Análisis de componentes principales

Lección 21

Dr. Pablo Alvarado Moya

CE5506 Introducción al reconocimiento de patrones
Área de Ingeniería en Computadores
Tecnológico de Costa Rica

II Semestre, 2019

Contenido

Introducción

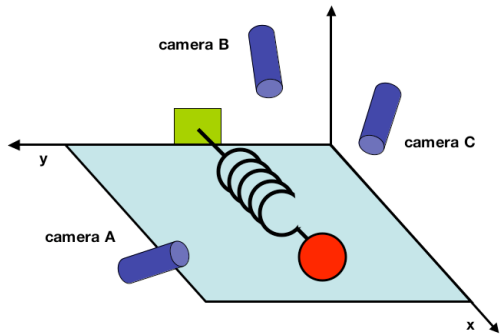
- El Análisis de Componentes Principales (ACP, o PCA) es una herramienta estándar del análisis de datos.
- Sus aplicaciones se encuentran desde neurociencias hasta gráficos por computador.
- Su tarea principal es **reducir** el número de dimensiones de vectores de entrada, con el afán de eliminar redundancia.
- Referencia: Shlens, J. A Tutorial on PCA.

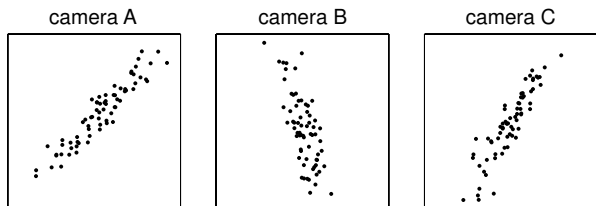
Experimento de partida (1)

- Supóngase que realizamos un experimento en donde se miden varias cantidades para analizar un fenómeno (p. ej. espectros, tensiones eléctricas, velocidades, etc.)
- Con esos datos queremos realizar clasificación, aglomeración, o encontrar la distribución que les subyace.
- Los datos medidos pueden ser difusos, redundantes, o poco claros.
- Datos así obstaculizan los proceso de aprendizaje, principalmente por la maldición de la dimensión.

Experimento de partida (2)

- Partamos del siguiente montaje para estudiar el movimiento de un resorte ideal:





Una bola de masa m está sujeta a un resorte sin masa y sin fricción. La bola se libera a una pequeña distancia alejada del equilibrio. La bola oscila entonces en el eje x a una determinada frecuencia.

Experimento de partida (3)

- Supóngase que se ignora cuáles ejes son importantes para la medición, así que se toman medidas en el espacio tridimensional utilizando tres cámaras de alta velocidad que toman 120 cuadros por segundo.
- Las cámaras se colocan en tres posiciones y orientaciones arbitrarias.
- La pregunta es cómo obtener de todas las mediciones la ecuación que ponen en evidencia la única dependencia de x.
- En el mundo real, las mediciones tomadas están además contaminadas con ruido, y por efectos de condiciones no ideales.

Cambio de base (1)

- La meta del ACP es identificar la base con "mayor sentido" para reexpresar los datos.
- Se espera además que en esta nueva base sea fácil filtrar los efectos del ruido y encontrar estructuras ocultas en los datos (por ejemplo, descubrir de las mediciones cuál $\underline{\mathbf{x}}$ es la dimensión de interés.
- Como únicamente utiliza los vectores $\underline{\mathbf{x}}$, es un método de reducción de dimensiones no supervisado

Una base canónica (1)

- Se tratará a cada dato tomado en el experimento como un elemento del conjunto de datos.
- A este dato se le denomina **muestra** $\underline{\mathbf{x}}^{(i)}$.
- El dato es un vector conformado por todas las mediciones tomadas (tensión eléctrica, posición, etc.)
- En el ejemplo, la cámara A reporta una posición $(x_A^{(i)}, y_A^{(i)})$, así que una muestra en el ejemplo está dada por:

$$\underline{\mathbf{x}}^{(i)} = \begin{bmatrix} x_A^{(i)} & y_A^{(i)} & x_B^{(i)} & y_B^{(i)} & x_C^{(i)} & y_C^{(i)} \end{bmatrix}^T$$

