Aprendizado de Máquina - Relatório

Prática 4 - PCA

(Análise de Componentes Principais)

1 - Conceitos Iniciais

Durante o desenvolvimento dessa prática alguns conceitos foram necessários, conceitos estes que foram de grande ajuda para o compreendimento e desenvolvimento da mesma. Ao longo desta seção iremos seção iremos fornecer uma breve definição dos conceitos utilizados.

Covariância

Quando duas variáveis aleatórias X e Y não são independentes, geralmente é de interesse avaliar quão fortemente estão relacionadas uma com a outra. A covariância dá uma ideia da dispersão dos valores da variável bidimensional (X,Y) em relação ao ponto (E(X), E(Y)), onde E(Z) é a operação de tomar a esperança do atributo Z, ou seja, é o vetor dos valores esperados de cada elemento de Z.

Definição:

Seja (X,Y)uma variável aleatória bidimensional. A covariância de X e Y que geralmente é denotada por Cov(X,Y) é definida por:

$$Cov(X,Y) = E[(X - E(X))(Y - E(Y))]$$

Um ponto a se observar é que se X e Y são variáveis aleatórias independentes Cov(X,Y) = 0, ou seja, as variáveis X e Y não possuem relação uma com a outra.

Contudo, a covariância será positiva se as duas variáveis tendem a variar no mesmo sentido, isto é, valores de X acima da sua média estão associados a valores de Y acima de sua média, o mesmo ocorrendo para valores de ambos inferiores à média. Deste modo, a covariância será negativa se valores acima da média de uma variável estão associados a valores inferiores à média da outra.

Matriz de Covariância

A matriz de covariância é uma matriz quadrática, simétrica de dimensões NxN, onde N é a quantidade de variáveis. Nela é armazenada a covariância entre todas as variáveis, assim nesta matriz uma posição A_{ij} é igual a covariância da variável i em relação a variável j. A diagonal principal da matriz contém as variâncias e as demais posições a correlação entre as direções(demais variáveis).

Nesta prática em específico ela é utilizada para o desenvolvimento do PCA(Análise de Componentes Principais), mas o uso dela não se restringe somente a isto.

PCA (Análise de Componentes Principais)

A Análise de Componentes Principais - PCA (do inglês Principal Component Analysis) é um método que tem por finalidade básica, a análise dos dados usados visando sua redução, eliminação de sobreposições e a escolha das formas mais representativas de dados a partir de combinações lineares das variáveis originais.

O PCA é um dos métodos estatísticos de múltiplas variáveis mais simples. O PCA é considerado uma transformação linear ótima, dentre as transformadas de imagens, sendo muito utilizada pela comunidade de reconhecimento de padrões.

Ele transforma variáveis discretas em coeficientes descorrelacionados. É uma maneira de identificar a relação entre as características extraídas de dados, é bastante útil quando os vetores de características têm muitas dimensões, quando uma representação gráfica não é possível, porém também pode ser útil em dimensões menores.

O componente principal é o arranjo que melhor representa a distribuição dos dados e a componente secundária é perpendicular a ela.

Assim, o PCA consiste em promover uma transformação linear nos dados de modo que os dados resultantes desta transformação tenham suas componentes mais relevantes nas primeiras dimensões, em eixos denominados principais. A matriz de transformação utilizada para o cálculo da PCA consiste em uma matriz cujas linhas são os autovetores da matriz de covariância estimada dos dados.

O cálculo desta matriz de transformação é baseado na matriz de covariância e esses autovetores são inseridos nessa matriz de acordo com seus respectivos autovalores, do maior para o menor. Os autovetores desta matriz de fato formam uma nova base que segue a variação dos dados. O PCA, portanto consiste em uma mudança de base.

2 - Desenvolvimento do PCA

A criação do PCA foi feita de acordo com a sequência de passos descrita no enunciado, a qual está descrita abaixo.

Inicialmente foi separado os atributos que pertencem a classe especificada dos demais no conjunto de dados.

Passo 1:

De acordo com esse novo conjunto de dados foi calculado a média para cada atributo.

Passo 2:

Para cada atributo do novo conjunto de dados e utilizando as médias obtidas no passo 1, foi efetuado uma alteração no valor de cada atributo. Para cada atributo foi subtraído de seu valor a sua média. Ou seja, cada atributo do conjunto de dados possui agora o valor resultante da subtração de sua média em seu valor anterior.

Passo 3:

Após realizar o passo 2 foi calculada a matriz de covariância desse conjunto de dados com o intuito de saber a relação das variáveis entre si, e assim poder utilizar-se dela para um cálculo posterior dos autovalores e autovetores.

Passo 4:

Baseado na matriz calculada no passo anterior foi calculado os seus respectivos autovalores e autovetores, que nos dão a informação dos atributos mais representativos daquele conjunto e quais direções devemos seguir.

