacionadas, mas não podemos concluir que elas tenham uma relação causal.



* **Entropia de Shannon**

A entropia quantifica a quantidade de incerteza envolvida no valor de uma variável aleatória ou na saída de um processo aleatório.

Entropia de 1 variável: H(x) = log N

Entropia de 2 variáveis:

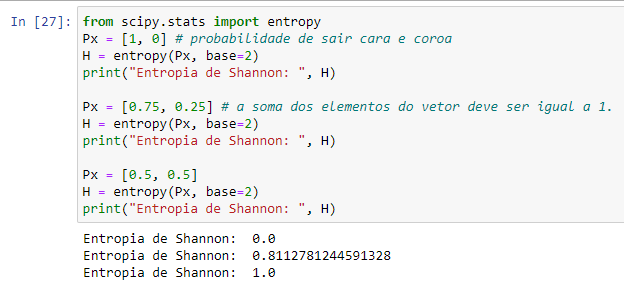
H(x,y) =

Por exemplo, se as **variáveis** representam X ={quente, frio} e Y ={seco,úmido} e p(quente, seco) = 1/2, p(quente, úmido) = 1/4, p(frio, seco) = 1/4 e p(frio, úmido) = 0.

Então H[X,Y] =−[1/2 log 1/2 + 1/4 log 1/4 + 1/4 log 1/4 +0log0] = 3/2 .

\*\*\* Ver material professor: “correlação.pdf”

Notem que essa entropia é máxima, visto que a distribuição uniforme é a que oferece maior informação e maior dificuldade na previsão. Para um vetor com um único valor, a entropia é mínima (ou seja, um dado com faces iguais). Vejam o exemplo abaixo para uma moeda cujas probabilidade de sair cara e coroa variam. Modifique os valores e veja o que acontece com a entropia.



No primeiro exemplo acima (px=[1,0]) uma moeda sempre cai como cara (1) e nunca como coroa (0), logo sua entropia é Zero, pois é 100% cara. Mas veja no 3º caso, onde tem 50% de chances de cara e 50% coroa (moeda justa), a entropia é 1. Ou seja, quanto maior a incerteza, maior o índice de entropia e quanto maior a certeza, menor a entropia.

**Portanto a Entropia mede o nível de certeza que eu tenho em uma predição. Quanto mais fácil a predição, menor é a entropia e quanto mais difícil a predição, maior a entropia.**

**Aula 2:**

* **Aprendizado Supervisionado**

No aprendizado supervisionado, o objetivo é ajustar um modelo preditivo a partir de um conjunto de exemplos de modo que o modelo seja capaz de prever dados não observados.

y = f(X, θ) + ϵ

y = Variável de saída, p.e. a classe de um paciente: doente, não doente;

X = Matriz, que representa os atributos/variáveis que utilizaremos na predição da classe y, p.e. nível de glicose, pressão arterial, batimento cardíaco etc;

Θ = Serão os parâmetros que iremos ajustar no modelo;

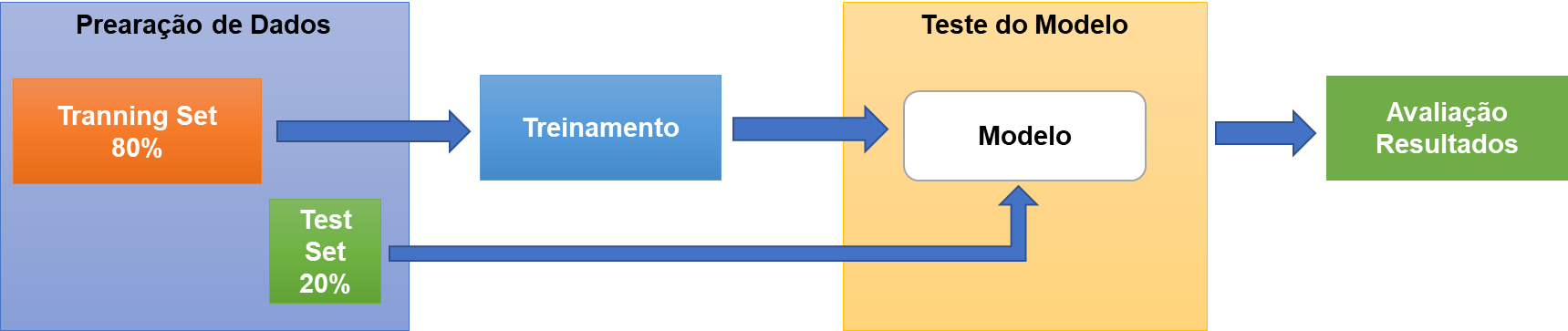
ϵ = Erro / Ruído que representa a informação que não está presente no modelo. São supostas variáveis que poderiam ajudar melhorar a precisão do modelo, mas que não foram utilizadas para gerar o modelo. P.e. na previsão da altura de um filho, não conseguiram medir o fator genético da mãe ou a qualidade da alimentação do filho ao longo da vida.

Essa função mapeia as entradas (x) nas saídas correspondentes (y), neste caso, o algoritmo preditivo aprende a aproximação (treinamento), que permite estimar valores de f para novos valores de X.

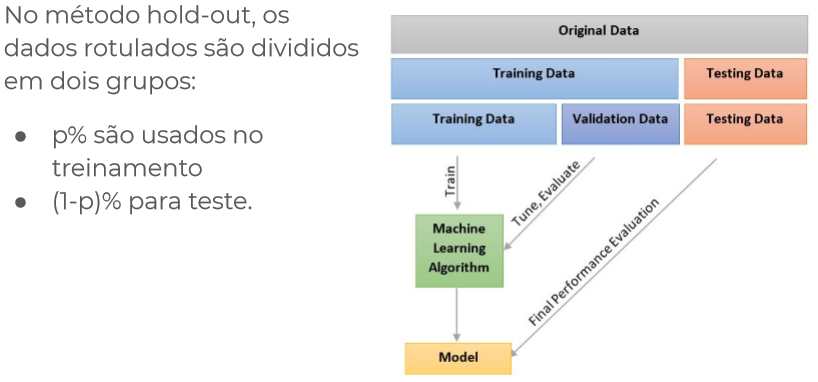
Se a classe a ser predita for um rótulo (lable, valor inteiro ou string), yiϵ {C1,C2,...,Cn}, definimos que é um problema de **Classificação.**

Se a classe a ser predita for um valor real, yiϵ R, definimos que é um problema de **Regressão.**

**Etapas:**

****

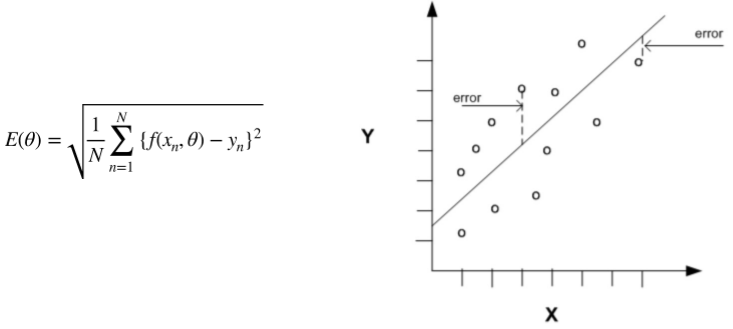
**Método hold-out:**



* **Erro (ou Risco) Quadrático Médio (EQM):**

Por definição, o EQM tem como principal objetivo encontrar a diferença média de um valor e o seu parâmetro inicial. De maneira mais prática, o seu uso é destinado a compreender um "erro de previsão".

Para encontrar o Erro Quadrático Médio, deve-se trabalhar com a soma de todos os resultados tidos como "erros" em relação à previsão inicial e, posteriormente, dividi-los pela quantidade de valores somados. Ou seja, apurar o quanto alguns resultados passam a se afastar de uma média aguardada inicialmente.



Exemplo: Imaginando que o índice Ibovespa seja a reta de referência. E existem o Fundo A e Fundo B; qual fundo se aproxima mais do índice Bovespa?

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mês 1 | Mês 2 | Mês 3 |
| Previsão Ibovespa | 1,2% | 2,1% | 1,8% |
| Fundo A | 3% | 6% | 1,8% |
| Fundo B | 1% | 2% | 1,8% |

E(fA) = [ (3-1,2)2 + (6-2,1)2 + (1,8 – 2)2 ] / 3 => [ 3,24 + 15,21 + 0,04 ] / 3 => 6,16

E(fB) = [ (1-1,2)2 + (2-2,1)2 + (1,8 – 2)2 ] / 3 => [ 0,04 + 0,01 + 0,04 ] / 3 => 0,03

Portanto, como o erro é menor no fundo B, podemos dizer que este se aproxima mais do índice Bovespa do que o funcho A.

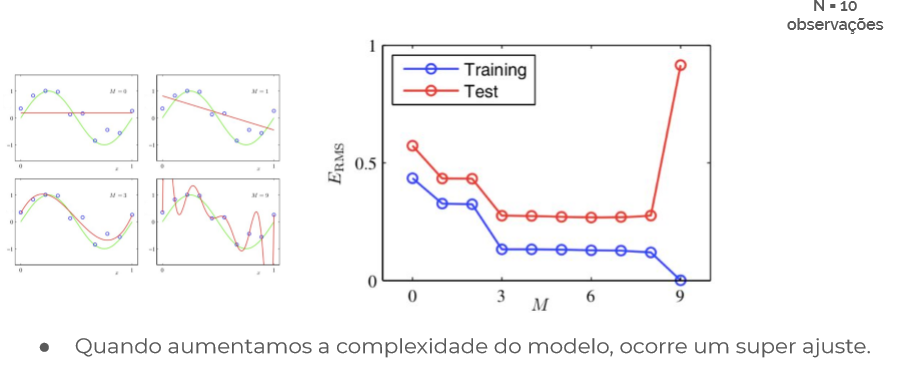
* **Overfitting**

Quando estamos analisando o melhor modelo para atender um determinado problema, buscamos ajustar o valor de θ (pesos/parâmetros atribuídos no modelo) para que minimizem a função custo/erro. Neste caso, devemos fazer a seguinte pergunta:

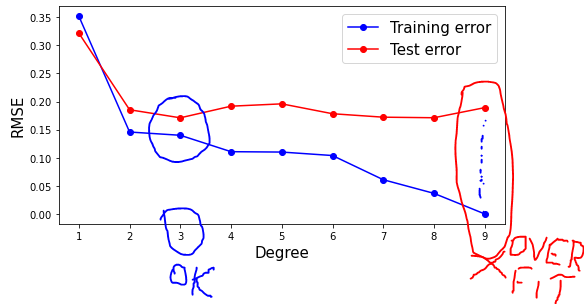
* + Quais são os melhores valores dos parâmetros do modelo (θ) que permitam generalizar e prever dados desconhecidos com precisão?