- Si se grabaran estas posiciones durante 10 minutos a 120 fps se tendrían entonces $i = 1 \ldots 72000$ muestras.
- Cada muestra $\underline{\mathbf{x}}^{(i)}$ es un vector del espacio $\mathbb{R}^n$

Cambio de base (1)

- La pregunta es ahora ¿Existe alguna combinación **lineal** de la base original que exprese "mejor" el conjunto de datos?
- Sea $\mathbf{X}$ la matriz de diseño (cada fila es un dato).
- En el ejemplo anterior $\mathbf{X}$ es una matriz 72000×6.
- Sea $\mathbf{Y}$ otra matriz resultante de la proyección o transformación lineal $\mathbf{P}$:

$$\mathbf{Y}^T = \mathbf{P}\mathbf{X}^T \qquad \Rightarrow \qquad \mathbf{Y} = \mathbf{X}\mathbf{P}^T$$

- Sea $\underline{\mathbf{p}}_i^T$ una fila de $\mathbf{P}$ (o $\underline{\mathbf{p}}_i$ una columna de $\mathbf{P}^T$)
- Sea $\underline{\mathbf{x}}^{(j)}$ la j-ésima columna de $\mathbf{X}^T$
- Sea $\underline{\mathbf{y}}^{(j)}$ la j-ésima columna de $\mathbf{Y}^T$

Cambio de base (2)

- $\mathbf{P}$ es una matriz que transforma $\mathbf{X}^T$ en $\mathbf{Y}^T$
- Geométricamente, $\mathbf{P}$ es una rotación y estiramiento que transforma columnas de $\mathbf{X}^T$ en columnas de $\mathbf{Y}^T$
- Las filas de $\mathbf{P}$ son la **nueva base** del espacio vectorial, lo que se observa con

$$\mathbf{P}\mathbf{X}^T = \begin{bmatrix} \underline{\mathbf{p}}_1^T \\ \vdots \\ \underline{\mathbf{p}}_n^T \end{bmatrix} \begin{bmatrix} \underline{\mathbf{x}}^{(1)} & \cdots & \underline{\mathbf{x}}^{(m)} \end{bmatrix}$$

$$\mathbf{Y}^T = \begin{bmatrix} \underline{\mathbf{p}}_1^T \underline{\mathbf{x}}^{(1)} & \cdots & \underline{\mathbf{p}}_1^T \underline{\mathbf{x}}^{(m)} \\ \vdots & \ddots & \vdots \\ \underline{\mathbf{p}}_n^T \underline{\mathbf{x}}^{(1)} & \cdots & \underline{\mathbf{p}}_n^T \underline{\mathbf{x}}^{(m)} \end{bmatrix}$$

Cambio de base (3)

es decir, la i-ésima columna de $\mathbf{Y}^T$

$$\underline{\mathbf{y}}^{(i)} = \begin{bmatrix} \underline{\mathbf{p}}_1^T \underline{\mathbf{x}}^{(i)} \\ \underline{\mathbf{p}}_2^T \underline{\mathbf{x}}^{(i)} \\ \vdots \\ \underline{\mathbf{p}}_m^T \underline{\mathbf{x}}^{(i)} \end{bmatrix}$$

se conforma por la **proyección** del i-ésimo vector original $\underline{\mathbf{x}}^{(i)}$ sobre cada uno de los vectores $\underline{\mathbf{p}}_j$

- La preguntas que quedan por responder son ¿cómo deben ser elegidos los vectores de la nueva base? ¿qué es bueno rescatar de $\mathbf{X}$?, o similar ¿qué debería contener $\mathbf{Y}$?

Varianza y valor medio (1)

- El valor medio o promedio $\underline{\boldsymbol{\mu}}$ del conjunto de datos se estima como:

$$\underline{\boldsymbol{\mu}} = \frac{1}{m}\sum_{i=1}^{m}\underline{\mathbf{x}}^{(i)}$$

- La varianza σ^2 se estima como el promedio de desviaciones de los datos con respecto a su valor medio

$$\sigma^2 = \frac{1}{m}\sum_{i=1}^{m}\|\underline{\mathbf{x}}^{(i)} - \underline{\boldsymbol{\mu}}\|_2^2 = \frac{1}{m}\sum_{i=1}^{m}(\underline{\mathbf{x}}^{(i)} - \underline{\boldsymbol{\mu}})^T(\underline{\mathbf{x}}^{(i)} - \underline{\boldsymbol{\mu}})$$

