Problemas Clássicos da Computação

Redução de Dimensionalidade

Felipe Augusto Lima Reis

felipe.reis@ifmg.edu.br

Sumário

Introdução

Introdução

- Trabalhar com conjuntos de múltiplas dimensões oferecem alguns desafios:
  - A medida em que aumentam-se o número de dimensões, os algoritmos necessitam de mais treinamento;
  - Ao analisar, interpretar e plotar um conjunto de dados, estamos limitados a, no máximo, 3 dimensões
    - Muitas vezes, inclusive, optamos por plotar resultados em duas dimensões, para análise mais simples;
- Para diminuir essas desvantagens, existem técnicas de redução de dimensionalidade [Marsland, 2014].

Introdução

- Trabalhar com conjuntos de múltiplas dimensões oferecem alguns desafios:
  - A medida em que aumentam-se o número de dimensões, os algoritmos necessitam de mais treinamento;
  - Ao analisar, interpretar e plotar um conjunto de dados, estamos limitados a, no máximo, 3 dimensões
    - Muitas vezes, inclusive, optamos por plotar resultados em duas dimensões, para análise mais simples;
- Para diminuir essas desvantagens, existem técnicas de redução de dimensionalidade [Marsland, 2014].

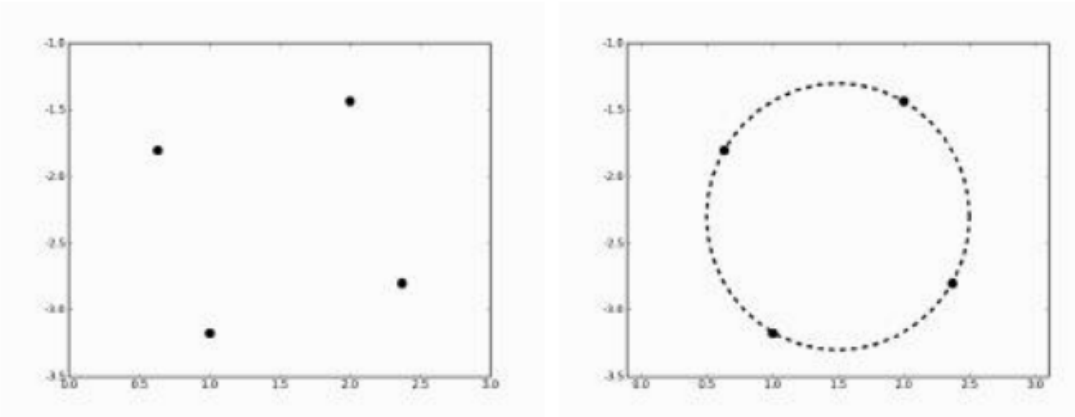
Introdução

- Redução de dimensionalidade possui as seguintes vantagens:
  - Auxilia na remoção de ruídos;
  - Auxilia a remoção de características irrelevantes ou redundantes;
  - Aumenta o desempenho dos algoritmos de treinamento;
    - Quanto menor o número de dimensões, mais rápido o algoritmo irá treinar;
    - Quanto maior o número de *features*, maior a quantidade de parâmetros a serem ajustados e maior a possibilidade de *overfitting*, devido aos ajustes;
  - Facilita o entendimento e uso do conjunto de dados [Marsland, 2014] [Richert and Coelho, 2013].

Introdução

- O entendimento correto do conjunto de dados permite melhor exploração do problema;
- Melhores decisões podem ser tomadas, com base em uma visualização adequada dos dados.

x	y
2.00	-1.43
2.37	-2.80
1.00	-3.17
0.63	-1.80



Fonte: [Marsland, 2014]

Classificação dos Métodos

- Os métodos de redução de dimensionalidade podem ser subdivididos em: [Marsland, 2014] [Richert and Coelho, 2013]
  - Seleção de *features*: analisa *features* que são úteis e correlacionadas ao problema
    - Demais *features* podem ser descartadas;
    - *Features* dependentes ou sobrepostas também podem ser descartadas;
  - Extração de *features*[1]: deriva *features* antigas em novas, por meio de transformações na base de dados
    - Podem ser aplicadas mudanças de coordenadas, deslocamento e rotação de eixos cartesianos (da representação);
  - Clusterização: agrupa *datapoints* similares de modo que uma menor quantidade de características sejam usadas.

[1] [Marsland, 2014] denomina essa técnica como Derivação de *features* (*feature derivation*).
[Richert and Coelho, 2013] subdivide os métodos em apenas seleção e extração de *features*.

Classificação dos Métodos

- Técnicas de seleção e extração de *features* não são excludentes
  - Em geral, utilizam-se primeiro técnicas seleção de *features*;
  - Em seguida, são usadas técnicas de extração para transformação e redução extra de dimensionalidade;
- Algoritmos podem ainda serem classificados em:[2]
  - Construtivos: inicia sem qualquer *feature* e adiciona-as iterativamente, avaliando o erro à medida em que novas *features* são adicionadas;
  - Destrutivos: *features* são removidas e, em seguida, é avaliado o comportamento da operação [Marsland, 2014].

[2] Os métodos aqui descritos podem também ser utilizados para a construção de árvores de decisão.

Conceitos Básicos de Estatística

Média, Moda e Mediana

- Média[3]: soma de todos os valores do conjunto de dados e dividido pelo número de elementos do conjunto

$$Me = \frac{x_1 + x_2 + x_3 + ... + x_n}{n}$$

- Moda: valor mais frequente de um conjunto de dados
  - Seja $X = \{1, 3, 4, 5, 4, 2, 0, 1, 4\}$. A moda $Mo = 4$, pois 4 é o numeral que mais aparece no conjunto de dados (3 vezes);

[3] Conceito referente à média aritmética.

Média, Moda e Mediana

- Mediana: representa o valor central de um conjunto
  - Deve-se ordenar o conjunto (ordem crescente ou decrescente);
  - Seja $X = \{1, 3, 4, 5, 4, 2, 0, 1, 4\}$. $X_0 = \{0, 1, 1, 2, 3, 4, 4, 4, 5\}$. A mediana $Md = 3$, pois 3 é o elemento central do conjunto.
  - Para conjuntos de cardinalidade par, obtém-se a média dos elementos centrais.

Tipos de Médias

- Média Aritmética: soma dos valores do conjunto de dados e dividido pela cardinalidade (visto previamente)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Média Ponderada: média aritmética com um coeficiente k_i ponderando os valores

$$MP = \frac{\sum_{i=1}^{n} k_i x_i}{\sum_{i=1}^{n} x_i}$$

A média aritmética é denotada frequentemente pelo símbolo μ.

Tipos de Médias

- Média Geométrica: n-ésima raiz do produto de todos os valores

$$MG = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

- Média Harmônica: quantidade de elementos no conjunto, divida pela soma do inverso dos elementos do conjunto
  - Utilizada para grandezas inversamente proporcionais, cálculo de velocidade média, circuitos e densidade.

$$MH = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}}$$

Valor Esperado (Esperança Matemática)

- Para uma variável aleatória discreta x_i, e suas respectivas probabilidades $p(x_i)$, o valor esperado pode ser calculado por:

$$E[X] = \sum_{i=1}^{n} x_i p(x_i)$$

- Se todos os eventos tiverem igual probabilidade, o valor esperado é a média aritmética.
- Exemplo:
  - Considere o valor médio de um dado, jogado aleatoriamente infinitas vezes:

$$E[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} \rightarrow E[X] = 3.5$$

Valor Esperado (Esperança Matemática)

- Para uma variável aleatória discreta x_i, e suas respectivas probabilidades $p(x_i)$, o valor esperado pode ser calculado por:

$$E[X] = \sum_{i=1}^{n} x_i p(x_i)$$

- Se todos os eventos tiverem igual probabilidade, o valor esperado é a média aritmética.
- Exemplo:
  - Considere o valor médio de um dado, jogado aleatoriamente infinitas vezes:

$$E[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} \rightarrow E[X] = 3.5$$

Valor Esperado (Esperança Matemática)

- Exemplo: Adaptado de [Marsland, 2014]
  - Considere uma raspadinha de R$1,00, com prêmio é R$10.000;
  - Serão vendidas 20.000 unidades e existirá um único ganhador;
  - O valor esperado do bilhete é dado por:

$$E[X] = -1 \times \frac{19999}{20000} + 9999 \times \frac{1}{20000} \rightarrow E[X] = -0.5$$

  - Valores:
    - -1: preço pago pela raspadinha;
    - 9999: prêmio, menos o custo da raspadinha;
    - 19999/20000: probabilidade de perder;
    - 1/20000: probabilidade de ganhar;
  - O valor esperado da raspadinha é -0.5, correspondente à perda, independentemente do evento que ocorrer.

