Universidade Federal do Paraná
Departamento de Informática

Reconhecimento de Padrões

PCA

Luiz Eduardo S. Oliveira, Ph.D.
http://lesoliveira.net

Objetivos

- Introduzir os conceitos de PCA e suas aplicações para extração de características
  - Revisão dos conceitos básicos de estatística e álgebra linear.

Estatística

- Variância
  - Variância de uma variável aleatória é uma medida dispersão estatística, indicando qual longe em geral os seus valores se encontram do valor esperado.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2,$$

  - Desvio padrão é a raiz da variância
    - O resultado do desvio se dá na mesma medida dos dados da população ou amostra.

Estatística

- Covariância
  - Variância é uma medida unidimensional.
  - É calculada de maneira independente pois não leva em consideração as outras dimensões.
  - Covariância por sua vez, é uma medida bi-dimensional. Verifica a dispersão, mas levando em consideração duas variáveis aletórias.

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)}$$

Estatística

- Matriz de covariância
  - Para 3 variáveis aleatórias, x, y e z, o cálculo de todas as covariâncias (x-y, x-z e y-z) pode ser acomodada em uma matriz, a qual denomina-se matriz de covariância.

$$C = \begin{pmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{pmatrix}$$

Cov(x,y) = cov(y,x)

Cov(z,z) = var(z)

Álgebra

- Autovetores
  - Como sabe-se duas matrizes podem ser multiplicadas se elas possuem tamanhos compatíveis. Autovetores são casos especiais neste contexto.

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 1 \\ 3 \end{pmatrix} = \begin{pmatrix} 11 \\ 5 \end{pmatrix}$$

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 12 \\ 8 \end{pmatrix} = 4 \times \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$ → Múltiplo do vetor resultante

Autovetores

- Nesse caso (3,2) representa um vetor que aponta da origem (0,0) para o ponto (3,2).
- A matriz quadrada, pode ser vista como uma matriz de transformação.
- Se esta matriz for multiplicada por outro vetor, a resposta será outro vetor transformado da sua posição original.
- É da natureza desta transformação que surgem os autovetores.

Autovetores

- Propriedades
  - Podem ser achados somente em matrizes quadradas.
  - Nem todas as matrizes possuem autovetores.
  - Para uma data *n* x *n* matriz, existem *n* autovetores.
  - Se o vetor for multiplicado por uma constante, ainda obteremos o mesmo resultado

$$2 \times \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 6 \\ 4 \end{pmatrix}$$

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 6 \\ 4 \end{pmatrix} = \begin{pmatrix} 24 \\ 16 \end{pmatrix} = 4 \times \begin{pmatrix} 6 \\ 4 \end{pmatrix}$$

Apenas fazemos o vetor mais longo, mas não mudamos a direção.

Autovetores/Autovalores

- Todos os autovetores são ortogonais (perpendiculares), ou seja os dados podem ser expressos em termos destes vetores.
- O valor pelo qual o vetor é multiplicado é conhecido como autovalor
  - Um autovetor sempre possui um autovalor associado.

Análise dos Componente Principais (PCA)

- Uma maneira de identificar padrões em dados, colocando em evidência suas similaridades e diferenças.
- Ferramenta importante para altas dimensões, onde não podemos fazer uma análise visual.
- Uma vez encontrados esses padrões, podemos comprimir os dados sem grande perda de qualidade.
- Extrator de características (representação)

PCA Tutorial

- 1) Escolha um conjunto de dados.
- 2) Normalize esses dados, subtraindo-os da média.

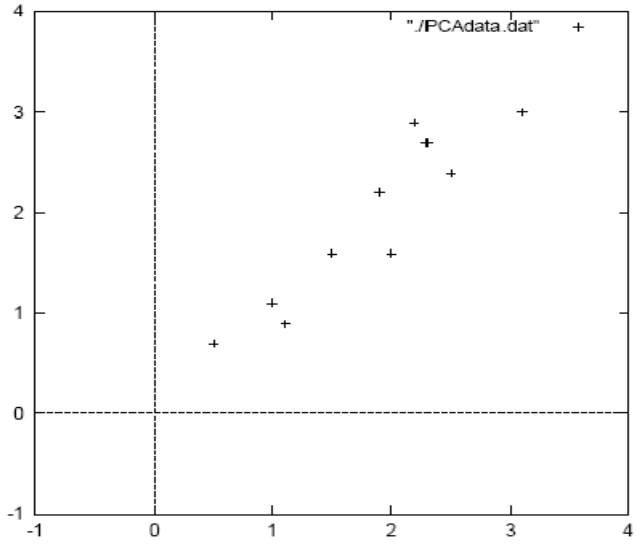
Dados

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

Dados Normalizados

x	y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01

PCA Tutorial



PCA Tutorial

- 3) Calcule a matriz de correlação para os dados normalizados. Uma vez que os dados possuem duas dimensões, teremos uma matriz 2x2

$$cov = \begin{pmatrix} .616555556 & .615444444 \\ .615444444 & .716555556 \end{pmatrix}$$