- El cálculo de estas magnitudes se puede realizar con **un** solo recorrido por el conjunto de datos.
- **¿Cómo?**

Varianza y valor medio (2)

$$
\begin{aligned}
\sigma^2 &= \frac{1}{m}\sum_{i=1}^{m}(\underline{\mathbf{x}}^{(i)} - \underline{\boldsymbol{\mu}})^T(\underline{\mathbf{x}}^{(i)} - \underline{\boldsymbol{\mu}}) \\
&= \frac{1}{m}\sum_{i=1}^{m}\left((\underline{\mathbf{x}}^{(i)})^T\underline{\mathbf{x}}^{(i)} - (\underline{\mathbf{x}}^{(i)})^T\underline{\boldsymbol{\mu}} - \underline{\boldsymbol{\mu}}^T\underline{\mathbf{x}}^{(i)} + \underline{\boldsymbol{\mu}}^T\underline{\boldsymbol{\mu}}\right) \\
&= \frac{1}{m}\sum_{i=1}^{m}\left(\|\underline{\mathbf{x}}^{(i)}\|^2 - 2\underline{\boldsymbol{\mu}}^T\underline{\mathbf{x}}^{(i)} + \|\underline{\boldsymbol{\mu}}\|^2\right) \\
&= \frac{1}{m}\sum_{i=1}^{m}\|\underline{\mathbf{x}}^{(i)}\|^2 - 2\underline{\boldsymbol{\mu}}^T\frac{1}{m}\sum_{i=1}^{m}\underline{\mathbf{x}}^{(i)} + \|\underline{\boldsymbol{\mu}}\|^2 \\
&= \frac{1}{m}\sum_{i=1}^{m}\|\underline{\mathbf{x}}^{(i)}\|^2 - \|\underline{\boldsymbol{\mu}}\|^2
\end{aligned}
$$

Ruido y rotación (1)

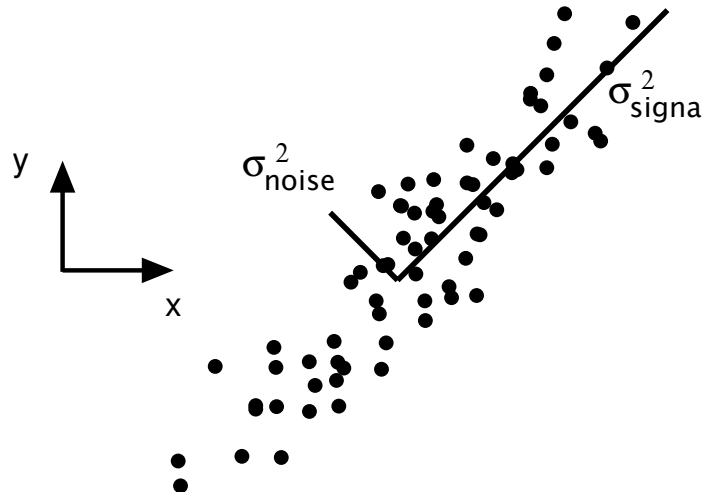
- El ruido de medición en cualquier conjunto de datos debe ser lo suficientemente bajo para poder rescatar la información en ellos.
- La calidad de los datos (o señal) se mide con la tasa de señal a ruido (SNR), descrita usualmente a través de una tasa de varianzas:

$$SNR = \frac{\sigma^2_{\text{señal}}}{\sigma^2_{\text{ruido}}}$$

  es decir, el ruido se mide respecto a la señal, usualmente "potencia" o energía entre ambas.
- SNR $\gg 1$ indica mediciones de alta precisión
- Baja SNR indica datos ruidosos.

Ruido y rotación (2)

- Volviendo al ejemplo del resorte.