Variância

- A variância, σ^2, de um conjunto é uma medida de quão dispersos estão os valores em relação ao valor esperado;
- É calculada como a soma das distâncias quadradas entre cada elemento e o valor esperado (média aritmética, $\bar{x}$)

$$var(x_i) = \sigma^2(x_i) = E((x_i - \bar{x})^2) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}$$

- Pode ser descrita como "a média do quadrado da distância de cada ponto até a média".

Desvio Padrão

- O desvio padrão, σ, corresponde à raiz quadrada da variância;
- Em estatística, indica uma medida de dispersão dos dados em torno de média amostral;
  - Baixo desvio padrão: indica que os pontos tendem a estar próximos da média ou do valor esperado;
  - Alto desvio padrão: indica que os pontos estão espalhados e/ou longe do valor esperado.

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Covariância

- A covariância mede como duas variáveis variam ao mesmo tempo;
- Pode ser também entendida como o grau de interdependência numérica[4] entre duas variáveis aleatórias;
- A covariância entre variáveis X e Y, com valores esperados $E[X] = \mu_X$ e $E[Y] = \mu_Y$, é dada por:

$$\boxed{cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]}$$

[4] O quão dependente as variáveis são (lembrar do conceito de variáveis dependentes e independentes).

Covariância

- Se duas variáveis forem totalmente independentes, sua covariância é igual a 0 (variáveis não correlacionadas);
- A covariância positiva indica que ambas as variáveis aumentam ou diminuem conjuntamente;
- A covariância negativa indica que a variação ocorre em sentidos opostos [Marsland, 2014].
- A covariância entre todos os pares de variáveis de um conjunto pode ser dado por uma matriz de covariância.

$$\Sigma = \begin{pmatrix} E[(\mathbf{x}_1 - \boldsymbol{\mu}_1)(\mathbf{x}_1 - \boldsymbol{\mu}_1)] & E[(\mathbf{x}_1 - \boldsymbol{\mu}_1)(\mathbf{x}_2 - \boldsymbol{\mu}_2)] & \dots & E[(\mathbf{x}_1 - \boldsymbol{\mu}_1)(\mathbf{x}_n - \boldsymbol{\mu}_n)] \\ E[(\mathbf{x}_2 - \boldsymbol{\mu}_2)(\mathbf{x}_1 - \boldsymbol{\mu}_1)] & E[(\mathbf{x}_2 - \boldsymbol{\mu}_2)(\mathbf{x}_2 - \boldsymbol{\mu}_2)] & \dots & E[(\mathbf{x}_2 - \boldsymbol{\mu}_2)(\mathbf{x}_n - \boldsymbol{\mu}_n)] \\ \dots & \dots & \dots & \dots \\ E[(\mathbf{x}_n - \boldsymbol{\mu}_n)(\mathbf{x}_1 - \boldsymbol{\mu}_1)] & E[(\mathbf{x}_n - \boldsymbol{\mu}_n)(\mathbf{x}_2 - \boldsymbol{\mu}_2)] & \dots & E[(\mathbf{x}_n - \boldsymbol{\mu}_n)(\mathbf{x}_n - \boldsymbol{\mu}_n)] \end{pmatrix}$$

Fonte: [Marsland, 2014]

Distribuição de Probabilidades

- A distribuição de probabilidades descreve as probabilidades de um evento ocorrer em meio a um conjunto de eventos possíveis [Marsland, 2014];
- Podem ser subdividas em:
  - Distribuições Discretas: distribuição quando o conjunto de eventos possíveis é contável[5];
  - Distribuições Contínuas[6]: distribuição quando o conjunto de eventos possíveis não é contável.

[5] Conjunto finito ou que tem a mesma cardinalidade do conjunto de inteiros positivos (pode ser mapeado no conjunto dos números inteiros positivos) [Rosen, 2019].

[6] Também denominada, por alguns autores, de distribuição absolutamente contínua (neste caso, é definido o conceito de distribuição singular - contínua, mas não absolutamente contínua).

Distribuição Normal

- A distribuição normal ou gaussiana é um tipo de probabilidade contínua para variáveis aleatórias;
- É definida como:

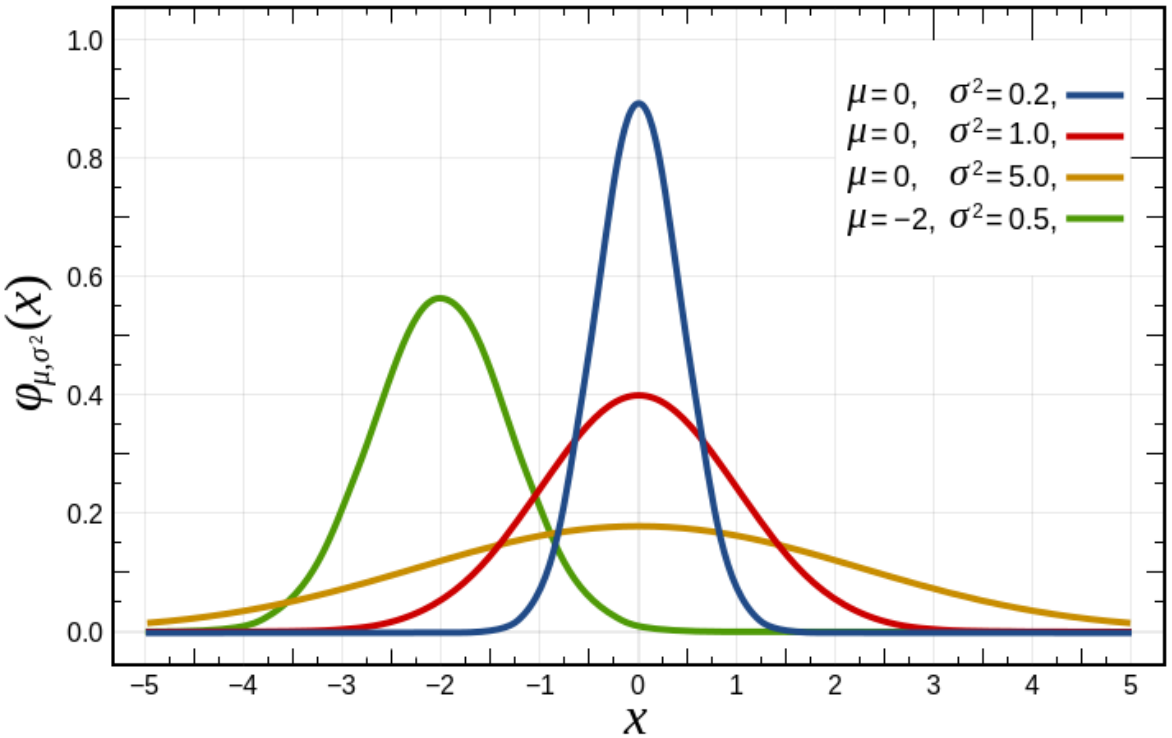
$$p(x) = \frac{1}{\sqrt{2\pi\sigma}}\ exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) = \frac{1}{\sqrt{2\pi\sigma}}\ e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

onde

- μ: média ou valor esperado;
- σ: desvio padrão.