Porém, se o modelo estiver tão ajustado aos dados de treinamento, de forma que cheguem ao erro muito próximo de zero, é bem possível que este modelo esteja causando **overfitting**.

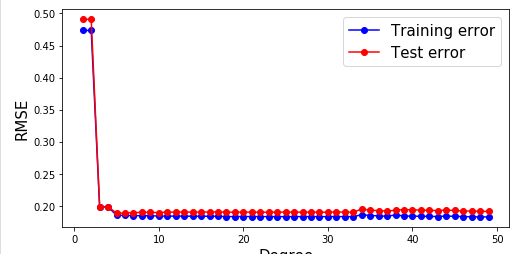
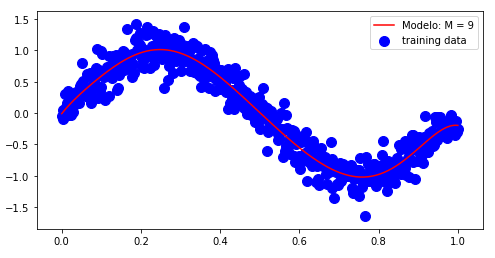
Neste caso, o overfitting ocorre quando um modelo se ajusta muito bem ao conjunto de dados anteriormente observado (dados em que foi treinado), mas se mostra ineﬁcaz para prever novos resultados. Ou seja, o modelo está “viciado” nos dados de treinamento e é incapaz de prever com precisão situações diferentes das que foi treinado. Ver exemplos aula do professor.



Repare que no gráfico abaixo, caso em que temos apenas 10 registros no conjunto de dados de treinamento e 20 no de teste, a diferença do erro encontrado entre os dados de treinamento e de teste se mantém estável e próximos até o polinômio de grau 3, o que é bom. Porém, a partir do grau 4 a distância entre os erros cresce, indicando que, já no grau 9, apesar do erro ser baixo no treinamento, ao aplicar o modelo nos dados de teste o erro foi bem maior do que o de treinamento; gerando assim o overfit. Assim, o modelo que melhor atende ao caso, é o de grau 3.



Porém, quando aplicamos o mesmo modelo em uma amostra maior, no caso 500 registrosno conjunto de dados de treinamento e 500 no de teste, o cenário muda. A curva resultante, mesmo utilizando polinômio de grau 9, o overfitting deixa de existir, se ajustando muito bem aos dados.



Portanto, o overfitting está relacionado ao número de observações e o número de atributos que eu tenho. Se tivermos um desbalanceamento muito grande entre poucas observações e muitos atributos, a chance de ter overfitting é muito alta.

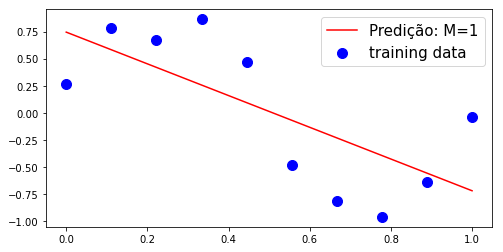
* **Viés-variância (bias-variance tradeoff)**

Conforme aumentamos a complexidade do modelo, buscando melhor precisão em seus resultados, passando-o de uma regressão linear para uma regressão polinomial de grau 2, 3, 4 em diante, as chances de overfitting aumentam. Portanto temos que fazer uma análise de “custo x benefício” de maneira que o modelo não fique tão simples (baixa performance/underfitting) e nem que causemos overfitting.

Então temos o Viés-Variância que trata este equilíbrio entre underfitting e overfitting.

**Underfitting:**

* + Alto Viés = Modelo muito simples, que não consegue prever nem no conjunto de treinamento. Alto erro.
  + Baixa Variância, sem muita variância no erro.

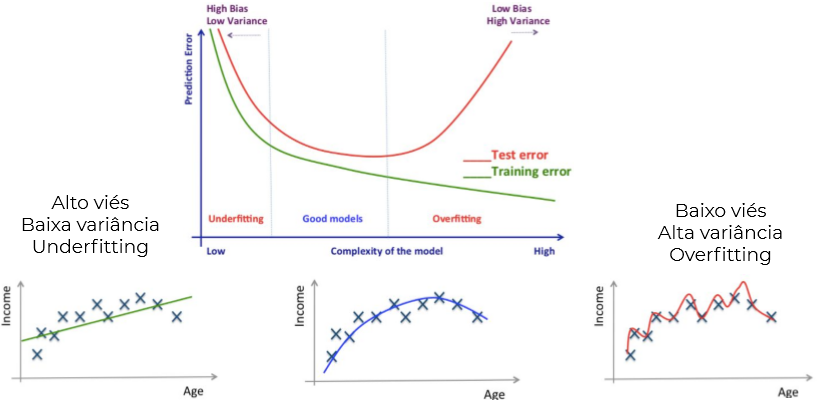


Underfitting

**Overfitting:**

* + Baixo Viés = Modelo mais complexo, com baixo erro.
  + Alta Variância, pois toda vez que colocar um novo elemento o erro vai variar muito.

Neste caso, o melhor modelo é uma zona intermediária, de equilíbrio, que não tenda a nenhum destes 2 lados.



Nesta avaliação, podemos perceber que, quando executamos as boas práticas de preparação dos dados, há uma relação entre a quantidade de dados e a complexidade do modelo que:

* + Quando temos um **grande volume de dados** para treinar e testar nosso modelo, **menos eu preciso conhecer**sobre o modelo e seus dados.

Neste caso, na prática, eu sei que um modelo de redes neurais vai atender a acurácia desejada.

* + Quando temos um **pequeno volume de dados** para treinar e testar nosso modelo, **mais eu preciso conhecer** sobre o modelo e seus dados.

Já neste caso, precisamos ser mais cuidadosos e utilizar, por exemplo, um modelo de árvores de classificação, para entender bem o que está acontecendo com os dados.

Observações na escolha do modelo:

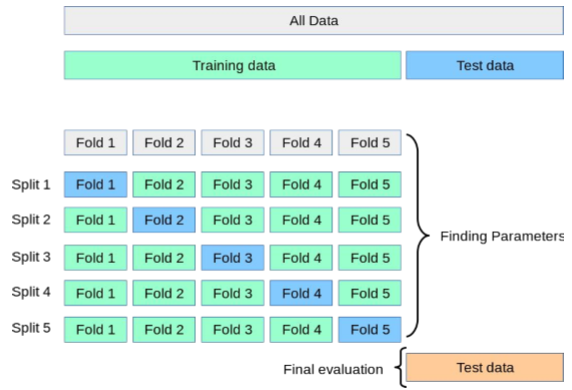
* Há relação entre o viés e variância e isso inﬂuencia na escolha do modelo.
* A escolha do melhor modelo envolve a redução da variância e viés.
* Infelizmente: não há nenhum método cientíﬁco padrão para isso.
* Como escolher complexidade ótima e conseguir erro mínimo no conjunto de teste?
* Erro no treinamento não é uma boa estimativa do erro no conjunto de teste.
* Podemos usar **validação cruzada** para definir o melhor modelo.
* **Validação Cruzada (Cross Validation)**

A validação cruzada é utilizada para reduzir a variância de resultados, permitindo, assim, que obtenhamos um modelo mais ajustado.

O método de validação cruzada é usado na escolha do modelo. Além disso, podemos usar o método para escolher o melhor conjunto de parâmetros de um classificador ou regressor. Notem que os dados são divididos em dois conjuntos, um de teste e outro de treinamento. Aplicamos validação cruzada ao conjunto de treinamento, para posterior avaliação do modelo no conjunto de teste.

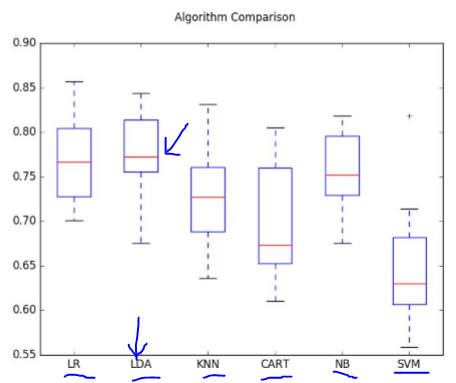
Trata-se de uma técnica em que:

1. Pegamos o conjunto de treinamento e o dividimos em partes, p.e. 5 partes (Folds);
2. Pegamos 4 destes Folds (2,3,4 e 5) para treinar o modelo e o Fold 1 será utilizado para validar e definir qual foi o erro.
3. Em seguida repetimos o mesmo processo, porém utilizando os Folds 1,3,4 e 5 para treinar o modelo e o Fold 2 para validar e definir qual foi o novo erro.
4. E assim sucessivamente até utilizarmos os Folds 1,2,3 e 4 para treinar e o Fold 5 para validar e verificar o erro.
5. Ao final, você pode comparar os resultados dos modelos,conseguindo identificar o modelo que melhor consegue reduzir a variância, e, assim, selecionar os melhores atributos e modelos.
6. Uma vez selecionado o melhor modelo treinado, fazemos uma validação final com todo o conjunto de dados de teste/validação para confirmar a acurácia desejada.

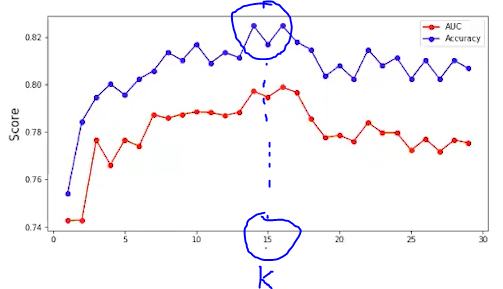


Portanto, na validação cruzada:

1. **Todos os dados rotulados são usados.**
2. Pode haver **grande variação** nos resultados de cada classiﬁcação, pois uma fração 1/k dos dados são colocadas no conjunto de teste.
3. A média de todas as classiﬁcações **reduz a variância** de todo o processo.
4. **Validação não serve para determinar a precisão do modelo**, mas para escolher os atributos e modelos.
5. Após a validação, usamos **todo o conjunto de dados** para ajustar o método de classiﬁcação ou regressão, para aplicar no conjunto de teste.
6. **Usamos validação para as seguintes tarefas:** 
   1. **Comparar modelos.**



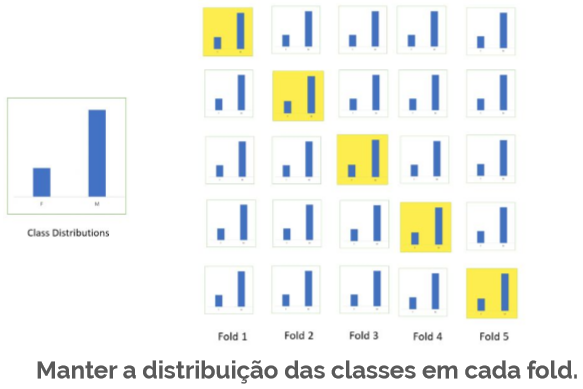
* 1. **Escolha dos parâmetros do modelo (ex. grau do polinômio).**



K = Quantidade de vizinhos que serão utilizados como parâmetros do modelo para fazer a classificação.