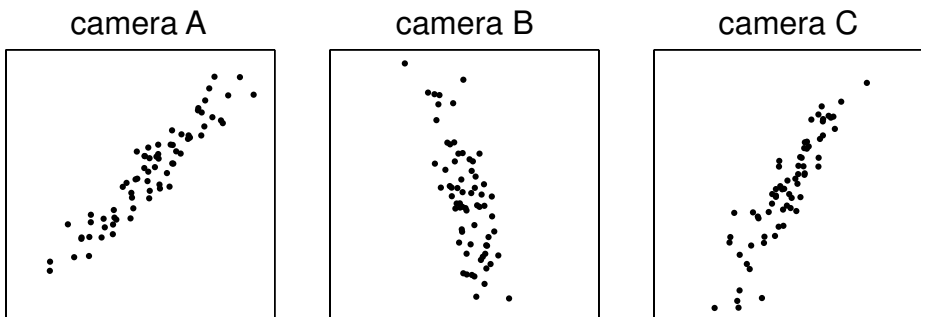
- Se supone que el resorte mueve la bola en una línea recta, así que las tres cámaras deberían observar un movimiento en línea recta.
- Toda desviación de la línea recta es ruido.

Ruido y rotación (3)

- La varianza asociada a la señal y al ruido se demarcan en el diagrama.
- Si se asume que la señal es suficientemente fuerte, entonces la dirección de la **mayor varianza** coincide con la señal.
- Se busca entonces una rotación de los vectores de la base canónica para alinearlos con los ejes de mayor varianza.

Redundancia (1)

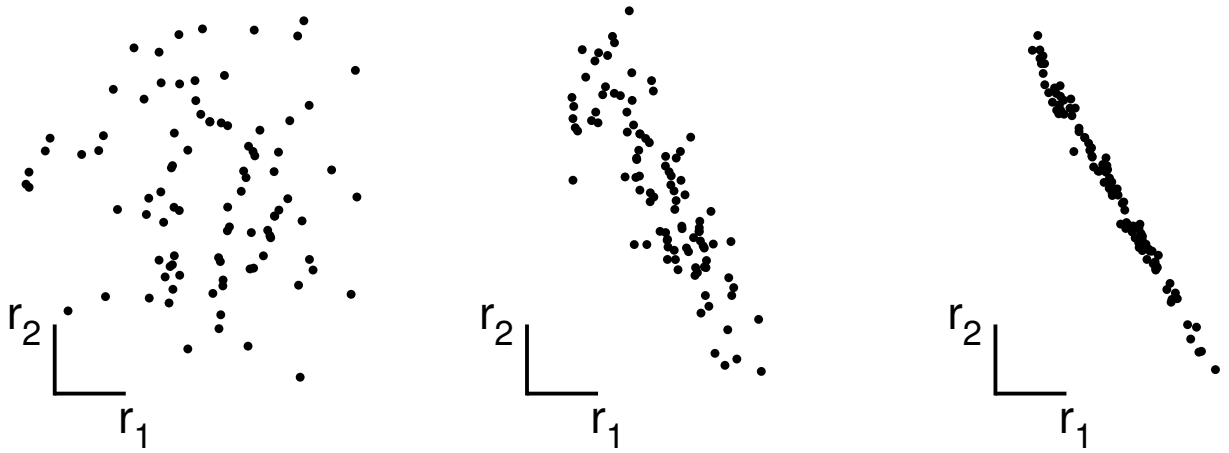
- Otro factor a considerar es la **redundancia**.
- En el ejemplo de las cámaras es notorio, pues cada cámara registra la misma información.



- Obsérvese que incluso en una misma imagen, si sube x, también sube y.

Redundancia (2)

- Existen varios tipos de dependencia entre dos variables r_1 y r_2:



baja redundancia alta redundancia

- En imagen anterior con una dimensión es suficiente para describir los datos altamente redundantes (se puede reducir la dimensión), pero dos dimensiones son necesarias si los datos en r_1 y r_2 no redundan.

Matriz de covarianza (1)

- Considérense dos conjuntos de mediciones con media cero

$$A = \{a_1, a_2, \ldots, a_m\} \qquad B = \{b_1, b_2, \ldots, b_m\}$$

- Las varianzas de los datos A y B por separado son

$$\sigma_A^2 = \frac{1}{m}\sum_i a_i^2, \qquad \sigma_B^2 = \frac{1}{m}\sum_i b_i^2$$

Matriz de covarianza (2)

- La **covarianza** entre A y B se generaliza como

$$\sigma_{AB}^2 = \frac{1}{m} \sum_i a_i b_i$$

- La covarianza mide el grado de dependencia lineal entre dos variables:
  - un valor positivo grande indica datos correlacionados positivamente
  - un valor negativo grande indica datos correlacionados negativamente
  - es cero si A y B no están correlacionados (no hay redundancia)
  - $\sigma_{AB}^2 = \sigma_A^2$ si $A = B$
- El valor absoluto de la covarianza indica el grado de redundancia