Distribuição Normal

- A curva da distribuição normal é denominada, informalmente, Curva de Gauss ou Gaussiana.



Fonte: [Wikipedia contributors, 2021]

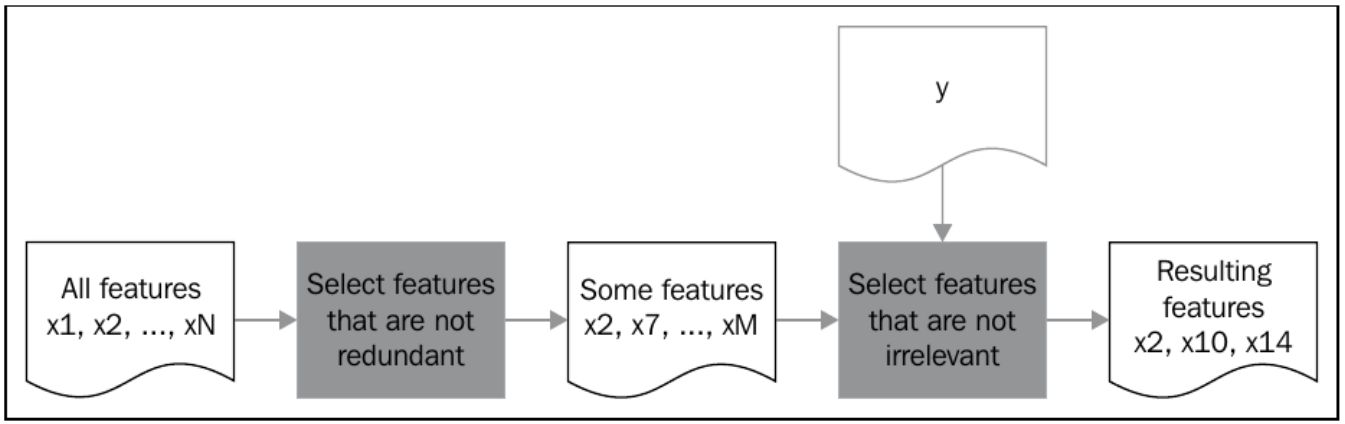
Distribuição Normal

- Distribuições normais são frequentemente utilizadas em ciências sociais e naturais, para representar valores aleatórios de distribuições desconhecidas;
- A distribuição normal também é utilizada no Teorema Central do Limite
  - O teorema indica que, sob algumas condições, a média de muitas amostras com média e variância finita converge para uma distribuição normal conforme o número de amostras aumenta [Marsland, 2014].

Seleção de Features

Seleção de Features

- O objetivo da seleção de *features* é garantir que:
  - As *features* sejam independentes umas das outras;
  - As *features* sejam relacionadas ao resultado a ser predito.
- Relações entre *features* podem ser detectadas usando análise estatística ou gráficos de dispersão (detecção visual) [Richert and Coelho, 2013].



Fonte: [Richert and Coelho, 2013]

Correlação entre Features

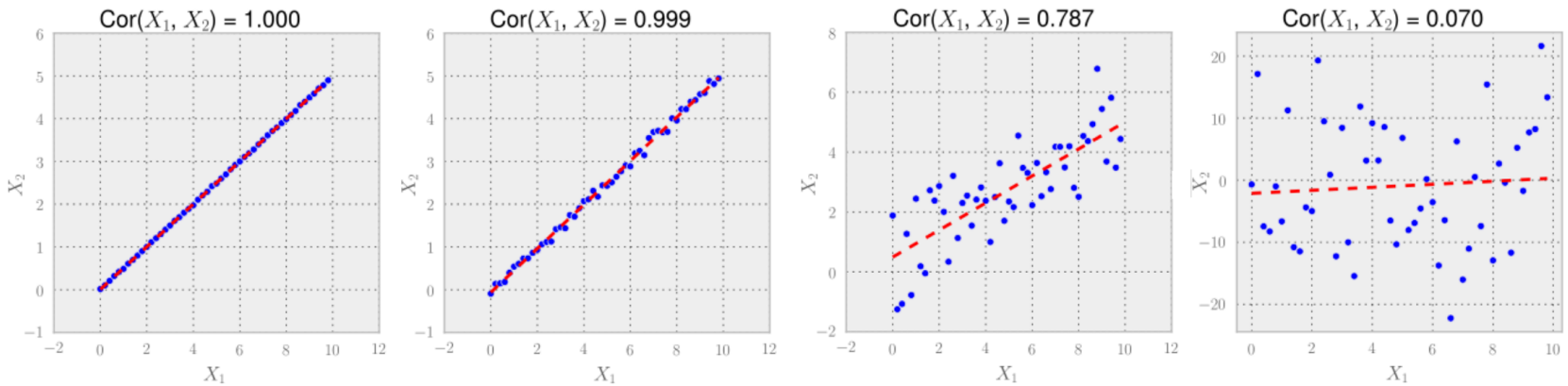
- A partir de um par de *features* é possível identificar a relação entre modelos usando linhas retas;
- Os graus de correlação e a potencial dependência linear podem ser visualizada nos gráficos;
- Frequentemente utiliza-se o coeficiente de correlação de Pearson[7] para indicar a correlação dois conjuntos
  - Mede a correlação linear entre dois conjuntos de dados;
  - Medida da covariância de duas variáveis, dividido pelo desvio padrão

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

[7] *Pearson Correlation Coefficient* (PCC) ou r-Pearson.

Correlação entre Features

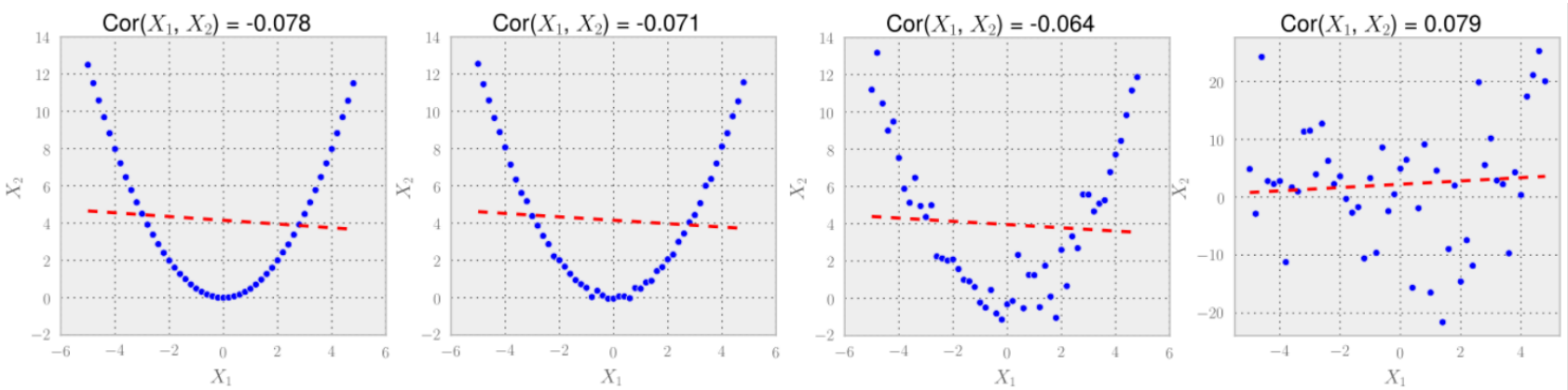
- Coeficientes altos de correlação indicam que as variáveis convergem em direção a uma informação idêntica ou similar;
- À medida em que a correlação diminui, as variáveis não convergem em direções similares [Richert and Coelho, 2013].