* **Validação Cruzada Estratificada**

É uma validação cruzada, porém na separação dos Folds, mantém-se a proporção dos dados de cada variável em cada Fold. Ex: Classe Sexo no exército. Como há muito mais homens do que mulheres, esta proporção de casos tem que ser mantidas em cada Fold.



Esta técnica sempre tem que ser usada quando há o desbalanceamento da variável, caso contrário, podem ocorrer Folds em que existam apenas homens na classe sexo, prejudicando o modelo.

**Aula 3:**

* **Tratamentos e transformação de dados**

1. **Problemas nos dados**

* **Ruídos nos dados;**
* **Amostragem**

Ex: diminuição na qualidade dos dados em cada amostra.

* **Dados Incompletos**

Como tratá-los? Eliminando-os? Substituindo-os pela média dos valores?....

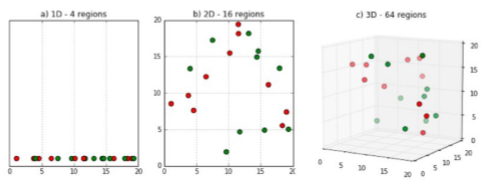
* **Maldição da Dimensionalidade**

Se dividirmos uma região do espaço em células regulares, o número células cresce exponencialmente com a dimensão do espaço.

Assim, o número de amostras deve crescer exponencialmente para garantir que nenhuma célula ﬁque vazia.

O poder preditivo de um método aumenta com o número de atributos, mas depois diminui (fenômeno Hughes).

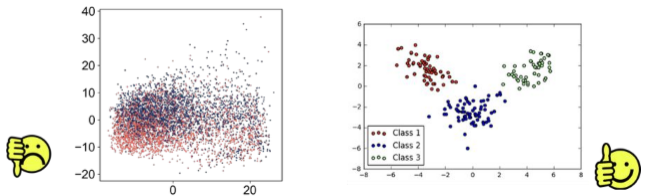
No exemplo abaixo, conforme aumentamos as dimensões 1D (4 regiões), 2D (16 regiões) e 3d (64 regiões), o número de regiões em que os pontos podem estar aumenta exponencialmente; onde teremos uma exparcidade muito grande, que prejudicará os métodos de aprendizado de máquina. Neste caso, ocorre o fenômeno de Hughes, onde ao aumentar a quantidade de atributos iremos melhorar a classificação, mas chegará a um ponto em que sua precisão começará a decrescer.



* **Outros problemas**, como:
  + Dados correlacionados – dados extremamente relacionados um com o outro como peso e altura;
  + Dados duplicados – a mesma entrada/registro possui os mesmos valores;
  + Combinação dados de diferentes tipos (nominais com racionais, etc) – como salários e nível de escolaridade;
  + Dados com diferentes escalas – p.e. dados variando de 0 a 1 e outro entre 0 e 1.000;
  + Dados irrelevantes para a análise – p.e. CPF;
  + Atributos que não contribuem para o aprendizado – p.e. cor do olho;
  + Dados desbalanceados – p.e. mais dados de uma classe do que de outra;
  + Número elevado de atributos (maldição da dimensionalidade);
  + Etc.

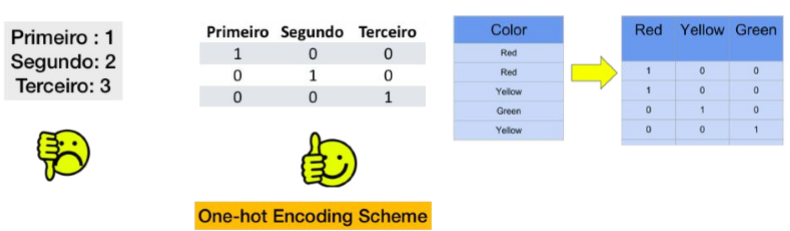
1. **Técnicas de tratamentos dos dados**

* **Eliminação Manual de Atributos**
  + Alguns atributos não possuem relação com o problema sendo solucionado;
  + Atributos que possuem o mesmo valor em todos os objetos/registros;
  + Atributos irrelevantes cuja identiﬁcação pode ser feita através de técnicas de seleção de atributos. Como a identificação de atributos onde suas classes estão sobrepostas, como no 1º exemplo abaixo, que acabarão causando ruídos no modelo. Ex: Idade x Profissões.



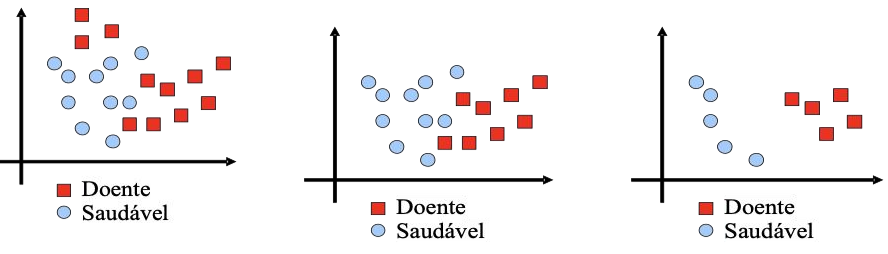
No exemplo da direita, as classes estão bem agrupadas, portanto são bons atributos para utilizarem.

* **Transformação de Valores**
  + Vários algoritmos de aprendizado de máquina têm diﬁculdades em usar os dados em seu formato original.
    - Ex. transformação de valores simbólicos para numéricos.



A técnica de **One-hot Encoding Scheme** é bastante útil quando formos considerar dados que são nominais e tivermos métodos que só aceitas atributos de tipo numérico.

* **Amostragem**
  + Algoritmos de aprendizado de máquina podem ter diﬁculdades em lidar com um grande volume de dados, como:
    - A saturação de memória do computador utilizado para treinamento de grandes volumes de dados;
    - Aumento do tempo computacional para ajustar os parâmetros do modelo; o que pode demandar muito tempo de treinamento. Por isto que, nestes casos utilizamos um número reduzido de amostras.
  + Contudo, quanto mais dados, maior tende a ser a acurácia do modelo.
  + **Nestes casos, o ideal é fazer uma análise de custo x benefício e buscar um balanço entre a eficiência computacional, descrita acima, e a acurácia do modelo. Para definir se precisa amostrar os dados e até quanto preciso amostrar.**
  + A distribuição das amostras também é um ponto importante de atenção, pois sua distribuição pode afetar os resultados dos modelo. Ex:



**Tipos de amostragens:**

* + **Amostragem Aleatória Simples**

Variações: com e sem reposição de exemplos (semelhantes quando tamanho da amostra é bem menor que o do conjunto original).

* + **Amostragem Estratiﬁcada**

Quando classes têm propriedades diferentes (ex. números de objetos diferentes).

Variações: manter o mesmo número de objetos para cada classe ou manter o número proporcional ao original.

* + **Amostragem progressiva**

Começa com amostra pequena e vai aumentando enquanto acurácia preditiva continuar a melhorar. Quando começar a piorar eu paro.

**Problema de Balanceamento na amostragem**

* + Em muitas aplicações de aprendizado de máquina, o número de objetos varia para as diferentes classes.
    - Ex. 80% dos pacientes que vão a um hospital estão na classe Doentes e, assim, possuo uma baixa amostra da classe Saudáveis. Sem um balanceamento das classes na amostra, o modelo tente a prevalecer a classe majoritária.
    - Problema na geração/coleta dos dados.
  + Vários algoritmos de aprendizado de máquina têm o desempenho prejudicado para dados muito desbalanceados, pois tendem a favorecer a classiﬁcação na classe majoritária.
  + Solução: balancear os dados. -> A técnica de Validação Cruzada Estratificada ajuda a mitigar este problema.
* **Balanceamento de Dados**
  + Acréscimo/eliminação de exemplos na classe minoritária/majoritária
    - **Acréscimo:**
      * P.e. em uma amostra com 80 Homens e 20 Mulheres, aumentar a proporção de mulheres chegando a uma proporção mais equilibrada, 80H e 80M.
      * Problema: Risco de objetos que não representam situações reais e overﬁtting
    - **Eliminação:** 
      * P.e. diminuir a amostra de homens para 20.
      * Problema: Risco de perda de objetos importantes e underﬁtting.
  + Usar custos de classiﬁcação diferentes para as classes.
    - Diﬁculdades: deﬁnição dos custos, incorporar custos em alguns algoritmos de AM.
    - Pode apresentar baixo desempenho quando muitos objetos da classe majoritária são semelhantes.
* **Limpeza dos Dados ----Parei aqui 15min**

Exemplos de problemas:

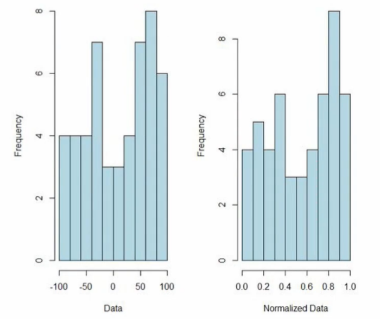
* Ruídos: erros ou valores diferentes do esperado.
  + Ex: Idade com valores negativos.
* Inconsistências: não combinam/contradizem valores de outros atributos no mesmo objeto.
  + Ex: pessoa com 2m pesando 10 Kg.
* Redundâncias:
  + Ex: objetos/atributos com mesmos valores. -> Que acaba “forçando” o modelo a aprender 2x aquela entrada, viciando-o.
* Dados incompletos: ausência de valores de atributos.
  + Ex: Atributo salário para crianças.
* **Dados incompletos: Soluções**
* Eliminar os objetos com valores ausentes. -> Caso não tenha o atributo; mas estará perdendo informação....
* Deﬁnir e preencher manualmente os valores ausentes. -> Ex: Alguns campos de salários vazios posso preenche-los com a média...
* Utilizar método/heurística para deﬁnir valores automaticamente.
* Empregar algoritmos de AM que lidam internamente com valores ausentes. -> Ex: No caso da falta de salário, crio o algoritmo para predizer o valor do salário e preencho o campo vazio com esta predição.
* **Atributos redundantes**
* Valores podem ser estimados a partir de pelo menos um dos demais atributos.
* Atributos com a mesma informação preditiva.
  + Ex. atributos idade e data de nascimento
  + Ex. atributos quantidade de vendas, valor por venda e venda total
* Atributo redundante pode supervalorizar um dado aspecto dos dados. -> podendo viciar a predição.
* Pode também tornar mais lento o processo de indução.

