Universidade Federal do Ceará - Campus de Quixadá

Prática 04
BCC0019 - Aprendizado de Máquina

Prof. Carlos Igor Ramos Bandeira

Outubro de 2016

Aluno (Nome/Matrícula): ____

A PONTUAÇÂO TOTAL QUE PODE SER CONQUISTADA COM ESTE TRABALHO É DE 10,0.

PRAZO DE ENTREGA: DEZEMBRO/2016.

1. (100 points) Aplicar PCA aos vetores de atributos da classe Hérnia de Disco (arquivo "coluna-com-label-binario.dat") para gerar um novo conjunto de vetores cujos os atributos são não-correlacionados. Os passos para executar esta tarefa são os seguintes, assumindo que os vetores de atributos estão organizados na forma de vetor-linha:

   Passo 1: Seja $\mathbf{X}$ uma matriz $n \times p$ cujas linhas correspondem aos n vetores de atributos e as colunas aos p atributos. Estimar o vetor médio $\bar{\mathbf{x}}$ do conjunto $\mathbf{X}$ original. Pode-se usar o comando mean do Matlab/Octave.

   Passo 2: Subtrair o vetor de cada linha da matriz $\mathbf{X}$, gerando a matriz $\mathbf{X}^c$. Em outras palavras, realizar a seguinte operação para cada linha da matriz $\mathbf{X}$: $\mathbf{x}_i^c = \mathbf{x}_i - \bar{\mathbf{x}}$, em que $\mathbf{x}_i$ corresponde à i-ésima linha de $\mathbf{X}$.

   Passo 3: Estimar a matriz de covariância $\mathbf{C}_x$ usando a matrix $\mathbf{X}_c$. Pode-se usar o comando *cov* do Matlab/Octave.

   Passo 4: Determinar os p autovalores (λ_i) e os p autovetores ($mathbfv_i$) associados à matriz de covariância $\mathbf{C}_x$. Pode-se usar o comando *eig* do Matlab/Octave.

   Passo 5: Ordenar os autovalores por ordem decrescente de magnitude, mantendo correspondência com o autovetor correspondente.

$$\lambda_1(\mathbf{v}_1) < \lambda_2(\mathbf{v}_2) < \cdots < \lambda_p(\mathbf{v}_p) \quad (1)$$

Passo 6a (Descorrelação): Assumindo que os autovetores são vetores-coluna, montar a matriz $\mathbf{Q} = [\mathbf{v}_1|\mathbf{v}_2|\ldots|\mathbf{v}_p]_{p\times p}$, cujas colunas (montadas da esquerda para a direita) são formadas pelos autovetores associados aos autovalores ordenados do maior para o menor.

Passo 6b (Descorrelação e Redução de Dimensão): Montar a matriz $\mathbf{Q} = [\mathbf{v}_1|\mathbf{v}_2|\ldots|\mathbf{v}_q]_{p\times q}$, cujas q ($q < p$) colunas (montadas da esquerda para a direita) são formadas pelos autovetores associados aos q maiores autovalores. O parâmetro q é a dimensão dos vetores transformados.

Observação 2: O parâmetro q corresponde ao número de componentes principais e geralmente é determinado com base na variância explicada pelas q primeiras componetes, VE(q):

$$\mathrm{VE}(q) : \frac{\sum_{k=1}^{q} \lambda_k}{\sum_{k=1}^{p} \lambda_k}. \tag{2}$$

Note que a variância explicada pelas q primeiras componentes nada mais é do que a razão entre a soma das variâncias devido aos q maiores autovalores pela variância total.

Assim, deve-se escolher um valor de q tal que a variância explicada pelos q maiores autovalores corresponda a um certo valor pré-definido (e.g. 0,90 ou 0,95) da variância total. A pergunta de interesse, por exemplo, pode ser colocada nos seguintes termos: Qual o valor de q que explica (ou conserva) 90% da variância total dos dados originais?

Passo 7: Gerar uma nova matriz de dados $\mathbf{Y}$ por meio da seguinte transformação linear: $\mathbf{Y} = \mathbf{XQ}$.

Observação 3: A operação acima pode ser aplicada individualmente a cada linha da matriz $\mathbf{X}$, gerando cada linha da matriz $\mathbf{Y}$ individualmente: $\mathbf{y}_i = \mathbf{x}_i\mathbf{Q}, i = 1, \ldots, n$.

Observação 4: Note que se os vetores de atributos estivessem organizados na forma de vetor-coluna, a operação linear acima seria dada por $\mathbf{Y} = \mathbf{Q}^T\mathbf{X}$. Aplicada individualmente a cada coluna de $\mathbf{X}$, teríamos $\mathbf{y}_i = \mathbf{Q}^T\mathbf{x}_i, i = 1, \ldots, n$.

Passo 8: Estimar a matriz de covariância $\mathbf{C}_y$ usando a matrix $\mathbf{Y}$. Pode-se usar o comando *cov* do Matlab/Octave.