Fonte: Adaptado de [Richert and Coelho, 2013]

Correlação entre Features

- Apesar do bom desempenho para correlações lineares, as seleções de features baseadas em correlações são incapazes de identificar correlações não-lineares [Richert and Coelho, 2013];
- Para esses casos, a observação visual de correlações pode auxiliar na tomada de decisões.



Fonte: Adaptado de [Richert and Coelho, 2013]

Informação Mútua

- Informação mútua calcula a quantidade de informação que duas *features* possuem em comum;
- Ao contrário da correlação, é dependente da distribuição e não de uma sequência de dados;
- Pode utilizar, por exemplo, uma medida de entropia da informação, definida por: [Richert and Coelho, 2013]

$$H(X) = -\sum_{i=1}^{n} p(X_i) \log_2 p(X_i)$$

Informação Mútua

- A informação mútua é definida por:

$$I(X, Y) = -\sum_{i=1}^{m}\sum_{j=1}^{n} P(X_i, Y_j) \log_2 \frac{P(X_i, Y_j)}{P(X_i)P(Y_j)}$$

- O cálculo de P é feito pela categorização[8] de valores de recursos, seguidos pelo cálculo da fração dos valores em cada categoria [Richert and Coelho, 2013].

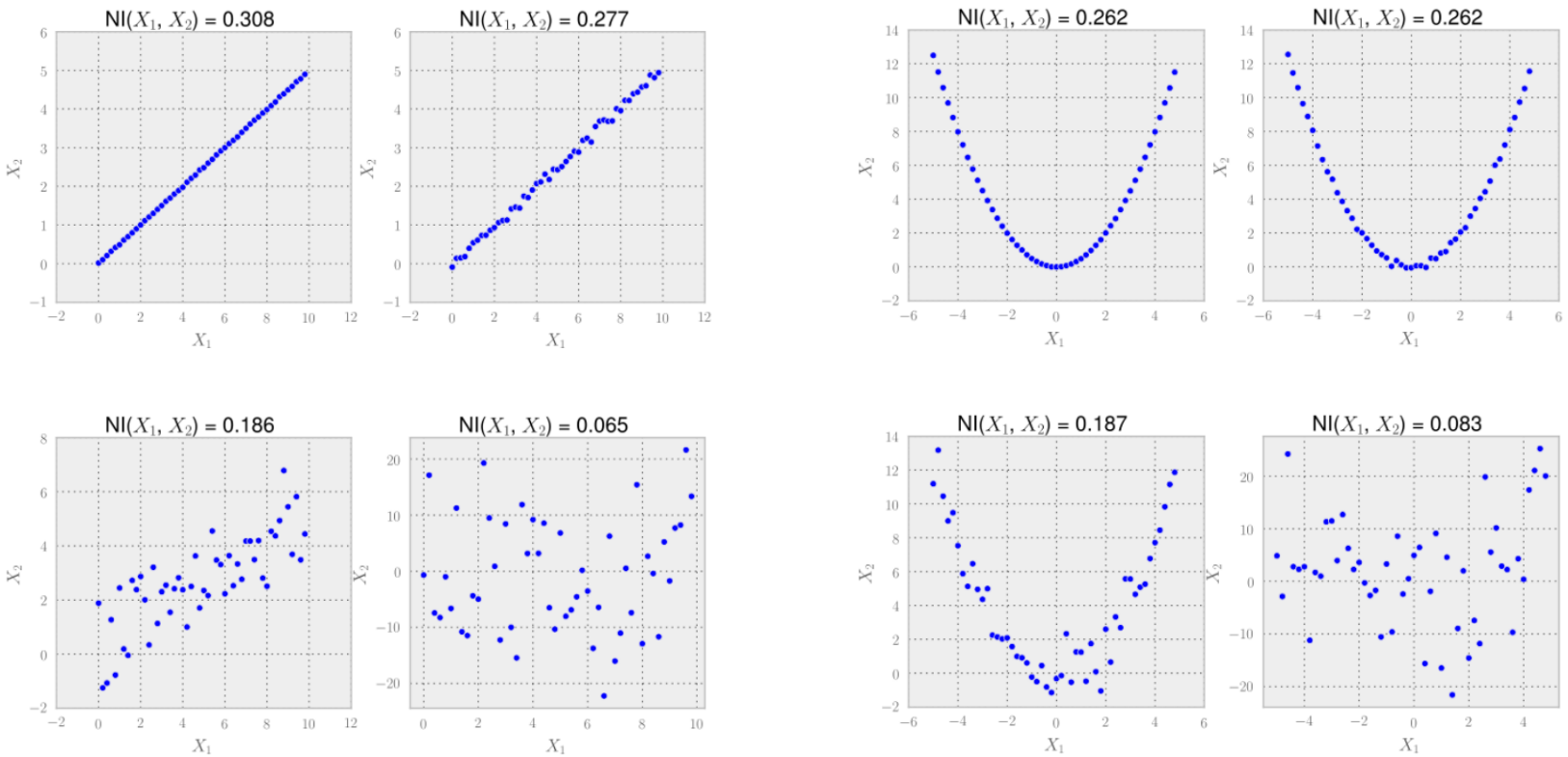
- No intervalo $[0, 1]$, a informação mútua é dada por:

$$NI(X, Y) = \frac{I(X, Y)}{H(X) + H(Y)}$$

[8] Tradução de *binning*, ou divisão em grupos/classes.

Informação Mútua

- A informação mútua pode ser definida para relações não lineares:



(a) (b)

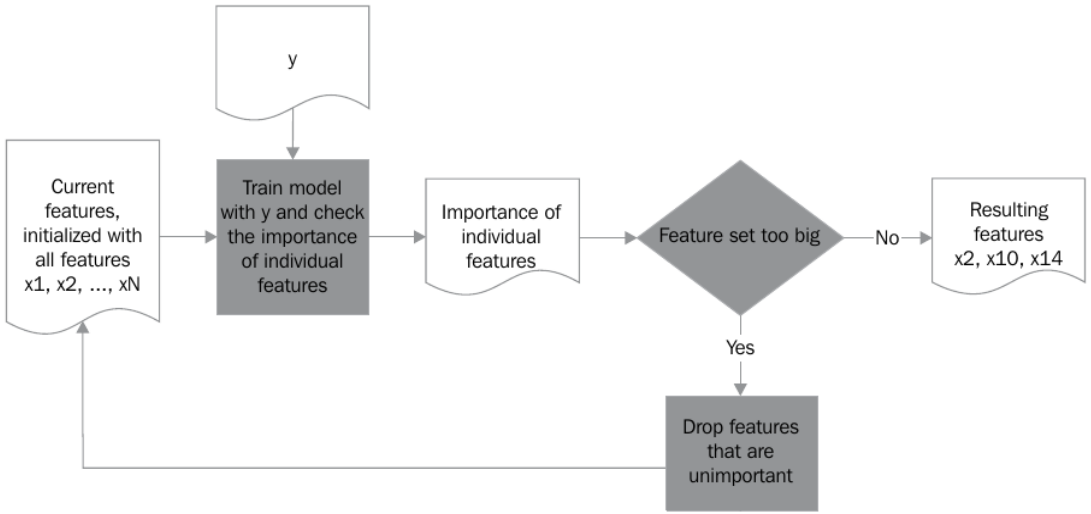
Fonte: [Richert and Coelho, 2013]

Informação Mútua

- Para múltiplas relações, é possível calcular a informação mútua normalizada para pares de variáveis;
  - Em pares de variáveis com valores altos, uma delas pode ser descartada;
  - Em alguns cenários, entretanto, múltiplas variáveis juntas tendem a retornar melhores resultados que variáveis separadas, não sendo recomendada a exclusão de uma delas
    - Essa condição pode ocorrer mesmo que as variáveis aparentem ser independentes;
  - Devido à complexidade, o uso da Informação Mútua em cenários de múltiplas variáveis pode ser proibitivo [Richert and Coelho, 2013].