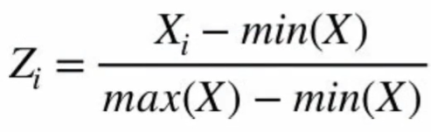
1. **Transformação de Atributos**

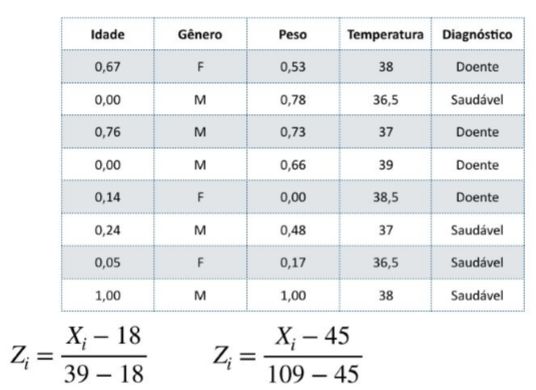
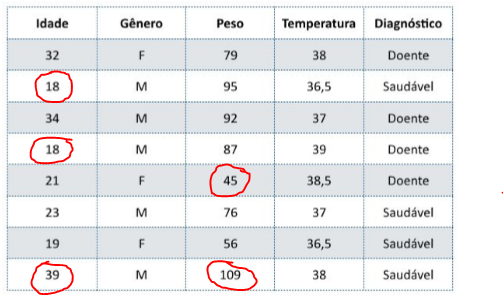
Algumas vezes é necessário transformar o valor de um atributo numérico em outro valor numérico.

* Por exemplo, quando o intervalo de valores são muito diferentes, levando a grande variação.
* Quando vários atributos estão em escalas diferentes. -> Ex: atributos em cm e outros em km.

As transformações (métodos) mais comuns são:

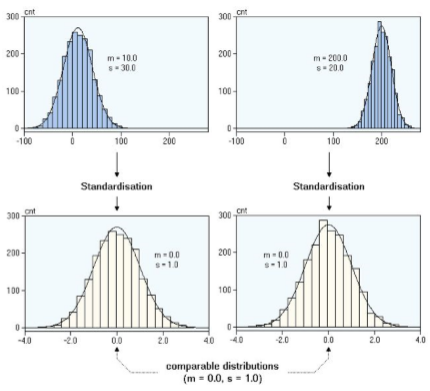
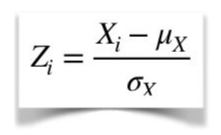
* Normalização (min-max scaling),
* Padronização,
* PCA,
* Softmax.
  1. **Normalização (min-max scaling)**
* Os dados serão ajustados de forma que o valor máximo seja igual a um e o menor, a zero.
* Usado em redes neurais.
* **Desvantagem: sensível a outliers.**





* 1. **Padronização (Z-score Normalization)**

Os dados ajustados apresentam média igual a zero e desvio padrão igual a um.



**Normalização e Padronização:**

<https://medium.com/data-hackers/normalizar-ou-padronizar-as-vari%C3%A1veis-3b619876ccc9>

* 1. **Outro tipo de transformação: TRADUÇÃO**
* Valor é traduzido por um mais facilmente manipulável
  + Ex. converter data de nascimento para idade
  + Ex. converter temperatura de F para C
  + Ex. localização por GPS para código postal.
  1. **Análise dos Componentes Principais (PCA)**
* Muitos atributos podem ser correlacionados, o que não contribui para a discriminação.
* Um número elevado de atributos pode levar à maldição da dimensionalidade.
* A simpliﬁcação dos dados, sem perder informações importantes, ajuda no processamento, pois reduz o tempo computacional e complexidade de algoritmos.
* PCA permite transformar os dados de modo a eliminar redundâncias e preservar informações importantes.

Muitos atributos são correlacionados, como, p.e., salário e nível educacional da pessoa. Portanto, como reduzir esta correlação entre os atributos?

Se eu tenho 100 atributos e quero trabalhar apenas com 5, como seleciono estes 5 atributos de maneira que a variância entre os dados seja mantida?

Com o PCA faremos uma transformação nos dados de forma a obter um novo conjunto de atributos, onde suas variáveis serão não correlacionadas e ortogonais (perpendiculares) uma a outra.

A análise dos componentes principais é um procedimento para reduzir a dimensão do espaço de variáveis através da transformação dos dados de modo a obter um conjunto de eixos ortogonais (não correlacionados) que capturam grande parte da variabilidade original dos dados.

Ou seja, queremos transformar os dados, sem perder muita informação, de forma em que estes dados não estejam correlacionados.



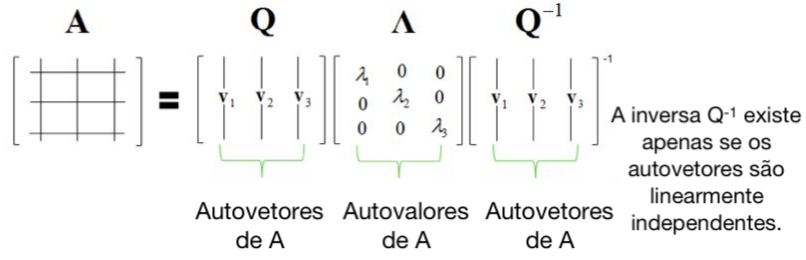
Onde o eixo 1 é o eixo principal, onde se encontra a maior variação dos dados, e o eixo 2, ortogonal ao eixo 1, será o 2º eixo de maior variação. Após a transformação, grande parte da variância estará representada em uma dimensão.

Uma vez que temos a matriz de covariância dos dados originais ( A = cov(X,Y) ), temos que obter uma outra matriz de covariância de forma que ela seja diagonal, que, por ser diagonal, tem a covariância entre as variáveis iguais a zero.

Lembrar que a diagonal de uma matriz de covariância é a própria variância, onde A=cov(X,X) é a variância de Xe A=cov(X,Y) é a covariância de X e Y.

E, assim, para **diagonalizarmos uma matriz** (encontrar os autovalores e autovetores), usamos o método chamado **decomposição em valores singulares (SVD decomposition).**

Na decomposição em valores singulares pegamos a matriz original de valores e achamos uma combinação de outras matrizes onde teremos uma matriz com os Autovalores de A (**Λ**) e as matrizes com os autovetores Q e Q-1.



**Passos:**

1. Centralizar os dados: Zi = Xi - µX
2. Calcular a matriz de covariância.
3. Calcular os autovalores e autovetores da matriz de covariância.
4. Ordenar os autovetores de acordo com o valor dos autovalores.
5. Obter os componentes: multiplicar os dados originais pelos principais (maiores) autovetores.

**PCA:**<https://www.youtube.com/watch?v=FgakZw6K1QQ>

**Aula 4:**

* **Técnicas de Agrupamento de Dados**

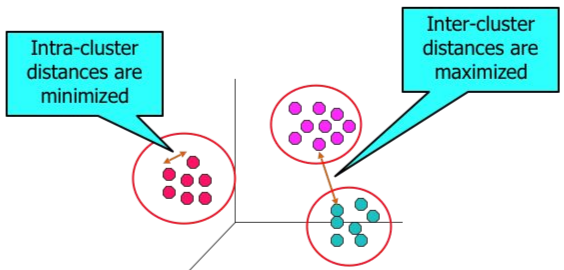
**Agrupamento de Dados**

A primeira questão a ser respondida no processo agrupamento de dados é em definir qual o critério em que separará os dados, em grupos/clusters, onde este critério irá interferir nos resultados.

Limitação: Não há uma deﬁnição clara sobre o signiﬁcado de “cluster” e como encontrá-los. Porém existem alguns métodos que nos ajudam na identificação destes clusters.

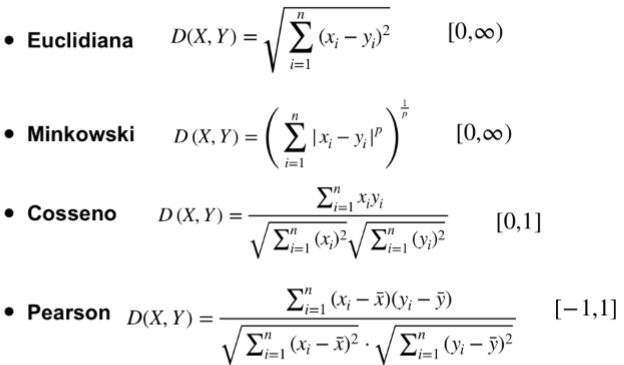
Para encontrarmos os clusters, precisamos alcançar 2 objetivos:

1. Minimizar a distância entre os objetos dentro do cluster, para identificarmos os que estão próximos/relacionados;
2. Maximizar a distância dos objetos entre os clusters, para segregar melhor os clusters e facilitar a separação entre eles.



Portanto, precisamos definir uma medida de proximidade/similaridade entre os objetos. Neste caso:

* Medida de **Similaridade**:
  + **d(Xi, Xi) é MÁXIMA**– onde distância entre o mesmo atributo é máxima.
  + Exemplo: Número de amigos compartilhados em uma rede social.
  + s(p, q) = 1 (ou máximo de similaridade) se p = q,
  + s(p, q) = s(q, p) para todo p e q, onde s(p, q) é a similaridade entre os objetos p e q.
* Medida de **Dissimilaridade**:
  + **d(Xi,Xi) = 0** – quanto mais próximos, mais perto de zero.
  + Exemplo: distância entre cidades (distância Euclidiana).
  + d(p, q) ≥ 0 para todo p e q, e d(p, q) = 0 se, e somente se, p = q,
  + d(p, q) = d(q,p) para todo p e q,
  + d(p, r) ≤ d(p, q) + d(q, r) para todo p, q e r, onde d(p, q) é a distância de dissimilaridade entre os pontos (objetos) p e q. -> Como se fosse um triângulo (3 obj.).
* **Métricas de Distância:**



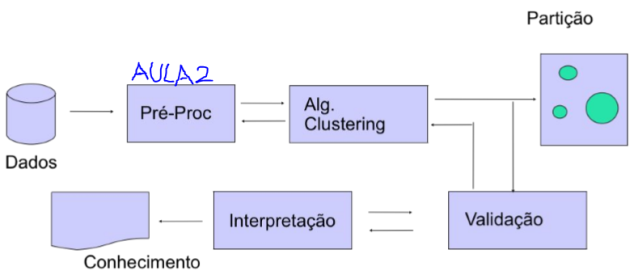
**Observações:**

* A distância de Minkowski é uma variação da distância Euclidiana (ex: se substituir p por 2, é a mesma fórmula);
* Nas distâncias de Minkowski e Euclidiana quanto mais similar, mais próximo de zero. **Portanto, são medidas de Dissimilaridade**.
* Já o Cosseno e Coeficiente de Pearson, quanto mais próximo de 1, mais similares são. **Portanto, são medidas de Similaridade.**
* A escolha de uso de cada uma depende do problema apresentado.

