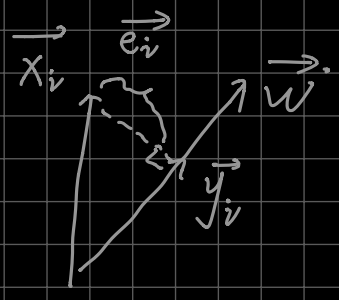
Lecture 8 PCA

Derivation of PCA:

I. Through MSE (Mean Square Error)



$y = \underbrace{W^T}_{\text{PCA变换矩阵}} X \Rightarrow$ 降维后数据

$\vec{e_i} = \vec{x_i} - \underbrace{\vec{y_i} \cdot w}_{\text{重构数据}} \Rightarrow$ 重构损失.

MLE: $J(w) = \frac{1}{m} \cdot \sum_{i=1}^{m} (x_i - (x_i^T w) \cdot w)^2$

$s.t. \quad \|w\| = 1$

Lagrangian Multiplier Method:

$$J(w, \lambda) = \frac{1}{m} \cdot \sum_{i=1}^{m} (x_i - (x_i^T w) \cdot w)^2 - \lambda \cdot (w^T w - 1)$$

$$\frac{\partial J(w, \lambda)}{\partial w} = -2 \cdot \underbrace{\sum_{i=1}^{m} x_i \cdot x_i^T}_{\Sigma_x} w - 2 \cdot \lambda \cdot w = 0$$

$$\Sigma_x \cdot w = -\lambda \cdot w$$

w 为 Σ_x 的 eigenvector (对上式作特征分解得到)

因此 PCA components 即 Σ_x 的 eigen vectors,

也即 w.

II. Through Eigen value Decomposition (特征分解)

$$\Sigma_X = \sum_{i=1}^{N} \cdot X_i \cdot X_i^T \Rightarrow \text{covariance matrix (协方差矩阵)}$$

$$\Sigma_X = U \cdot \Lambda \cdot U^T \Rightarrow \text{特征分解}$$

U 即 PCA components.

III. Through SVD (Singular Value Decomposition)

$$X = U \cdot \Sigma \cdot V^T \Rightarrow SVD$$

where: U is the eigenvector of $X \cdot X^T$ and $U \cdot U^T = I$

and V is the eigenvector of $X^T X$ and $V^T \cdot V = I$

$$\Rightarrow ① X \cdot X^T = U \cdot \Sigma \cdot V^T \cdot V \cdot \Sigma \cdot U^T = U \cdot \Sigma^2 \cdot U^T = \underbrace{U \cdot \Lambda \cdot U^T}_{\text{特征分解}}$$

$$② X^T X = V^T \cdot \Sigma \cdot U^T \cdot U \cdot \Sigma \cdot V = \underbrace{V^T \cdot \Sigma^2 \cdot V}_{\text{特征分解}}$$

U 即 PCA components.

IV. Hebbian learning. Oja. LMSER

$$J(w) = \frac{1}{m}\sum_{i=1}^{m}\underbrace{\left(x_i - (x_i^T w)\cdot w\right)^2}$$

⇓ 实际上为，上式只是另一种写法

$$\left\| x_i - (x_i^T w)\cdot w \right\|_2^2$$

$$\left\| x_i - (x_i^T w)\cdot w \right\|_2^2 = tr\left((x_i - (x_i^T w)\cdot w)^T \cdot (x_i - (x_i^T w) w) \right)$$

$(AB)^T = B^T A^T$

$(ABC)^T = C^T B^T A^T$

$$= tr\left(x_i^T - w^T w^T x_i \right)\left(x_i - (x_i^T w)\cdot w \right)$$

$$= tr\left(x_i^T \cdot x_i - w^T w^T x_i \cdot x_i - x_i^T x_i^T w w + w^T w^T x_i \cdot x_i^T w w \right)$$

tr中可移动

$$= tr(x_i^T x_i) - tr(w^T w^T x_i x_i) - tr(x_i^T x_i^T w w) + tr(w^T x_i x_i^T w)$$

$w^T x_i \, x_i^T w = w^T x_i \cdot x_i^T w$ ； $w \cdot x_i^T x_i^T w = w^T x_i \cdot x_i^T w$

$$= \underbrace{tr(x_i^T x_i)}_{\Downarrow \text{常数}} - tr(w^T \cdot x_i \cdot x_i^T \cdot w)$$

$$= C - \left\| w^T x_i \right\|_2^2$$

$$= C - y_i \cdot y_i^T$$

$$\therefore J(w) = \frac{1}{m}\cdot\sum_{i=1}^{m}\left(x_i - (x_i^T w)\cdot w\right)^2$$

$$= C - \sum_{i=1}^{m} y_i \cdot y_i^T$$

$$= C - \underbrace{\Sigma_y}_{\text{covariance of } y}$$

因此 PCA minimize $J(w)$ 相当于 maximize 变换后数据的方差.