Wrappers

- Apesar da existência de relações claras entre variáveis, algumas relações não possuem dependência evidente
  - Um exemplo simples é a operação XOR
- Uma técnica para seleção de *features* é o uso de *wrappers*
  - Modelos de aprendizado de máquina são usados para indicar quais *features* são importantes.



Fonte: [Richert and Coelho, 2013]

Recursive Feature Elimination (RFE)

- Recursive Feature Elimination (RFE)[9] é um algoritmo do tipo *wrapper* para seleção de *features*;
  - É um algoritmo fácil de configurar e usar;
  - Eficaz na seleção de *features* mais relevantes na previsão da saída[10];
  - Admite como parâmetro a quantidade de *features* desejadas, podendo treinar até que um subconjunto com essa cardinalidade seja obtido [Richert and Coelho, 2013].

[9] Tradução literal: Eliminação recursiva de *features* (características, recursos).

[10] À partir de um conjunto de treinamento.

Extração de Features

Extração de Features

- Após a seleção de *features*, o número de características restantes ainda pode ser alto;
- A etapa de extração de *features* é utilizada para redução extra
  - Nela objetiva-se a manutenção das *features* mais relevantes na previsão da saída [Richert and Coelho, 2013].
- Existem na literatura diversos algoritmos para essa tarefa
  - A maior parte deles são não-supervisionados [Marsland, 2014].

Análise de Discriminantes Lineares (LDA)

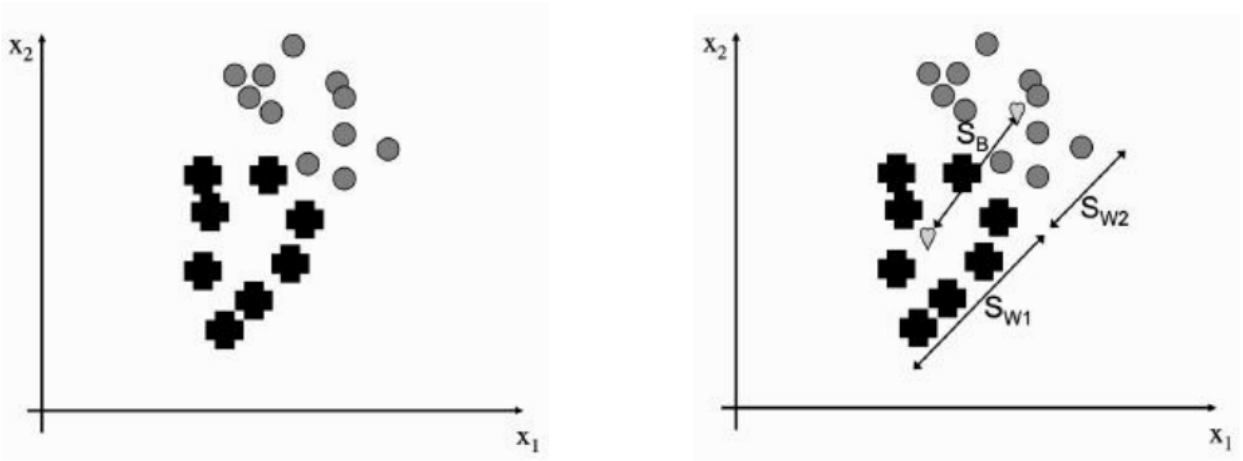
Análise de Discriminantes Lineares

- A Análise de Discriminantes Lineares (LDA)[11] é capaz de indicar, a partir da matriz de covariâncias, o quão dispersos estão os dados [Marsland, 2014]
  - A dispersão é dada pela multiplicação da covariância pela probabilidade de classe p_c.
- O LDA é recomendado para conjuntos rotulados de dados.

[11] Em inglês, *Linear Discriminant Analysis*.

Análise de Discriminantes Lineares

- Considere um conjunto de dados X com média μ
  - O conjunto pode ser dividido em duas classes x_1 e x_2, com respectivas médias μ_1 e μ_2;
  - A covariância de cada classe é dada por: $\sum_j (x_j - \mu)(x_j - \mu)^T$;
  - As dispersões internas S_{Wi} das classes e a dispersão entre classes S_B podem ser calculadas [Marsland, 2014].



(a) (b)

Fonte: [Marsland, 2014]

Na figura (b), a dispersão entre classes é calculada com base na média das classes. Ver slides seguintes.

Análise de Discriminantes Lineares

- Dispersão Interna da classe[12]: soma de todos os valores de todas as classes c [Marsland, 2014]

$$S_W = \sum_c \sum_{j \in c} p_c(x)(x_j - \mu_c)(x_j - \mu_c)^T$$

- Dispersão entre Classes[13]: define a distância entre classes, com objetivo que as classes estejam o mais longe possível

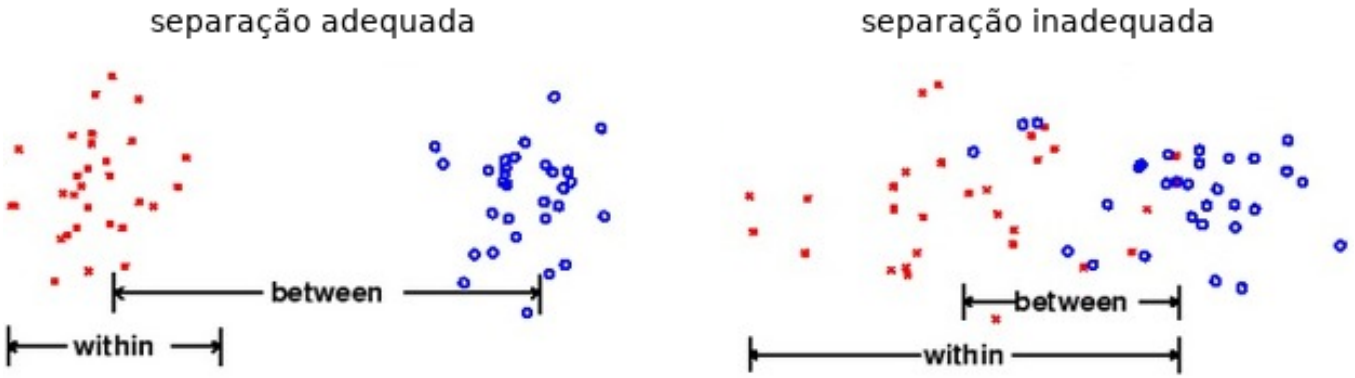
$$S_B = \sum_c (\mu_c - \mu)(\mu_c - \mu)^T$$

[12] Tradução direta de *within-class scatter*.

[13] Tradução direta de *between classes scatter*.