**Estágios dos Agrupamentos de Dados**

1. Seleção de atributos
   1. Os atributos devem ser selecionados de modo que ocorra o mínimo de redundância entre eles.
   2. Esta seleção, p.e., pode ser feita usando o PCA (Análise dos Componentes Principais), p.e., projetando os dados de 10 dimensões para 3 dimensões, já é uma seleção que você está fazendo, embora esteja transformando os dados.
2. Medida de Proximidade
   1. Esta medida deve quantiﬁcar o quão similar ou dissimilar são os objetos, para agrupar os mais semelhantes.
3. Critério de Clusterização
   1. Consiste de uma função custo ou algum tipo de regra.
   2. Se eu tenho, supostamente, 2 grupos, como vou juntá-los? Utilizar a distância mínima ou a máxima ou a média...
4. Algoritmo de Clusterização
   1. Consiste de um conjunto de passos para revelar a estrutura dos dados, baseados na medida de similaridade e no critério adotado.
5. Validação dos resultados.
6. Interpretação dos resultados.
   1. O que estes grupos representam, ou querem dizer, no mundo real, no caso de uso?

Passos:



1. **Algoritmo K-MEANS**

Amplamente usado na prática:

* Simplicidade;
* Interpretabilidade;
* Eﬁciência computacional.

Funcionamento do algoritmo:

1. Selecione k pontos como centróides iniciais.
   1. Inicialmente podemos fazê-lo aleatoriamente, mas se não der bons resultados, podemos empregar algumas técnicas para melhorar a localização do centroide.
2. Repita até que os centróides não mudem.
   1. Forme k grupos associado todos os pontos aos centróide mais próximos (através do cálculo das distâncias, já mencionado);
   2. Calcule o centróide de cada grupo obtido (Este cálculo pode ser, p.e., a média da distância entre os pontos).



**Ver exemplo dos notebooks.**

**Elbow Method (Método do Cotovelo)**

Um dos problemas iniciais que nos desafia é a definição de quantos clusters utilizar para o nosso caso.

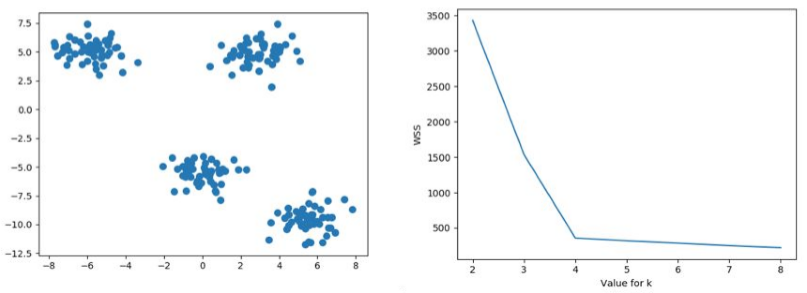
Portanto, o método Elbow é utilizado para encontrar o melhor valor de k, onde podemos usar a distância média dos pontos dentro de um cluster até o seu centróide (within-cluster sum of squares) para diferentes valores de **k (quantidades de clusters)**.

onde Ci é um grupo e Nc é o número de grupos.

O WSS é a distância total de cada ponto ao seu centroide. Se eu tenho, p.e., 2 clusters:

* Pego cada ponto do cluster 1 e calculo a distância até o centroide do cluster 1;
* Somo a distância de todos os pontos do cluster 1.
* Faço a mesma coisa para os pontos do cluster 2
* Somo o total da distância do cluster 1 e 2.

Assim, **WSS** pode ser entendida como uma medida de compactação. Que neste caso, calculamos o WSS com 2 clusters, depois com 3 clusters, 4 e assim em diante até identificarmos a menor quantidade de clusters para atender o caso em questão. Ver figura abaixo:

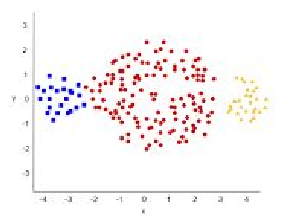
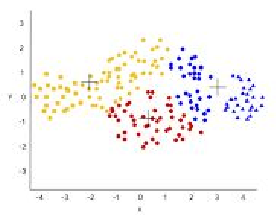


No caso acima, a distância quadrática média tende a descer até que, a partir de 4 clusters, há uma modificação na tendência de queda na distância, formando o “cotovelo”, nos mostrando, assim, o número ideal de clusters para atender o caso, que são 4 clusters.

**Observações / Limitações do KMEANS:**

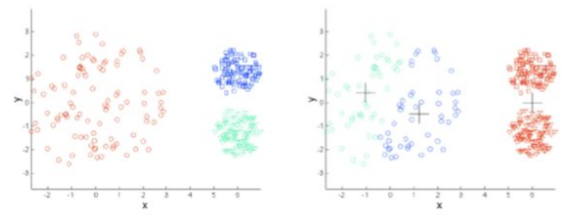
* Na Inicialização do algoritmo:
* O algoritmo é sensível à posição inicial das sementes.
* Importante rodar o algoritmo diversas vezes para obter resultados signiﬁcativos.
* É bastante susceptível a problemas quando clusters são de diferentes tamanhos.

**Antes Depois**

* É bastante susceptível a problemas quando clusters são de diferentes densidades.

**Antes Depois**



* É bastante susceptível a problemas quando clusters são de diferentes formatos (em geral não globulares). Tende a não identificar clusters em “meia lua”, **pois tende a identificar clusters esféricos**, principalmente quando utilizamos a distância Euclidiana (menor distância do centroide).

**Antes Depois**



* [K-means](http://en.wikipedia.org/wiki/K_means) é um algoritmo de agrupamento que tenta particionar um conjunto de pontos em conjuntos K (agrupamentos), de modo que os pontos em cada agrupamento tendem a estar próximos um do outro. Não é supervisionado porque os pontos não têm classificação externa.

[K vizinhos mais próximos](http://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm) é um algoritmo de classificação (ou regressão) que, para determinar a classificação de um ponto, combina a classificação dos K ​​pontos mais próximos. É supervisionado porque você está tentando classificar um ponto com base na classificação conhecida de outros pontos.

1. **Agrupamento Hierárquico**

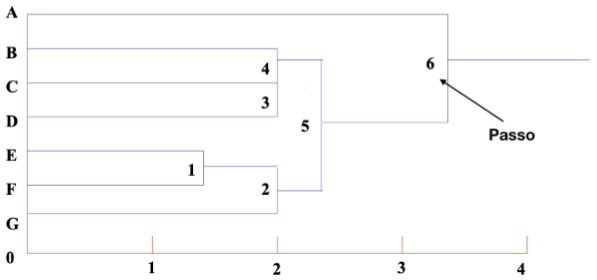
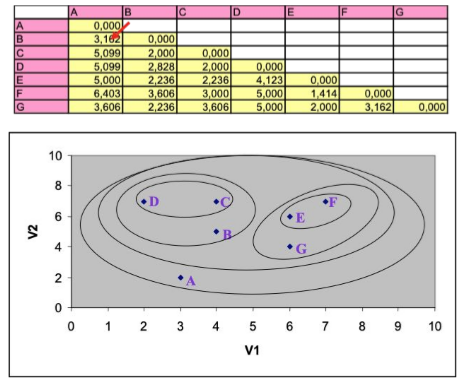
Trata-se do agrupamento em hierarquia por proximidade, no caso, representado por um **Dendograma**.

Um algoritmo de agrupamento hierárquico gera uma estrutura aninhada (árvore) de X, H={H1, H2, ..., HQ} (K ≤ N), tal que:

* Ci ∈ Hm e Cj ∈ Hl (m > l), implica que:
  1. Ci ⊆ Cj
  2. Ci ∩ Cj = ∅ (i,j,m,l=1,...,Q e i ≠ j)



* Tratamento:
  1. Calcula-se a distância entre os pontos;
  2. Repita: Busca o menor valor de distância e agrupa os valores, até relacionar todos os valores em clusters.



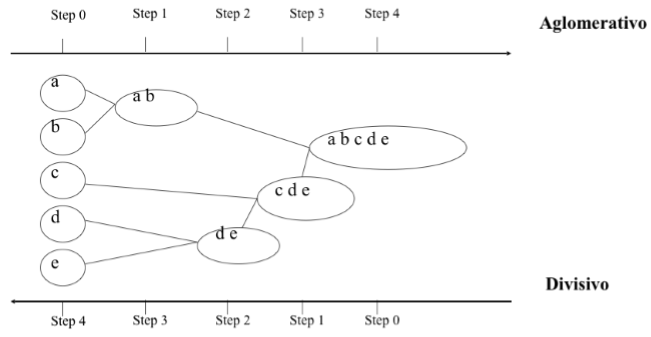
**DENDOGRAMA**

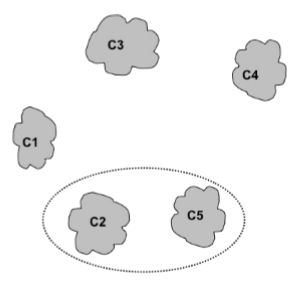
* Este é um procedimento Hierárquico e Aglomerativo.
* Ver exemplo da apresentação da aula e vídeo da aula (16:00min).

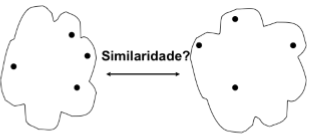
O agrupamento hierárquico nós podemos fazer em 2 diferentes níveis:

1. **Processo Aglomerativo**: Começando a partir das observações separadas e vai agrupando-as ao longo do processo;
2. **Processo Divisivo**: Começando com todos os clusters definidos/conectados e vai separando os objetos até todos estarem separados.