Passo 5:

Como o passo anterior retorna duas listas, uma para os autovalores e outra para os autovetores de modo que um elemento da primeira está diretamente relacionado com um elemento da segunda na mesma posição.

Assim, a lista de autovalores foi ordenada de modo não crescente e simultaneamente a lista de autovetores também foi ordenada.

Passo 6:

Processo de criação da matriz de transformação, a qual é composta pelos autovetores(como vetor-coluna) do maior para o menor de acordo com os autovetores e autovalores dos passos 4 e 5. Deste modo, efetuou-se uma descorrelação dos dados de modo que pode-se agora realizar uma diminuição de dimensão sem perda de informação, visto que como dito anteriormente isso é possível via combinação linear.

Para efetuar essa redução de dimensão é necessário o cálculo de uma constante k (corresponde ao número de componentes principais e geralmente é determinado com base na variância explicada pelas k primeiras componentes). É recomendável escolher um valor de k, tal que a variância explicada pelos k maiores autovalores corresponda a um certo valor pré-definido entre 90% e 95% da variância total.

Passo 7:

Uma nova matriz Y é criada de acordo com uma transformação linear, $Y = X*Q$. Onde X é a matriz de dados originais, e Q é a matriz resultante do passo 6. Assim, dependendo como foi considerado a matriz que essa transformação pode ser alterada. Se os vetores de atributos estivessem organizados na forma de vetor-coluna, a operação linear acima seria dada por $Y = Q' X$. Onde Q' significa é a matriz transposta de Q.

Passo 8:

Foi calculado a matriz de covariância da matriz Y resultante do passo anterior, e assim foi obtida a correlação das variáveis, porém agora com uma dimensão a menos do que a matriz anterior.

Portanto, como foi descrito anteriormente e o que foi percebido durante a prática, é que o PCA é um método de análise de componentes com o intuito de descobrir quais deles representa melhor o conjunto de dados, e assim realizar uma redução. Por isso realiza a descorrelação dos dados, pois retira do conjunto de atributos aqueles que podem ser obtido através dos demais via combinação linear, e assim permitindo reduzir a quantidade de componentes significativos.

3 - Código

```
#-*- coding: utf-8 -*-
import numpy as np
from numpy import linalg as LA

def ordenar(lista, matriz):
  valores = []
  vetores = []
  while len(lista) != 0:
      valor = max(lista)
      pos = lista.index(valor)
      vetor = matriz[pos]
      valores.append(valor)
      vetores.append(vetor)
      lista.remove(valor)
      matriz.remove(vetor)
  return  valores,vetores

# Lendo o arquivo
data = np.genfromtxt("coluna_com_label_bipolar.txt", delimiter=",")

# Total de elementos da classe Hernia
qtd_hernia = 0
for linha in data:
  if int(linha[6]) == 1:
      qtd_hernia += 1

# Classe Hernia, seprando dados
hernia = np.zeros((qtd_hernia, 6))
i = 0
while i < len(data):
  if int(data[i][6]) == 1:
      hernia[i] = data[i][0:6].tolist()
  i += 1

# Passo 1: Cálculo da média da classe.
media = np.mean(hernia, axis=0)

# Passo 2: Subtração de cada linha do vetor da matriz pela média de seu atributo, gerando uma nova matriz
matriz = np.zeros((qtd_hernia, 6))
i = 0
while i < qtd_hernia:
  matriz[i] = (hernia[i] - media)
  i += 1

# Passo 3: Matriz de covariância..
matriz_cov = np.cov(matriz, None, 0)

# Passo 4: Autovalores e Autovetores.
a_valores, a_vetores = LA.eig(matriz_cov)
```

```
# Passo 5: Ordenar de modo não crescente
valores, vetores = ordenar(list(a_valores.tolist()), list(a_vetores.tolist()))

a_valores = valores
a_vetores = vetores

# Passo 6: Construindo a matriz Q
# 6 - A: Matriz de alto vetores do maior para o menor
# Como os alto vetores já estão ordenados de acordo com
n = len(a_vetores)
Q1 = np.zeros((n, n))
i = 0
while i < len(a_vetores):
  Q1[i] = a_vetores[i]
  i += 1


# Passo 6-B
#Encontra o valor de q e criar a nova matriz

soma = 0
var = 0
total = int(0.9*len(data))
i = 0
selecionados = []
while not var >= total:
  soma += a_valores[i]
  var += soma/(len(a_vetores))
  selecionados.append(i)
  i += 1

# Pegando os autovetores selecionados
Q = Q1[0:len(selecionados)]

# Passo 7:
# Como tata-se de vetores colunas
Y = np.dot(hernia, Q.transpose())

# Passo 8:
# Calculando a nova matriz de covariância.
n_matriz_cov = np.cov(Y, None, 0)

print "Matriz de Covariância Original: \n", matriz_cov , "\n"
print "Nova Matriz de Covariância: \n", n_matriz_cov , "\n"
```