Análise de Discriminantes Lineares

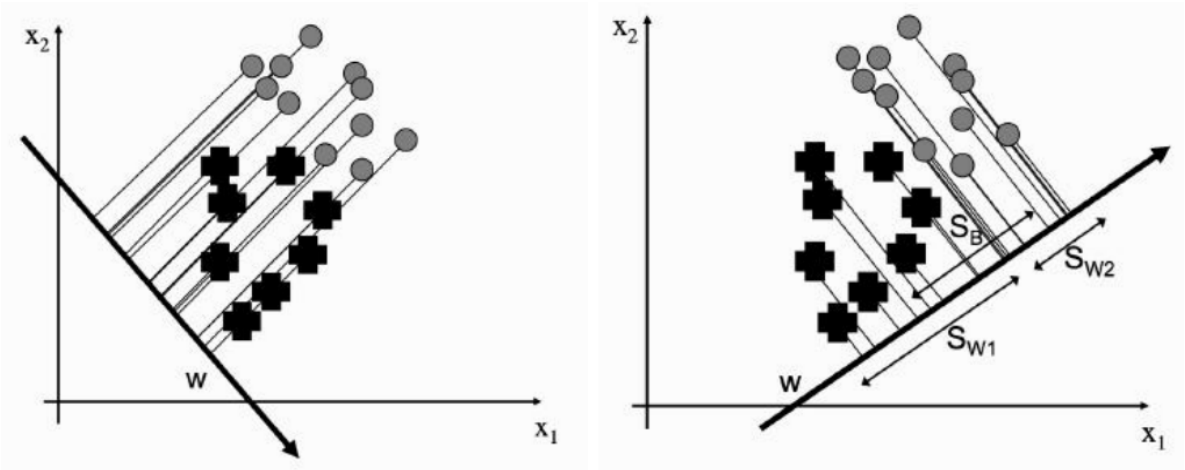
- Conjuntos que podem ser separados de forma adequada são chamados de discrimináveis [Marsland, 2014];
- A razão S_B/S_W deve ser o maior valor possível.



Fonte: Adaptado de [FunnyPR, 2014]

Análise de Discriminantes Lineares

- Podemos fazer uma projeção do conjunto de dados X, utilizado nesta seção
  - Dependendo da forma como a projeção é feita, ela pode, ou não, auxiliar na separação dos conjuntos;



Fonte: [Marsland, 2014]

Análise de Discriminantes Lineares

- A partir de arranjos matemáticos e com auxílio de um vetor de pesos w, podemos correlacionar as dispersões internas S_W e entre classes S_B, com a seguinte equação:

$$\boxed{S_w w = \frac{w^T S_w w}{w^T S_B w} S_B w}$$

- Para cálculo matemático, é necessário utilizar um autovalores e autovetores, que podem ser calculados com auxílio de um algoritmo [Marsland, 2014].

Análise de Discriminantes Lineares

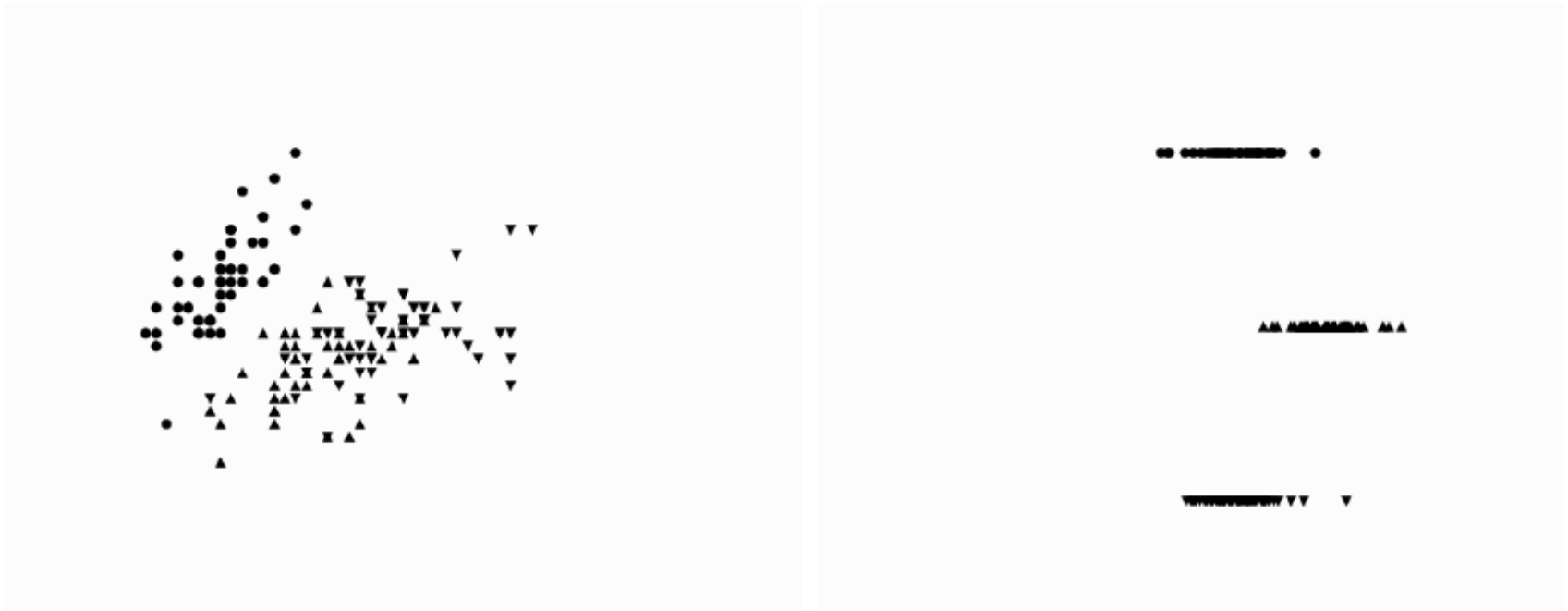
- O algoritmo a seguir pode ser usado para cálculo do LDA.

```
C = np.cov(np.transpose(data))

# Loop over classes
classes = np.unique(labels)
for i in range(len(classes)):
    # Find relevant datapoints
    indices = np.squeeze(np.where(labels==classes[i]))
    d = np.squeeze(data[indices,:])
    classcov = np.cov(np.transpose(d))
    Sw += np.float(np.shape(indices)[0])/nData * classcov

Sb = C - Sw
# Now solve for W and compute mapped data
# Compute eigenvalues, eigenvectors and sort into order
evals,evecs = la.eig(Sw,Sb)
indices = np.argsort(evals)
indices = indices[::-1]
evecs = evecs[:,indices]
evals = evals[indices]
w = evecs[:,:redDim]
newData = np.dot(data,w)
```

Fonte: [Marsland, 2014]

Análise de Discriminantes Lineares

- A figura a seguir contém o resultado de um conjunto de dados após aplicação do LDA.



Fonte: [Marsland, 2014]

Análise de Componentes Principais (PCA)

Análise de Componentes Principais

- Análise de Componentes Principais (PCA)[14] é possivelmente a primeira escolha como método de extração de *features* [Richert and Coelho, 2013];
  - Apesar de limitado a modelos lineares[15], tem alta capacidade de aprendizado;
- O PCA é adequado a conjuntos não rotulados
  - No entanto, nada impede que seja utilizado também em conjuntos rotulados [Marsland, 2014];
- O PCA é adequado tanto a tarefas de classificação quanto regressão [Richert and Coelho, 2013].

[14] Em inglês, *Principal Component Analysis*.