**\* Em ambos os casos, teremos a mesma estrutura de árvore.**



**Questão importante:** Após calculamos as distâncias entre os pontos, montamos a nossa matriz de proximidade e, depois de alguns passos, definimos os clusters. Após isto, questionamos: Como que eu atualizo a proximidade? Ou seja, se eu tenho o cluster 2 e 5, como eu calculo a distância entre eles?

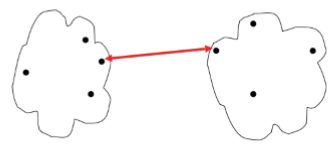


**Observar que:**Dependendo do método que você adotar o resultado pode ser muito diferente.

**Possibilidades:**

1. **MIN (single linkage)**

Pego todos os pontos dos 2 clusters, calculo todas as distâncias entre os pontos do cluster 1 e 2 e identifico a MENOR distância entre 1 ponto do cluster 1 e outro ponto do cluster 2. Esta será a distância entre os clusters.

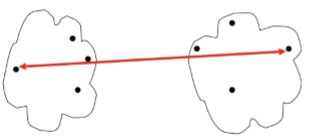


**Observação:** Este método é muito eficiente para aquelas “meia-luas” que o K-Means não conseguiu identificar. Pois o Single Linkage determina que a distância entre os clusters é a menor distância entre as observações. **Ouvir aula 4 do notebook, minuto 8.**



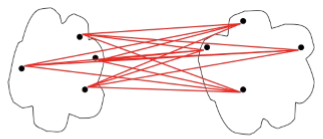
1. **MAX (complete linkage)**

Pego todos os pontos dos 2 clusters, calculo todas as distâncias entre os pontos do cluster 1 e 2 e identifico a MAIOR distância entre 1 ponto do cluster 1 e outro ponto do cluster 2. Esta será a distância entre os clusters.



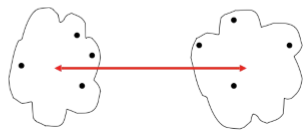
1. **Média dos grupos**

Pego todos os pontos dos 2 clusters, calculo todas as distâncias entre os pontos do cluster 1 e 2 e identifico a MÉDIA das distâncias entre os pontos do cluster 1 e os ponto do cluster 2. Esta será a distância entre os clusters.



1. **Distância entre centroides**

Calculo o centroide de cada cluster (que é o ponto médio do cluster) e a distância desejada será a distância entre estes centroides.

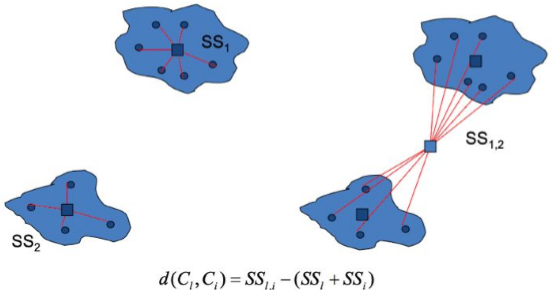


1. **Método de Weard’s usa erro quadrático médio.**

Este método visa minimizar/diminuir a perda de informação ao juntar 2 clusters. No caso, tenho 2 clusters e quero identificar se devo juntar os 2 ou não.

Verificação:

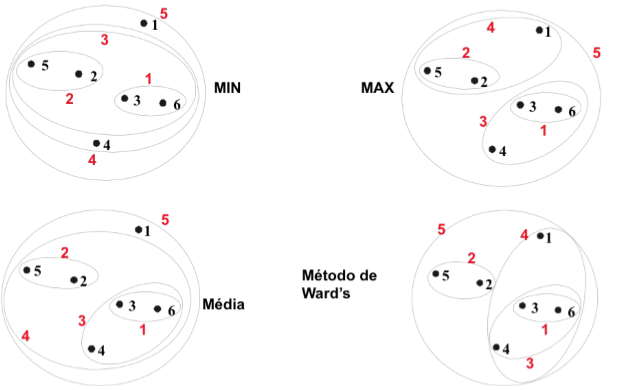
* 1. Para cada cluster, calculo a distância entre os pontos até o contróide;
  2. Crio um novo cluster, englobando os pontos dos 2 clusters originais, e calculo a distância de cada ponto, deste novo cluster, até o novo centroide.
  3. Verifico se a soma das distâncias do novo cluster é menor que a soma das distâncias dos clusters 1 e 2 originais. Se a distância do novo cluster for menos, decido agrupá-los, caso contrário não.



1. **Outros métodos que usam uma função objetivo.**

**Observações:**

* Dependendo do método que uso, teremos diferentes tipos de clusters. Ex:



* O método de Ward’s é o método mais utilizado na prática porque ele tenta deixar os clusters o mais compactos possíveis.

1. **Avaliando Agrupamentos**

Quão signiﬁcativo é o agrupamento?

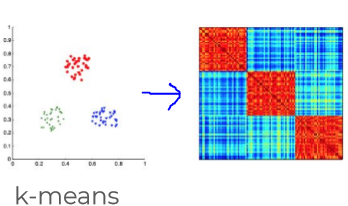
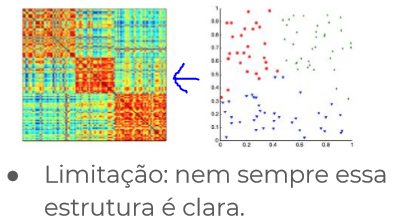
A avaliação de agrupamentos pode ser usada:

* Para evitar encontrar padrões em ruídos.
* Para comparar diferentes métodos de agrupamento.
* Para comparar clusters.

Medidas para avaliar agrupamentos são usadas em três casos:

1. **Índice externo:** Usado quando os rótulos/classes dos objetos são conhecidos e queremos avaliar se os clusters correspondem aos grupos originais. Ex: Se os casos de doentes estão nos cluster de doentes – Supervisionado.
   * Exemplo: Medidas de entropia.
2. **Índice interno:** Usado para avaliar um agrupamento sem usar informações externas, somente uma medida calculada. – Não supervisionado.
   * Exemplo: Soma do erro quadrático
3. **Índice relativo:** Usado para comprar agrupamentos ou grupos.
   * Exemplo: Índices internos ou externos são usados para esse ﬁm.

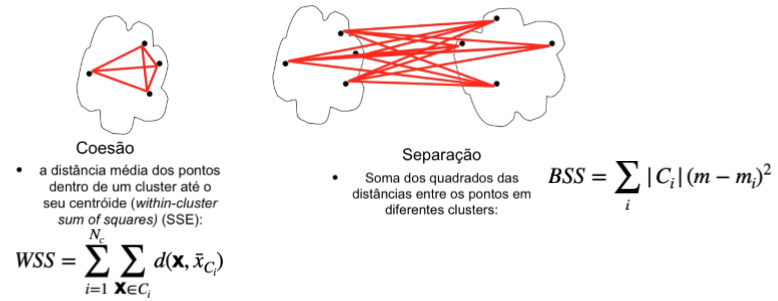
**Avaliação visual– Matriz de Similaridade:** ordena-se os objetos de acordo com os grupos e inspeciona-se visualmente. Cada elemento da matriz deﬁne a similaridade entre dois objetos (por exemplo, distância euclidiana entre dois vetores de atributos).

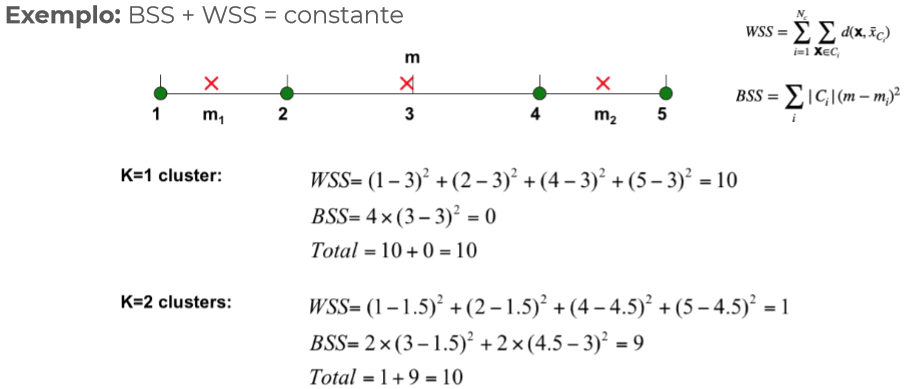
 

Cluster bem definido Cluster não bem definido

**Avaliação do Índice interno**: Usado para avaliar um agrupamento sem usar informações externas. Neste caso temos:

* A **Medida de Coesão (WSS)**, que me dirá o quão compacto é este cluster, sendo que quanto menor for este medida, mais compacto é o cluster.**Quanto menor melhor.**
* A **Medida de Separação (BSS)**, onde quanto maior for este valor, mais separados estão os clusters. **Quanto maior melhor.**

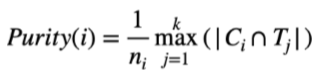




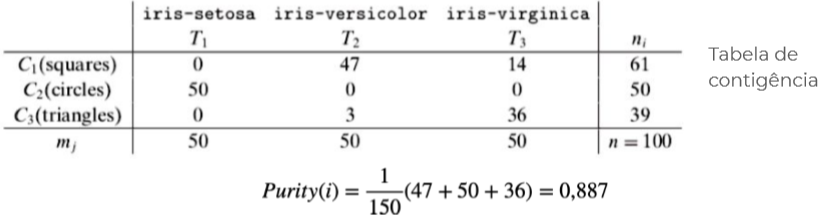
No exemplo acima, **utilizando 2 clusters (K=2) é melhor que 1 cluster (K=1)**, pois:

* WSS (Medida de Coesão) de K1 = 10 e K2=1 (quanto menor melhor) e
* BSS (Medida de Separação) de K1=0 e K2=9 (quanto maior melhor).

**Purity (Medida de Pureza)**: mede o quão “puro" é cada cluster, ou seja, o quanto um cluster é formado por apenas 1 tipo/classe. Quanto mais próximo de 1, maior o grau de pureza.



Onde Ci representa a classe obtida e Tj é a partição esperada.



No exemplo acima, como se pega apenas o máximo de cada linha, se todos Squares estivessem em T2 e todos Triangles estivessem em T3 o índice seria =1 (61+50+39 / 150).

**Importante:** A Medida de Pureza é utilizada como índice externo (já conheço as classes e posso comparar os resultados) e é importante para comparar métodos. P.e. se estou aplicando 2 métodos e quero verificar qual dos dois é o melhor, aplico a Medida de Pureza para verificar qual segmentou melhor as classes em seus devidos clusters, ou seja, verifico qual chega ao índice o “mais puro”.

**Índice externo: Outras medidas:**

* Maximum matching,
* f-Measure
* Normalized mutual information (NMI).
* Etc.