PCA Tutorial

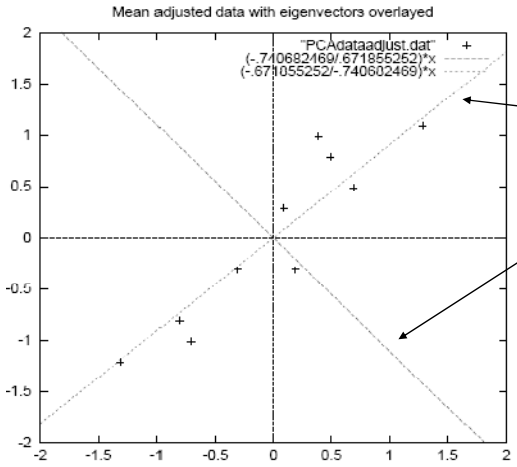
- 4) Encontre os autovetores e autovalores para a matriz de covariância.
  - Uma vez que a matriz de covariância é quadrada podemos encontrar os autovetores e autovalores.

$$eigenvalues = \begin{pmatrix} .0490833989 \\ 1.28402771 \end{pmatrix}$$

$$eigenveciors = \begin{pmatrix} -.735178656 & -.677873399 \\ .677873399 & -.735178656 \end{pmatrix}$$

O que esses valores significam ??

PCA Tutorial



$$\begin{pmatrix} -.735178656 & -.677873399 \\ .677873399 & -.735178656 \end{pmatrix}$$

PCA Tutorial

- 5) Escolhendo os componentes que vão formar o vetor
  - Como vimos, os autovalores são bastante diferentes.
  - Isso permite ordenar os autovetores por ordem de importância.
  - Se quisermos eliminar um componentes, devemos então eliminar os que tem menos importância.

$$FeatureVector = (eig_1\ eig_2\ eig_3\ ....\ eig_n)$$

PCA Tutorial

- No nosso exemplo temos duas escolhas
  - Manter os dois.
  - Eliminar um autovetor, diminuindo assim a dimensionalidade dos dados
    - Maldição da dimensionalidade
    - Quanto maior a dimensionalidade do seu vetor, mais dados serão necessários para a aprendizagem do modelo.

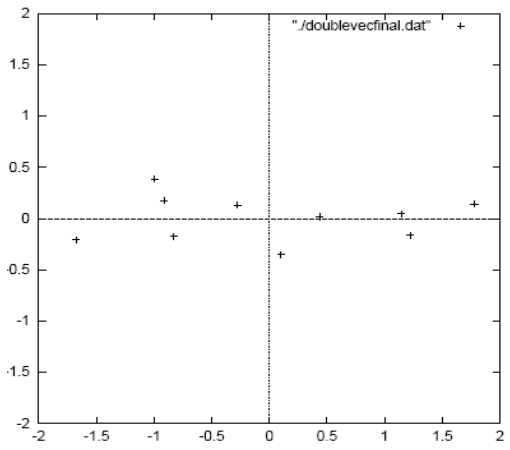
PCA Tutorial

- 5) Construindo novos dados.
  - Uma vez escolhidos os componentes (autovetores), nós simplesmente multiplicamos os dados pelo autovetor(es) escolhidos.
  - O que temos?
    - Dados transformados de maneira que expressam os padrões entre eles.
    - Os PCs (Principal Components) são combinações linear de todas as características, produzindo assim novas características não correlacionadas.

PCA Tutorial

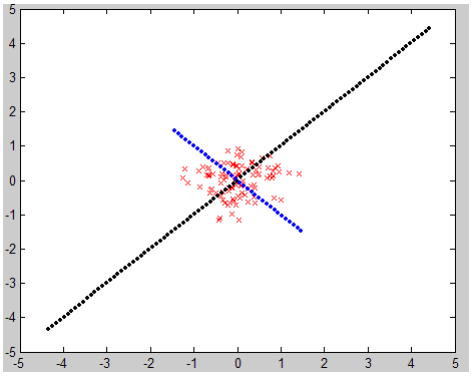
Dados transformados usando 2 autovetores

x	y
-.827970186	-.175115307
1.77758033	.142857227
-.992197494	.384374989
-.274210416	.130417207
-1.67580142	-.209498461
-.912949103	.175282444
.0991094375	-.349824698
1.14457216	.0464172582
.438046137	.0177646297
1.22382056	-.162675287



PCA Tutorial

- Exemplo

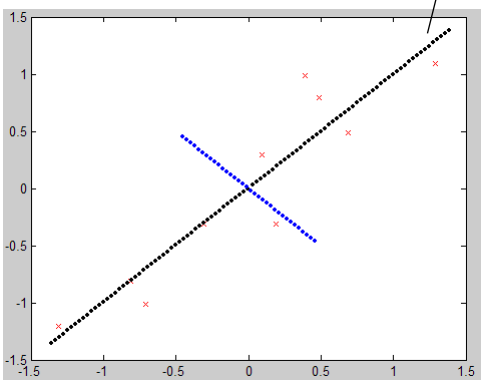


Usando a função pcacov do Matlab, três parâmetros são retornados.
-autovetores
-autovalores
-percentagem da variância total explicada pro cada modelo

PCA Tutorial

- Exemplo 2

$$y = \frac{-W_1 \times x - b}{W_2}$$



AUTOVETORES
-0.6779 -0.7352
-0.7352 0.6779

AUTOVALORES
1.2840
0.0491

VARIÂNCIA EXPLICADA
96.3181
3.6819

PCA Tutorial

- Exercício
  - Gere diferentes conjuntos de dados em duas dimensões e aplique PCA. Verifique e discuta a variância explicada em função da dispersão dos dados.