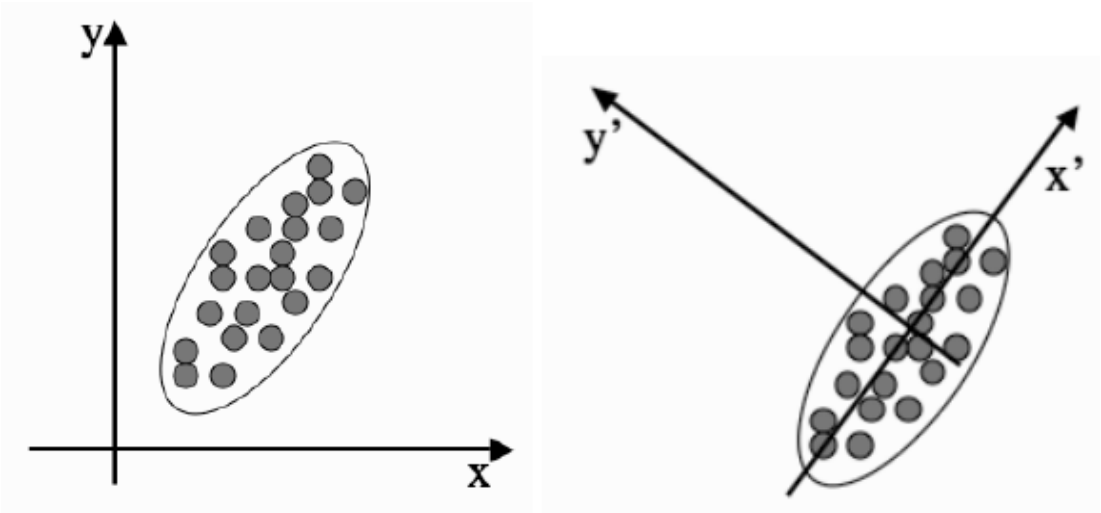
[15] Extensões do PCA permitem análise de modelos não lineares.

Análise de Componentes Principais

- A partir do espaço original de *features*, o PCA busca uma projeção linear que atenda às seguintes propriedades:
  - A variância conservada no processo dever ser maximizada;
  - Ao reconstruir as informações (voltar às *features* originais), o erro deve ser minimizado [Richert and Coelho, 2013].

Análise de Componentes Principais

- Considere o conjunto abaixo e a transformação (rotação + translação) realizada nas coordenadas:



Fonte: [Marsland, 2014]

- A partir da transformação realizada, é possível perceber de forma mais clara que o eixo y tem menor variabilidade.

Análise de Componentes Principais

- A ideia principal da Análise de Componentes é identificar a direção dos dados com maior variação;
- Etapas: [Marsland, 2014] [Richert and Coelho, 2013]
  1. Centralizar os dados, subtraindo a média;
  2. Escolher e criar um eixo na direção com a maior variação[16];
  3. Verificar a variação remanescente e encontrar um eixo ortogonal ao primeiro[17];
  4. Cobrir o máximo de itens possíveis com a variação restante;
  5. Iterar até esgotar todos os eixos possíveis.

[16] A escolha do eixo pode ser realizada a partir do cálculo da matriz de covariância.

[17] São realizados os cálculos do autovetores da matriz de covariância.

Análise de Componentes Principais

- O resultado do algoritmo indica que a variação ocorre ao longo dos eixos do conjunto de coordenadas
  - Em uma situação ideal, a matriz de covariância é diagonal;
  - Nesse cenário, cada nova variável está correlacionada somente a ela mesma [Marsland, 2014].

$$\text{cov}(\mathbf{Y}) = \text{cov}(\mathbf{P}^T\mathbf{X}) = \begin{pmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \lambda_N \end{pmatrix}$$

Fonte: [Marsland, 2014]

Análise de Componentes Principais

- A partir da matriz diagonal de covariância, é possível calcular autovalores e autovetores, gerando: [Marsland, 2014]

$$\boxed{V^{-1}CV = D}$$

onde

- C: matriz de covariância;
- V: autovetores da matriz C;
- V^{-1}: matriz inversa de autovetores de C;
- D: diagonal $M \times M$ da matriz de autovalores.

- As colunas de D são ordenadas de forma decrescente de autovalores, assim como as colunas de V;
- Autovalores inferiores a um limiar η são rejeitados, deixando L dimensões de dados [Marsland, 2014].

Análise de Componentes Principais

INSTITUTO FEDERAL
Minas Gerais

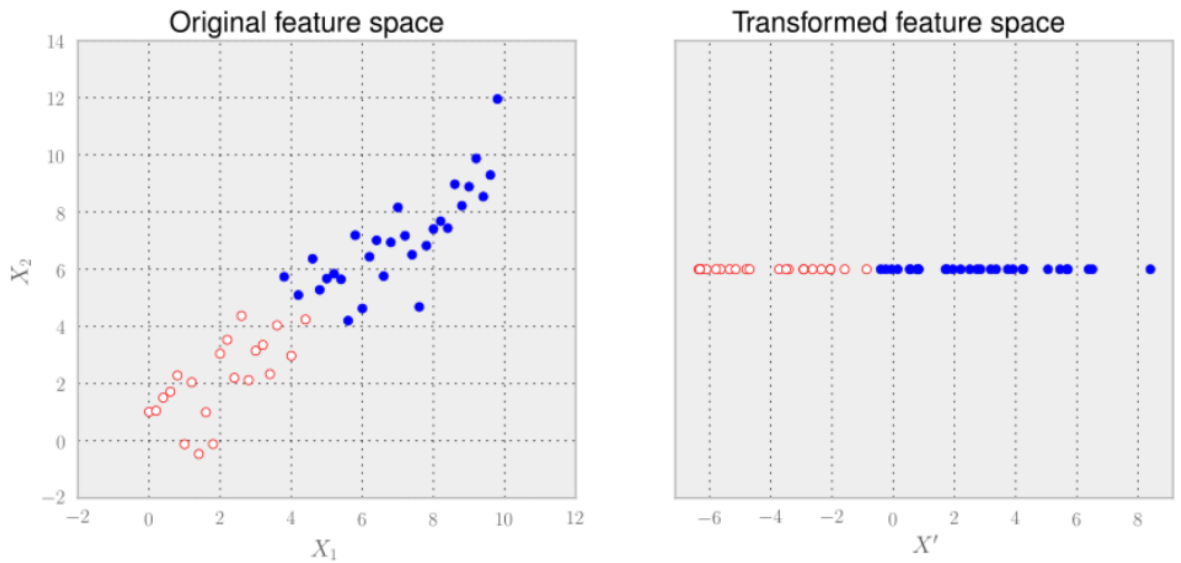
- O processo de cálculo do PCA pode ser resumido no seguinte algoritmo: [Marsland, 2014]

```
def pca(data,nRedDim=0,normalise=1):

    # Centre data
    m = np.mean(data,axis=0)
    data -= m

    # Covariance matrix
    C = np.cov(np.transpose(data))

    # Compute eigenvalues and sort into descending order
    evals,evecs = np.linalg.eig(C)
    indices = np.argsort(evals)
    indices = indices[::-1]
    evecs = evecs[:,indices]
    evals = evals[indices]

    if nRedDim>0:
        evecs = evecs[:,:nRedDim]

    if normalise:
        for i in range(np.shape(evecs)[1]):
            evecs[:,i] / np.linalg.norm(evecs[:,i]) * np.sqrt(evals[i])

    # Produce the new data matrix
    x = np.dot(np.transpose(evecs),np.transpose(data))
    # Compute the original data again
    y=np.transpose(np.dot(evecs,x))+m
    return x,y,evals,evecs
```

Fonte: [Marsland, 2014]

Análise de Componentes Principais

- A figura a seguir contém o resultado de um conjunto de dados após aplicação do PCA
  - Após a reconstrução do conjunto de dados, foi produzida, para o exemplo, a linha de dados, ao longo de um eixo X'.



Fonte: [Richert and Coelho, 2013]

Kernel PCA

- O PCA sempre assume que as direções de variação dos conjuntos são linhas retas;
  - Dependendo do conjunto, essa condição não ocorre;
- Uma extensão do PCA, chamada de Kernel PCA, permite transformações não lineares
  - Essa extensão possibilita que o PCA possa assumir direções diferentes de linhas retas [Richert and Coelho, 2013];
  - Utiliza-se uma função não linear $\Phi(\cdot)$ para cada *datapoint* x;
  - Os dados são transformados no espaço de *kernel* e o PCA é executado normalmente nesse espaço [Marsland, 2014].