Which vector is the PCA solution?

$$\because \quad C = -\mathrm{tr}(W^T X \cdot X^T W)$$

$$= C - \mathrm{tr}(W^T \cdot \Sigma_X \cdot W)$$

$$= C - \mathrm{tr}(W^T \cdot \lambda \cdot W)$$

上式的 $\Sigma_X \cdot W = (\lambda) W = \lambda W$

$$= C - \mathrm{tr}(\lambda \cdot W \cdot W^T)$$

$$= C - \lambda \quad (\|w\| = 1)$$

$\therefore$ minimize $J(w) \Leftrightarrow C - \lambda \Leftrightarrow$

maximize λ，因此 PCA 的 solution 为 eigen value

λ 的前 k 个。

PCA component.

△ 此也即：$\dfrac{\sum_{i=1}^{k} (x_i - W^T x)^2}{\sum_{i=1}^{m} x_i^2} \le 0.01 \Longleftrightarrow \dfrac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{m} \lambda_i} \ge 0.99$

的证明方法。

Lecture 9 FA

Why FA:

从概率生成模型的角度思考PCA.

y为一个低维的向量，x为原始的数据，若能找这个隐变量y使其线性变换后 $(x = A^T y + u + e)$ 能模拟原始的分布，那么这个y即可代表降维后的数据。

因此FA的目的即找此y满足上述，方法为EM。

Derivation:

$$y \sim G(0, \Sigma_y), \quad \Sigma_y \to I^2, \quad e \sim G(0, \sigma^2 I)$$

y为隐变量 ; e与y独立，为noise

$$P(x|\theta) \sim G(u, AA^T + \sigma^2 I)$$

$\Downarrow$

> $Y = AX + b$，若 $X \sim N(u, \Sigma)$，则 $Y \sim N(Au + b, A\Sigma A^T)$

$\because x = Ay + u + e$，因此依上式得此.

E step:

$$P(y|x) = \frac{P(y|\theta) \cdot P(x|y,\theta)}{P(x|\theta)} = \frac{G(0, \Sigma_y) \cdot G(Ay+u, \sigma^2 I)}{G(u, A^TA + \sigma^2 I)}$$

Conditional Gaussian:

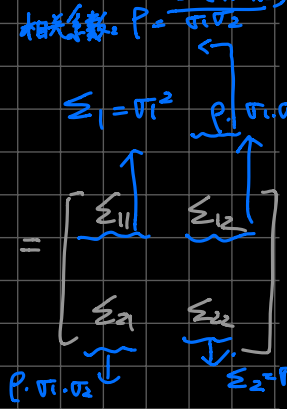
$$x_1 \sim N(u_1, \Sigma_1), \quad x_2 \sim N(u_2, \Sigma_2)$$

$(x_1 | x_2 = a) \sim N(\bar{u}, \bar{\Sigma})$，where:

条件高斯

$$\begin{cases} \bar{u} = u_1 + \Sigma_{12} \cdot \Sigma_{22}^{-1} (a - u_2) \\ \bar{\Sigma} = \Sigma_{11} - \Sigma_{12} \cdot \Sigma_{22}^{-1} \Sigma_{21} \end{cases}$$

$$\Rightarrow \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$



$\Rightarrow$

$E(y|x) = \bar{u} = 0 + A^T(AA^T + \sigma^2 I)^{-1} \cdot x$

A^T: $\mathrm{cov}(Y, X) = A^T$

$(AA^T + \sigma^2 I)^{-1}$ 令其为 W

$= Wx$

$E(yy^T|x) = \Sigma + E^2(y|x)$

$= I - W \cdot A + Wx \cdot (Wx)^T$

$(Wx)^T = W_x \cdot x^T \cdot W^T$

M step:

$\underset{\theta}{\arg\max}\ P(y, x|\theta)$

u, σ, A

$\sum_{i=1}^{N} \log \cdot P(y_i, x_i|\theta)$

$= \sum_{i=1}^{N} \cdot \log \cdot \frac{P(y_i, x_i|\theta) \cdot P(y_i|x_i;\theta)}{P(y_i|x_i;\theta)}$

$= \sum_{i=1}^{N} \cdot P(y_i|x_i;\theta) \cdot \log \frac{P(y_i|\theta) \cdot P(x_i|y_i;\theta)}{P(y_i|x_i;\theta)}$