Matriz de covarianza (3)

- Expresando A y B como vectores:

$$A \rightarrow \underline{\mathbf{a}}^T = [a_1, a_2, \ldots, a_m]$$
$$B \rightarrow \underline{\mathbf{b}}^T = [b_1, b_2, \ldots, b_m]$$

se replantea la covarianza como

$$\sigma^2_{\underline{\mathbf{ab}}} = \frac{1}{m} \underline{\mathbf{a}}^T \underline{\mathbf{b}}$$

donde recordemos que A y B tienen media cero.

Matriz de covarianza (4)

- Lo anterior se puede generalizar a más vectores y dimensiones.
- Dada la matriz $\mathbf{X}$ de dimensiones $m \times n$

$$\mathbf{X} = \begin{bmatrix} (\underline{\mathbf{x}}^{(1)})^T \\ (\underline{\mathbf{x}}^{(2)})^T \\ \vdots \\ (\underline{\mathbf{x}}^{(m)})^T \end{bmatrix} = \begin{bmatrix} \underline{\mathbf{x}}_{:,1} & \underline{\mathbf{x}}_{:,2} & \cdots & \underline{\mathbf{x}}_{:,n} \end{bmatrix}$$

  - Las **columnas** $\underline{\mathbf{x}}_{:,i}$ de $\mathbf{X}$ corresponden a todas las mediciones de un tipo particular (por ejemplo, eje x de la cámara A).
  - Cada **fila** $(\underline{\mathbf{x}}^{(j)})^T$ de $\mathbf{X}$ corresponde a un dato o muestra, compuesto de varias mediciones.

Matriz de covarianza (5)

- La matriz de covarianza $\mathbf{\Sigma_X}$ de $n \times n$ se define entonces como

$$\mathbf{\Sigma_X} = \frac{1}{m}\mathbf{X}^T\mathbf{X}$$

- El elemento (i,j) de la matriz $\mathbf{\Sigma_X}$ es el producto punto entre el i-ésimo tipo de medición y el j-ésimo tipo.
- Esta matriz es
  - Cuadrada $(n \times n)$ y simétrica
  - La diagonal está formada por las varianzas de cada tipo de medición
  - Los términos fuera de la diagonal son las covarianzas entre tipos de medición.
- Los valores de covarianzas reflejan niveles de ruido y redundancia en las mediciones.

Matriz de covarianza (6)

- En la diagonal, valores de gran magnitud indican estructuras relevantes
- Fuera de la diagonal, valores grandes en magnitud indican alta redundancia.

Diagonalización de la matriz de covarianza (1)

- Es de interés minimizar la redundancia y maximizar la señal de los datos de salida $\mathbf{Y}$, y por lo tanto
  - Todos los datos fuera de la diagonal de $\mathbf{\Sigma_Y}$ deben ser cero, por lo que $\mathbf{\Sigma_Y}$ es diagonal
  - Cada dimensión sucesiva de $\mathbf{Y}$ deberá estar ordenada de acuerdo a la varianza.
- La matriz $\mathbf{\Sigma_X}$ debe ser entonces **diagonalizada**
- El ACP asume que los vectores de la nueva base son ortonormales, y éstos constituyen los **componentes principales**.
- La varianza en cada componente indica qué tan importante es el componente.

Resumen de supuestos

- Se ha supuesto que el problema es **lineal**
  Un cambio de base expresado como producto de matrices puede solucionar el problema.
- Varianzas grandes indican estructura importante.
  Esto supone que se tiene alto SNR, para que las varianzas grandes estén asociadas con la señal.
- Componentes principales son ortogonales.
  Esto permite encontrar una solución utilizando álgebra lineal.