Kernel PCA

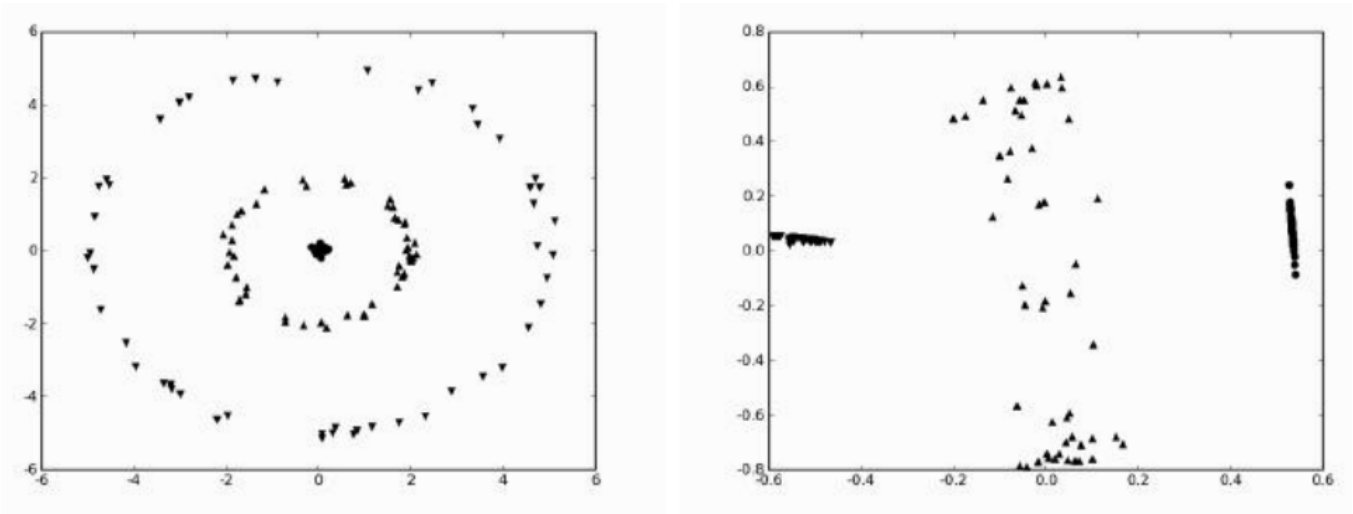
- O algoritmo do Kernel PCA pode ser resumido em: [Marsland, 2014]

```
K = kernelmatrix(data,kernel)

# Compute the transformed data
D = np.sum(K,axis=0)/nData
E = np.sum(D)/nData
J = np.ones((nData,1))*D
K = K - J - np.transpose(J) + E*np.ones((nData,nData))

# Perform the dimensionality reduction
evals,evecs = np.linalg.eig(K)
indices = np.argsort(evals)
indices = indices[::-1]
evecs = evecs[:,indices[:redDim]]
evals = evals[indices[:redDim]]

sqrtE = np.zeros((len(evals),len(evals)))
for i in range(len(evals)):
    sqrtE[i,i] = np.sqrt(evals[i])

newData = np.transpose(np.dot(sqrtE,np.transpose(evecs)))
```

Fonte: [Marsland, 2014]

Kernel PCA

- A figura a seguir contém o resultado de um conjunto de dados após aplicação do Kernel PCA (Gaussiano)
  - Os dados originais são três conjuntos, representados em círculos concêntricos [Marsland, 2014].



Fonte: [Marsland, 2014]

Outros Algoritmos

Análise Fatorial

- A Análise Fatorial[18] é uma técnica para criação de novas variáveis, a partir das variáveis originais;
- O método busca analisar se é possível explicar os dados por meio de um número menor de fatores não correlacionados ou variáveis latentes[19];
  - As variáveis são modeladas como uma combinalização de linear de fatores em potenciais, somadas a um error (ruído) inerente aos fatores;
- Essa técnica é utilizada comumente na psicologia e ciências sociais, onde os fatores possuem algum significado relevante [Marsland, 2014].

[18] Em inglês, *Factor Analysis*.

[19] Variáveis não diretamente observadas, que podem ser inferidas a partir de um modelo matemático.

Análise de Componentes Independentes (ICA)

- A Análise de Componentes Independentes (ICA) é um método que busca analisar componentes estatisticamente independentes
  - Pode ser considerado um tipo especial da método de Separação Cega de Sinais[20], correspondente à separação de um conjunto de sinais mistos, sem auxílio de informações sobre os sinais de origem ou processo de mistura.
  - Frequentemente associado ao "Problema da Festa"[21], no qual um conjunto de pessoas conversam simultaneamente em uma festa barulhenta e um ouvinte deve identificar claramente uma única conversa ("atenção seletiva") [Marsland, 2014].

[20] Tradução livre de *Blind Source Separation* (BSS).

[21] Tradução livre de *"The Cocktail Party Problem"*.

Multi-Dimensional Scaling (MDS)

- O método Multi-Dimensional Scaling (MDS) busca encontrar uma aproximação linear capaz de reduzir as dimensões e, ao mesmo tempo, preservar as distâncias entre pares de pontos
  - Em espaços Euclidianos, MDS e PCA (que tenta preservar as variâncias) possuem comportamentos idênticos.
- O MDS não se preocupa com os pontos de dados em si, focando nas dissimilaridades[22] entre eles
  - As dissimilaridades são interpretandas como distâncias;
  - O MDS funciona bem apenas em *manifolds*[23] planos [Richert and Coelho, 2013] [Marsland, 2014].

[22] Quanto maior o valor observado menos parecidos (dissimilares) são os objetos [Müller, 2015].

[23] Variação de um espaço topológico parecido a um espaço euclidiano nas vizinhanças de cada ponto.

Isomap

- O Isomap é um algoritmo proposto por Tenenbaum et al., em 2000, como uma variação do MDS;
- Assume que a distância ente pares de pontos são sempre adequadas [Marsland, 2014];
- Processo:
  - Supõe que em uma pequena distância a não linearidade do *manifold* não importa;
  - Constrói distâncias entre pontos distantes, encontrando caminhos que percorrem pontos próximos (vizinhos);
  - Utiliza, em seguida, o MDS convencional na matriz de distâncias [Marsland, 2014].

Referências I

FunnyPR (2014).
Linear discriminant analysis(lda) - 2 classes.
[Online]; acessado em 09 de Fevereiro de 2021. Disponível em:
https://funnypr.tistory.com/entry/Linear-Discriminant-AnalysisLDA.

Kopec, D. (2019).
Classic Computer Science Problems in Python.
Manning Publications Co, 1 edition.

Marsland, S. (2014).
Machine Learning: An Algorithm Perspective.
CRC Press, 2 edition.
Disponível em: https://homepages.ecs.vuw.ac.nz/ marslast/MLbook.html.

Müller, S. (2015).
AnÃ¡lise de agrupamentos (clusters) - notas de aulas.
[Online]; acessado em 10 de Fevereiro de 2021. Disponível em:
https://docs.ufpr.br/~soniaisoldi/ce090/8ANALISEAGRUPAMENTOS.pdf.

Richert, W. and Coelho, L. P. (2013).
Building Machine Learning Systems with Python.
Packt Publishing Ltd., 1 edition.

Rosen, K. H. (2019).
Discrete Mathematics and Its Applications.
McGraw-Hill, 8 edition.

Referências II

Shalev-Shwartz, S. and Ben-David, S. (2014).
Understanding Machine Learning: From Theory to Algorithms.
Cambridge University Press, 1 edition.
Disponível em: http://www.cs.huji.ac.il/ shais/UnderstandingMachineLearning.

Wikipedia contributors (2021).
Normal distribution.
[Online]; acessado em 27 de Janeiro de 2021. Disponível em:
https://en.wikipedia.org/wiki/Normal_distribution.