**Ver Aula 4 do Notebook, minuto 11 em diante**, para ver outras medidas de avaliação de acurácia de métodos, como o detalhamento do NMI.

**Aula 5:**

* **Modelos de Regressão**

Modelos de regressão são modelos matemáticos que relacionam o comportamento de uma variável Y com outra X. Neste caso, no processo de regressão, temos um conjunto de atributos (X) que serão usados para prever uma variável de saída (Y – resposta), que pode ser representado por:

y = f(X, θ) + ϵ

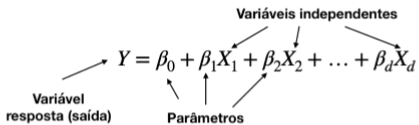
**Onde:**

* yϵ R
* f(X, θ)representa o modelo de regressão, que é uma função onde passaremos um conjunto de dados (p.e. altitude, humidade, nível pluviométrico, temperatura etc.) e que iremos predizer a variável y, que, p.e., pode ser o número de casos de dengue.
* A predição se dá em função do ajuste do parâmetros do modelo.
* ϵé o erro (ou ruído), que possui média 0 e variância σ2 , descrevendo o que não pode ser capturado pelo modelo. O erro nos ajuda a avaliar o quando meu modelo é adequado para prever a variável y.

Exemplo: Se possuímos um modelo que utiliza as variáveis mencionadas acima, mas que deixamos de tratar variáveis que são importantes ao modelo, como pressão atmosférica, o erro no modelo provavelmente será grande.

Modelos de regressão podem ser usados principalmente para duas tarefas:

* Prever dados desconhecidos a partir do modelo treinado.
* Determinar a importância de cada variável independente na previsão, ajudando a identificar as que mais impactam/influenciam na variável de resposta.



Quando a função f que relaciona duas variáveis é do tipo f (X) = a + b X temos o modelo de regressão simples. A variável X é a variável independente da equação enquanto Y = f (X) é a variável dependente das variações de X. O modelo de regressão é chamado de simples quando envolve uma relação causal entre duas variáveis. O modelo de regressão é multivariado quando envolve uma relação causal com mais de duas variáveis. Isto é, quando o comportamento de Y é explicado por mais de uma variável independe X1, X2, ....Xn.

Os modelos acima (simples ou multivariados) simulam relacionamentos entre as variáveis. Esse relacionamento poderá ser do tipo linear (equação da reta ou do plano) ou não linear (equação exponencial, geométrica, etc.). A análise de regressão compreende, portanto quatro tipos básicos de modelos:

* Linear simples;
* Linear multivariado ou múltipla;
* Não linear simples;
* Não linear multivariado.

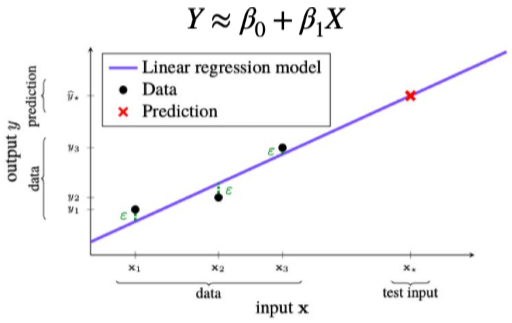
Para que serve determinar a relação entre duas variáveis?

1. Para realizar previsões sobre o comportamento futuro de algum fenômeno da realidade. Neste caso extrapola-se para o futuro as relações de causa-efeito – já observadas no passado – entre as variáveis. Pode-se, por exemplo, prever a população futura de uma cidade simulando a tendência de crescimento da população no passado.
2. Pesquisadores interessados em simular os efeitos sobre uma variável Y em decorrência de alterações introduzidas nos valores de uma variável X também usam este modelo. Por exemplo: de que modo a produtividade (Y) de uma área agrícola é alterada quando se aplica certa quantidade (X) de fertilizante sobre a terra.

No exemplo acima o pesquisador seleciona “n” pedaços de terra x1, x2, x3,....xn, aos quais são aplicadas quantidades definidas de fertilizante. Em seguida, medem-se as quantidades colhidas em cada pedaço de terra y1, y2, y3, ....yn, obtendo assim pares de valores (x1,y1) (x2,y2), ......(xn, yn) que podem ser plotados em um gráfico cartesiano chamado de **diagrama de dispersão**.

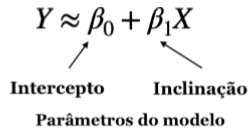
**Regressão Linear Simples**

No caso da Regressão Linear Simples, podemos ter, p.e., 2 variáveis, Temperatura e Número de Casos de Dengue, e tentar encontrar quais são os valores aplicáveis em ϐ0 e ϐ1 de forma que a reta, que é o nosso modelo e que está representada abaixo, seja ajustada de forma a minimizar o erro quadrático médio (ϵ), ou seja, a diferença entre o valor esperado/realizado e o valor obtido/predito.



Ver exemplode aplicação de RLS em um caso de Vendas de TVs nos slides 8 e 9 da aula 5.

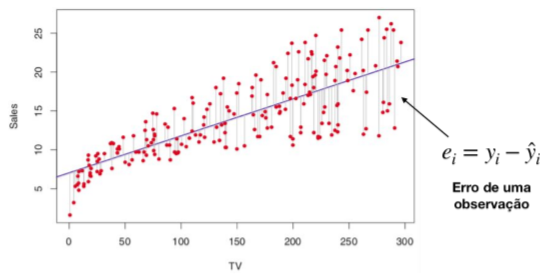
Neste exemplo:

 **=>** 

\* Intercepto: Local onde vai cruzar o eixo, quando x=0. Inclinação (da reta).

Na regressão linear simples, queremos encontrar os valores dos parâmetros tal que a reta obtida se aproxime o máximo possível dos pontos.

Mas como calcular/identificar estes parâmetros? Como ajustar a reta de forma a minimizar/diminuir o erro médio?





No caso, primeiramente calculamos o erro/resíduo dos pontos, que é deﬁnido pela diferença entre o valor real **i**(do conjunto de treinamento) e o valor predito**i**:



E o somamos seus quadrados (RSS – Soma dos Quadrados dos Resíduos/Erros dos Vizinhos):



OU

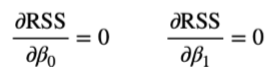
Como a predição da variável Y é dada por:



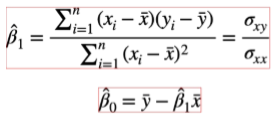
Ao substituir**e** e **i** obtemos:



E como queremos minimizar o erro, calculamos a derivada em beta, igualando-as a zero...

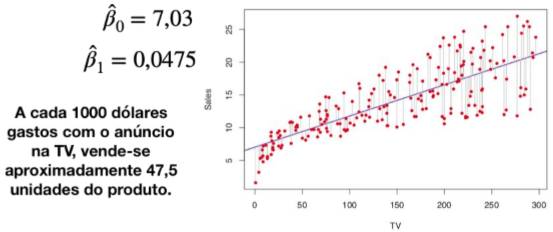


... obtendo:



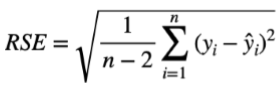
Esse é o método dos mínimos quadrados para estimar os coeﬁcientes do modelo de regressão.

E ao aplica-lo aos dados do exemplo citado obtemos:



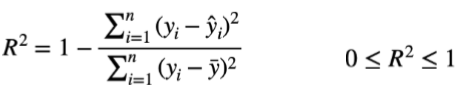
Para quantiﬁcar a acurácia do modelo, temos 2 medidas possíveis:

1. **O Erro Padrão Residual (Residual Standard Error)**, porém esta medida fornece o erro absoluto, medido em unidades de Y. **Ou seja, não consegue comparar 2 modelos, pois depende dos dados.**



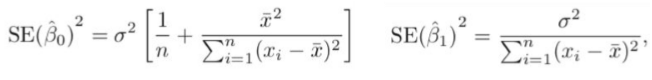
1. **Medida R2**: mede a proporção da variabilidade em Y que pode ser explicada a partir de X. Ou seja, uma vez que tenho os dados em X, quanto de variabilidade eu consigo predizer em Y. E quanto mais variabilidade eu consigo predizer, melhor é o meu modelo. Ou seja, **quanto mais R2 se aproximar de 1, melhor é o reajuste.** E se não explicar nada de variabilidade, R2 = 0.

**Portanto, a Medida R2é mais adequada do que o Erro Padrão Residual.**



* R2 é uma medida de relação linear entre X e Y.
* Valores de R2 próximo de 1 indicam que uma grande proporção da variabilidade dos dados são explicadas pelo modelo de regressão.
* Valores de R2 próximo de zero indicam que o modelo não explica muito da variabilidade, podendo ser que os dados não seguem uma relação linear ou erro é muito grande.
* No exemplo, R2 = 0,61, indicando que a variável TV explica apenas 2/3 da variabilidade nas vendas.

Com a medida R2 consigo verificar quão bom é o meu modelo. Mas daí surge uma nova pergunta: Será que este modelo é significativo ou não? Para responder isto, primeiro precisamos calcular o erro na estimação do 0 e 1...



... e com um intervalo de confiança de 95%, se 0 e 1 seguem uma distribuição normal, calculamos:

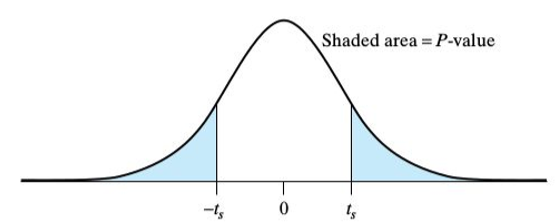
 ou seja, 1 + ou – (2\*Desvio Padrão de 1 )

ou seja, 0 + ou – (2\*Desvio Padrão de 0 )

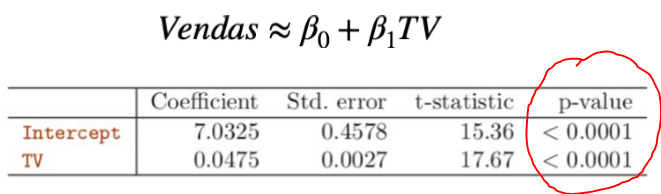
Aplicando estes conceitos em um teste de hipóteses, onde:

* H0: Não há relação entre X e Y (H0.: ϐ1 != 0)
* Ha: Há alguma relação entre X e Y (Ha: ϐ1 = 0)

Quanto mais próximo de zero for o valor p, maior é conﬁança em rejeitar H0,**ou seja, maior é a chance de eu ter uma relação entre X e Y, e assim predizer Y a partir de X**.