Derivación de diagonalización (1)

- El problema que se plantea es: dado la matriz de diseño $\mathbf{X}$ (matriz $m \times n$, cada fila representa una muestra de n dimensiones, se cuenta con m muestras), encuéntrese una matriz ortogonal $\mathbf{P}$ que transforma los datos a $\mathbf{Y}^T = \mathbf{P}\mathbf{X}^T$, de tal modo que

$$\mathbf{\Sigma_Y} = \frac{1}{n}\mathbf{Y}^T\mathbf{Y}$$

  sea una matriz **diagonal**.
- Las filas de $\mathbf{P}$ son los **componentes principales** de $\mathbf{X}$

Derivación de diagonalización (2)

- Se parte de $\mathbf{\Sigma_Y}$

$$\begin{aligned}
\mathbf{\Sigma_Y} &= \frac{1}{m}\mathbf{Y}^T\mathbf{Y} \\
&= \frac{1}{m}(\mathbf{PX^T})(\mathbf{PX^T})^T \\
&= \frac{1}{m}\mathbf{PX}^T\mathbf{XP}^T \\
&= \mathbf{P}\left(\frac{1}{m}\mathbf{X}^T\mathbf{X}\right)\mathbf{P}^T \\
&= \mathbf{P\Sigma_X P}^T
\end{aligned}$$

que relaciona las matrices de covarianza de datos de entrada y salida

Derivación de diagonalización (3)

- Multiplicando ambos lados a la izquierda con $\mathbf{P}^T = \mathbf{P}^{-1}$

$$\mathbf{P}^T\mathbf{P}\mathbf{\Sigma_X}\mathbf{P}^T = \mathbf{P}^T\mathbf{\Sigma_Y}$$
$$\mathbf{\Sigma_X}\mathbf{P}^T = \mathbf{P}^T\mathbf{\Sigma_Y}$$

  se tiene que las **columnas** de $\mathbf{P}^T$ son los **eigenvectores** de $\mathbf{\Sigma_X}$, y los elementos en la matriz diagonal $\mathbf{\Sigma_Y}$ los **eigenvalores**.

- En otras palabras, las **filas** de $\mathbf{P}$ son los eigenvectores, que se ordenan de acuerdo a sus correspondientes eigenvalores de forma descendiente.

Derivación de diagonalización (4)

- Los **componentes principales** de $\mathbf{X}$ están dados entonces por los eigenvectores de su matriz de covarianza $\mathbf{\Sigma_X} = \frac{1}{m}\mathbf{X}^T\mathbf{X}$
- El i-ésimo valor de la diagonal de $\mathbf{\Sigma_Y}$ es la varianza de los datos $\mathbf{X}$ a lo largo del componente principal $\mathbf{p}_i$.

Pasos del ACP

1. Substraer la media de todos los datos
2. Calcular la matriz de covarianza $\mathbf{\Sigma_X}$
3. Calcular los eigenvectores y eigenvalores de $\mathbf{\Sigma_X}$
4. Construir la matriz de proyección $\mathbf{P}$ compuesta por el número de componentes principales deseados (seleccionados de acuerdo a los eigenvalores)
5. Proyectar los datos con $\mathbf{Y} = \mathbf{X}\mathbf{P}^T$

Derivaciones de PCA

- Hay decenas de formas de derivar el PCA
- Aquí buscamos decorrelar datos produciendo $\mathbf{\Sigma_Y}$ diagonal
- También podríamos buscar el eje que minimiza distancias cuadradas a puntos
- Se puede ver como problema de optimización que busca eje que maximice varianza de proyección

Consideraciones prácticas

Aplicaciones

- Visualización (como en la tarea)
- Compresión
- Reducción de dimensiones para aprendizaje
- Detección de anomalías (distancia a subespacio principal)
- Emparejamiento/Cálculo de distancias (p.ej. similitud entre caras)

Normalización de los datos

- En aplicaciones con datos con componentes disímiles (rangos distintos), se evita sesgo de selección de componentes **normalizando** los datos.
- Por lo general se normaliza primero eliminando la media de cada vector, y luego normalizando el rango de cada columna de $\mathbf{X}$ por separado:

$$\underline{\bar{\mathbf{x}}}^{(i)} = \begin{bmatrix} \frac{1}{\sigma_1} & & & \\ & \frac{1}{\sigma_2} & & \\ & & \ddots & \\ & & & \frac{1}{\sigma_n} \end{bmatrix} (\underline{\mathbf{x}}^{(i)} - \underline{\boldsymbol{\mu}})$$

con σ_i^2 el i-ésimo elemento de la diagonal de $\boldsymbol{\Sigma}_{\mathbf{X}}$