General EM 可考虑的部分

$Q(y_i)$ / $Q(z_i)$

隐变量

$$\theta = \arg\max_{\theta} \sum_i \cdot \int_{z^{(i)}} Q_i(z^{(i)}) \cdot \left(\log \frac{P(x^{(i)}, z^{(i)}|\theta)}{Q_i(z^{(i)})}\right.$$

$$= \sum_i \cdot E_{z^{(i)} \sim Q(i)} \left[\log \frac{P(x^{(i)}, z^{(i)}|\theta)}{Q_i(z^{(i)})}\right] \Rightarrow \text{Gaussian form: } e^{(\cdot)}$$

$$= \left(\sum_i (\cdot)^2\right)'$$

⇓

Square form, can directly take $\Delta u \theta$

Lecture 10 ICA

- Independence: $P(y_1, y_2) = P(y_1) \cdot P(y_2)$ || $E(h_1(y_1) \cdot h_2(y_2)) = E(h_1(y_1)) \cdot E(h_2(y_2))$
  ⇒ no information can be gained from each other.
- Uncorrelatedness: $E(y_1 y_2) = E(y_1) \cdot E(y_2)$
- Conditional Independence: $P(X, Y | Z) = P(X|Z) \cdot P(Y|Z)$
- Dependence
  - Common cause
  - Causal relation
  - Common effect (conditional dependence)

Gaussian Distribution properties:

$E(XY) = E(X) \cdot E(Y)$

① Uncorrelatedness ⇔ Independence $\quad E(f(x) \cdot g(y)) = E(f(x)) \cdot E(g(y)), \forall f, g$

② Central Limit Theorem (CLT) means: $\sum_{i=1}^{n} X_i = S$,

if $X_1, X_2, \dots, X_n$ are independent random variables, the S is approximately Gaussian Distribution.

③ Cramer's decomposition theorem: $\xi = \xi_1 + \xi_2$.

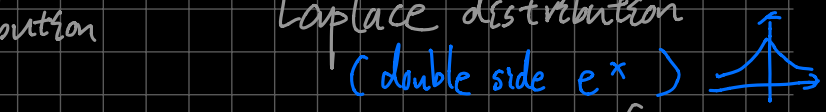
if ξ is normally distributed && ξ_1, ξ_2 are independent random variables

Gaussian distribution

, then s_1, s_2 are both normally distributed.

④ Sub-Gaussian and Super-Gaussian ⇒

e.g.: uniform distribution

Laplace distribution (double side e*) 

Vector is noiseless, then it's more Super-Gaussian;

Otherwise, if a large amount of noise exists, then it's more tend to Gaussian distribution.

⑤ A gaussian variable has the largest entropy among all random variables of equal variance. ⇒

More noise, larger entropy (according to ④)

Measures of non-Gaussian

- Kurtosis: $kurt(y) = E(y^4) - 3E^2(y^2)$

  ⇒ $kurt(y) > 0 \Leftrightarrow$ super-Gaussian

  $kurt(y) < 0 \Leftrightarrow$ sub-Gaussian

- Negentropy: $J(y) = H(y_{gauss}) - H(y)$, where $H(y) = -\int f(y)\cdot \log f(y)\cdot dy$ or $-\sum f(y)\cdot \log f(y)$.

  $\Downarrow$ $I(y_1, y_2 \cdots y_n) = C - \sum_{i=1}^{n} J(y_i)$

- Mutual Information: $I(y_1, y_2 \cdots y_n) = KL(P(y_1, y_2 \cdots y_n) \,|\, P(y_1)\cdot P(y_2)\cdots P(y_n))$

  $= \sum_{i=1}^{m} H(y_i) - H(y)$

C closer to 0 ⇒ more Gaussian / less non-gaussian

Independent Component Analysis (ICA)

Intuition:

$x = A \cdot y \Leftrightarrow A^{-1} \cdot x = y \Leftrightarrow y = W^T \cdot x = W^T \cdot A \cdot s = Z^T \cdot s$

- A: transform matrix
- x: row data (观测数据)
- y: original independent components (target data)
- Z^T: more Gaussian than any of the s_i

Thus, we want to find w that maximizes the non-Gaussianity measured by negentropy

$\Rightarrow$ ICA allows ≤ 1 Gaussian variables. (complete symmetry, no information on the direction of A) : 当 > 1 个原始 s_i 中

Fast ICA algorithm:

① Randomly initialize w

② while w is not converged:

③ $w = E[x \cdot g_1(w^T x)] - E[g_2(w^T x)]$

where $\begin{cases} g_1(x) = \frac{1}{a_1} \cdot \log\cosh(a_1 x) \\ g_2(x) = x \cdot e^{-\frac{x^2}{2}} \end{cases}$

get from the derivative of MLE.

④ $w = \frac{w}{\|w\|}$ # normalized

end

CDF (Cummulative Distribution Functions)

PDF (Probability Distribution Function)