No exemplo anterior, rejeitamos a hipótese nula, ou seja, há uma relação entre o investimento em anúncios na TV e número de itens vendidos. Como o valor de p é muito próximo de zero, podemos concluir que o tempo de TV influencia o Volume de Vendas.



**Regressão Linear Múltipla**

A regressão múltipla envolve três ou mais variáveis, ou seja, uma única variável dependente, porém duas ou mais variáveis independentes (explicativas).

A finalidade das variáveis independentes adicionais é **melhorar a capacidade de predição em confronto com a regressão linear simples**. Mesmo quando estamos interessados no efeito de apenas uma das variáveis, é aconselhável incluir as outras capazes de afetar Y, efetuando uma análise de regressão múltipla, por 2 razões:

1. Para reduzir os resíduos. Reduzindo-se a variância residual (erro padrão da estimativa), aumenta a força dos testes de significância;
2. Para eliminar a tendenciosidade que poderia resultar se simplesmente ignorássemos uma variável que afeta Y substancialmente.

Uma estimativa é tendenciosa quando, por exemplo, numa pesquisa em que se deseja investigar a relação entre a aplicação de fertilizante e o volume de safra, atribuímos erroneamente ao fertilizante os efeitos do fertilizante mais a precipitação pluviométrica.

O ideal é obter o mais alto relacionamento explanatório com o mínimo de variáveis independentes, sobretudo em virtude do custo na obtenção de dados para muitas variáveis e também pela necessidade de observações adicionais para compensar a perda de graus de liberdade decorrente da introdução de mais variáveis independentes.

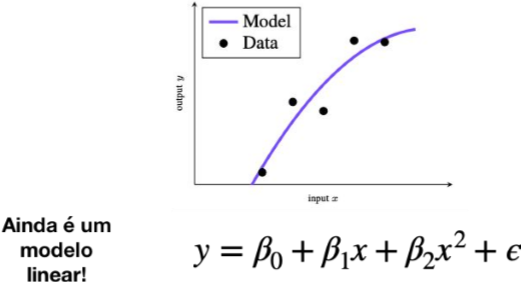
Suponha que temos d preditores distintos para a variável Y. Então, o modelo de regressão linear múltipla é:



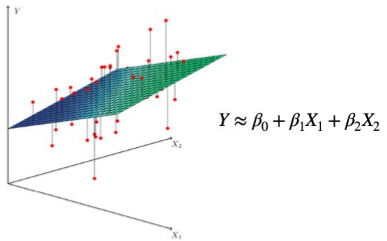
E adicionando novas variáveis ao exemplo anterior, podemos ter:



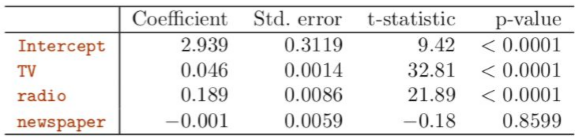
**Importante:** o modelo é linear **nos parâmetros e não nas variáveis**, que podem ser não lineares (ex. X2, sen(X), ln(X)). O modelo não precisa ter termos lineares em X, mas apenas nos parâmetros.



Enquanto uma regressão simples de duas variáveis resulta na equação de uma reta, um problema de três variáveis resulta um **plano**, e um problema de k variáveis resulta um **hiperplano**.



Ao aplicar o mesmo método que utilizamos anteriormente (RSS), onde objetivamos minimizar o erro, ao exemplo, podemos constatar que a variável rádio é a que mais contribui para aumentar o volume de vendas.



Neste caso, o maior coeficiente alcançado foi do Rádio, onde se aplicarmos 1.000 dólares, a divulgação através do Rádio venderia 189 e por TV, 46.

Note que a divulgação via Jornal não é significativa, pois o p-value foi muito alto. Mas porque ele não é significativo? Pois, provavelmente, o jornal é muito relacionado com as variáveis Radio e TV. Ou seja, se você usa as variáveis Radio e TV, a variável Jornal não está sendo útil para melhorar a acurácia do modelo.

Ver slides 25 a 32 e a vídeo aula 5 (Min 15:35 a 20:00).

**Regularização**

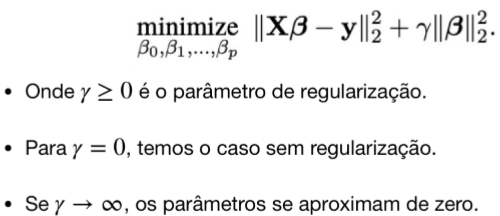
Após os ajustes nos modelos de regressão lineares ou múltiplos, surge a pergunta: Quão bom é este ajuste? Pois se fizermos um “super ajuste” no conjunto de treinamento, vamos incidir em overfitting...

A **Regularização** é uma das técnicas para evitar o surgimento deste overfitting.

De uma maneira bem direta, podemos entender regularização como a inserção de bias em um modelo. Ou em outras palavras, essa técnica desencoraja o ajuste excessivo dos dados, afim de diminuir a sua variância.

Dentro da regressão linear, Ridge e Lasso são formas de regularizarmos a nossa função através de penalidades. De forma simples, dentro de uma equação estatística dos dados, nós alteramos os fatores de forma a priorizar ou não certas parcelas da equação e, assim, evitamos ‘overfitting’ e melhoramos a qualidade de predição.

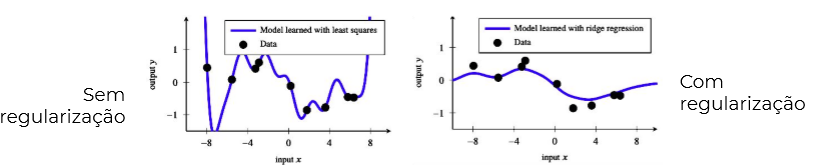
Um dos métodos mais populares é chamada **Ridge Regression (ou Tikhonov regularization)**. Ele é muito parecido com a regressão vista anteriormente, pois utiliza o Erro Quadrático Médio, porém adicionando um termo em que irá penalizar o “super ajuste” do modelo. Ou seja, se tivermos um super ajuste, alguns dos parâmetros de β possuirão valores muito altos, e, neste caso, este novo termo fará com que estes valores diminuam. E assim, quando enfrentarmos um problema de super ajuste, este novo termo fará com que esta curva “viciada” fique menos ajustada, menos “viciada”.



O Ridge Regression utiliza em seu novo termo a 2ª Norma, ou seja, a distância Euclidiana. A penalização consiste nos quadrados dos coeficientes, ao invés de seus módulos (como a seguir no LASSO).

Note que na equação acima, o valor de ϒ tem que ser > 0, senão o termo será = 0 e voltamos ao problema anterior. E se ϒ tender a infinito, o ajuste/parâmetros tenderão a zero...

Ou seja, agora também precisamos identificar qual o melhor valor de ϒ para ajustar os dados.



Ver slides 34 a 36 e a vídeo aula 5 (Min 20:00 a 22:50).

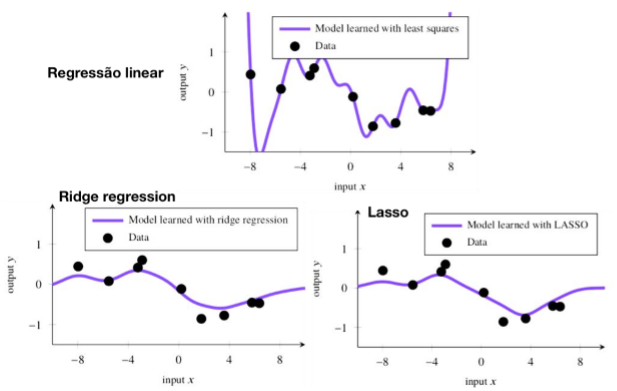
Outro método bastante popular é chamado **Least Absolute Shrinkage and Selection Operator (ou LASSO).**É semelhante ao método Ridge Regression, com a adição de um novo termo, porém este novo termo utiliza a norma N1, que é a soma dos módulos de β, ao invés da norma N2 do Ridge.

A regressão Ridge falha na parcimônia do modelo, pois ainda que as estimativas dos parâmetros sejam muito próximos de zero, elas nunca atingem exatamente esse valor. Assim, todos os pp preditores, ainda que com pouco peso, permanecem no modelo. A regressão Lasso é uma alternativa que contorna essa desvantagem. Os coeficientes Lasso, β̂Lassoλβ^λLasso, minimizam a quantidade.

Além de diminuir a variância do modelo, essa regularização tem uma outra importante aplicação em machine learning. Quando há múltiplas features altamente correlacionadas (ou seja, features que se comportam da mesma maneira) a regularização Lasso seleciona apenas uma dessas features e zera os coeficientes das outras, de forma a minimizar a penalização L1. Desse modo, dizemos que esse modelo realiza feature selection automaticamente, gerando vários coeficientes com peso zero, ou seja, que são ignorados pelo modelo. Isso facilita a interpretação do modelo, o que é uma enorme vantagem.

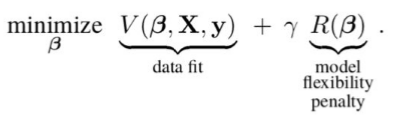
Notem que **os resultados obtidos por ridge regression e Lasso são diferentes**. Enquanto **Ridge Regression** mantém os valores dos parâmetros pequenos, **LASSO** tende a selecionar alguns valores para serem diferentes de zero, enquanto que outros são exatamente iguais a zero. Assim, LASSO pode ser usado para selecionar atributos.

No **Ridge Regression** tentaremos manter os betas pequenos, controlando-os para que nenhum deles fiquem muito alto e domine os demais. Enquanto que no **LASSO**, vou manter os parâmetros, que podem até ser grandes, mas alguns deles vou jogar para zero, p.e., se tiver 20 atributos podemos selecionar apenas 5, ou seja, você estará fazendo seleção de atributos.



Quando usamos cada um? O ideal é testar os métodos e verificar qual deles retorna o melhor erro possível.

Outros métodos podem ser deﬁnidos para regularização, como Elastic Net, HorseShoe etc. Mas de maneira geral, a função objetivo pode ser deﬁnida por:



* Temos um termo que descreve o quão bem o modelo se ajusta aos dados.
* Um termo que penaliza a complexidade do modelo.
* A regularização é obtida através de um balanço entre esses dois termos.

Portanto uma Regularização objetiva:

* Minimizar o nosso ajuste, a diferença entre o valor predito e o valor real;
* Utilizar uma função afim de minimizar o ajuste;
* E, assim, fazer um balanço entre não ocorrer um overfitting e não simplificar demais (underfitting), para alcançar o melhor modelo possível.