Enblanquecimiento
Whitening

- PCA se usa para reducir dimensiones de vectores de entrada a métodos supervisados y no supervisados de aprendizaje.
- La normalización anterior permite acercar distribuciones de mediciones de distinto rango en cada componente de $\underline{\mathbf{x}}^{(i)}$.
- El **enblanquecimiento** es un proceso de normalización basado en PCA que produce datos decorrelados de varianza 1:

$$\underline{\bar{\mathbf{y}}}^{(i)} = \mathbf{\Sigma}_{\mathbf{Y}}^{-\frac{1}{2}} \mathbf{P}(\underline{\mathbf{x}}^{(i)} - \underline{\boldsymbol{\mu}})$$

  lo que fuerza que $\bar{\mathbf{\Sigma}}_{\mathbf{Y}} = \mathbf{I}$.
- El vector de salida es un vector de "ruido" **blanco**.

Selección de número de dimensiones

- ¿Cómo seleccionar el número de dimensiones a la salida del PCA?
- Tradicionalmente se usa la **razón de varianza total** como medida de selección, dada por:

$$S_k = \frac{\sum_{i=1}^{k} \lambda_k}{\sum_{i=1}^{n} \lambda_k}$$

  Se busca entonces el menor k, tal que S_k supera un umbral de selección.
- A la curva de S_k en función de k se le denomina **espectro** de $\mathbf{\Sigma_X}$.

Recuperación de los datos (1)

- ¿Cómo recupero los datos en el espacio original a partir de sus componentes principales?
- En gran cantidad de aplicaciones es necesario recuperar el dato en el espacio de entrada a partir de su representación "comprimida" en el espacio de salida.
- Para ello basta con expresar el proceso de "compresión":

$$\underline{\mathbf{y}} = \hat{\mathbf{P}}\mathbf{N}(\underline{\mathbf{x}} - \underline{\boldsymbol{\mu}})$$

con $\underline{\mathbf{x}} \in \mathbb{R}^n$ el vector de entrada, $\mathbf{N} \in \mathbb{R}^{n \times n}$ la matriz de normalización, $\hat{\mathbf{P}} \in \mathbb{R}^{k \times n}$ la matriz con $k \leq n$ eigenvectores en sus filas, k componentes principales, y $\underline{\boldsymbol{\mu}}$ la media de los datos.

Recuperación de los datos (2)

- Revirtiendo esta fórmula tenemos el problema de que $\hat{\mathbf{P}}$ desecha información y por tanto no tiene inversa (es rectangular).
- Sin embargo, la matriz $\mathbf{P}$ original completa es ortogonal, con inversa $\mathbf{P}^{-1} = \mathbf{P}^T$.
- Vamos entonces a usar $\hat{\mathbf{P}}^T$ como pseudo-inversa de $\hat{\mathbf{P}}$, lo que nos dará la proyección en el subespacio de k dimensiones "visto" en el espacio de entrada.
- La reconstrucción la hacemos entonces con:

$$\underline{\tilde{\mathbf{x}}} = \mathbf{N}^{-1}\hat{\mathbf{P}}^T\underline{\mathbf{y}} + \underline{\boldsymbol{\mu}}$$

Cálculo de los eigensistemas (1)

- Para calcular los componentes principales requerimos métodos para encontrar los eigenvectores y eigenvalores de la matriz de covarianza $\mathbf{\Sigma_X}$
- Como la matriz de covarianza es simétrica y positiva definida, los métodos estándar para cálculo de eigensistemas producen resultados reales.
- Podemos usar además la descomposición por valores singulares (SVD) directamente:

$$\mathbf{\Sigma_X} = \mathbf{U\Lambda V}^T$$

pues para este tipo particular de matrices, $\mathbf{\Lambda}$ contiene los eigenvalores, y tanto las columnas de $\mathbf{U}$ como de $\mathbf{V}$ los eigenvectores.

Cálculo de los eigensistemas (2)

- Para matrices de gran tamaño (p. ej. $\mathbb{R}^{50\,000 \times 50\,000}$), lo anterior no es práctico, y necesitamos otro truco.
- Supongamos que $m \ll n$. Tomemos la SVD de $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$. Sabemos que las columnas de $\mathbf{U}$ y $\mathbf{V}$ son ortogonales entre sí y que $\mathbf{\Lambda}$ es diagonal.
- Tomemos la matriz de covarianza:

$$\begin{aligned}
\mathbf{\Sigma_X} = \frac{1}{m}\mathbf{X}^T\mathbf{X} &= \frac{1}{m}(\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T)^T\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T \\
&= \frac{1}{m}\mathbf{V}\mathbf{\Lambda}^T\mathbf{U}^T\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T \\
&= \mathbf{V}\left(\frac{1}{m}\mathbf{\Lambda}^T\mathbf{\Lambda}\right)\mathbf{V}^T \\
&= \mathbf{V}\mathbf{D}\mathbf{V}^T
\end{aligned}$$

Cálculo de los eigensistemas (3)

- Esto implica que las columnas de $\underline{\mathbf{V}}$ (¡de la SVD de $\mathbf{X}$!) corresponden también a los eigenvectores de $\mathbf{\Sigma_X}$.
- Los cuadrados de los valores singulares de la SVD de $\mathbf{X} \in \mathbb{R}^{m \times n}$ divididos por m corresponden a los eigenvalores de $\mathbf{\Sigma_X}$.
- (Columnas de $\mathbf{U}$ corresponden a los eigenvectores de $\mathbf{X}\mathbf{X}^T$)
- Basta entonces con seleccionar las primeras k columnas de $\mathbf{V}$ asociadas a los valores singulares mayores para obtener los componentes principales.
- Con $m \ll n$ lo anterior es mucho más eficiente que obtener el eigensistema de $\mathbf{\Sigma_X} \in \mathbb{R}^{n \times n}$.

Cálculo de los eigensistemas (4)

- **Precaución**: herramientas numéricas en este caso de matrices anchas suelen eliminar los bloques nulos de las matrices $\mathbf{U}$ y $\mathbf{\Lambda}$ para ahorrar tiempo y espacio.
- En general, las rutinas de SVD son más estables que las de cálculos de eigensistemas.
- En otra familia de algoritmos, la **regla de Oja** mapea el problema de componentes principales al entrenamiento de una red neuronal, cuyos pesos convergen a los componentes principales.

Cuándo usar qué método

	Modelado $p(\underline{\mathbf{x}})$	No probabilístico
Subespacio	Análisis de factores	PCA
Grupos	Mezcla de gaussianas	k-Medias

Ejemplos

PCA con textos (1)

Latent semantic indexing (LSI)

- La indexación semántica latente es la aplicación de PCA a textos.
- Usamos vectores $\underline{\mathbf{x}}^{(i)} \in \mathbb{R}^n$, con $n > 50000$, cada elemento $x_j^{(i)}$ denota la ocurrencia de la j-ésima palabra en el i-ésimo texto.
- Usualmente no normalizamos los datos.
- Queremos medir con frecuencia similitud entre dos documentos $\underline{\mathbf{x}}^{(i)}$ y $\underline{\mathbf{x}}^{(j)}$.
- Una opción es medir el ángulo entre los dos vectores:

$$\text{sim}(\underline{\mathbf{x}}^{(i)}, \underline{\mathbf{x}}^{(j)}) = \cos\theta = \frac{(\underline{\mathbf{x}}^{(i)})^T \underline{\mathbf{x}}^{(j)}}{\|\underline{\mathbf{x}}^{(i)}\| \|\underline{\mathbf{x}}^{(j)}\|}$$

PCA con textos (2)

Latent semantic indexing (LSI)

- El numerador en nuestra defición de similitud:

$$(\underline{\mathbf{x}}^{(i)})^T \underline{\mathbf{x}}^{(j)} = \sum_{k=1}^{n} x_k^{(i)} x_k^{(j)} = \sum_{k} 1\{\text{docs } i \text{ y } j \text{ tienen palabra } k\}$$

- Con LSI nueva base encuentra correlaciones entre palabras, no visible en los datos originales

Ejemplos

- Eigenfaces
  http://www.youtube.com/watch?v=5nBL_u4MF0k
- AAM http://www.youtube.com/watch?v=I3YsqHCQB4k

Resumen

Este documento ha sido elaborado con software libre incluyendo LaTeX, Beamer, GNUPlot, GNU/Octave, XFig, Inkscape, GNU-Make y Subversion en GNU/Linux