Dimensionality Reduction

2021年5月11日 16:06

P22 Background

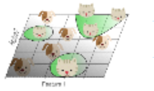
Overfitting → ① Add data ② Regularization ③ Dimensionality reduction

DR → ① Feature Selection ② Linear DR (PCA, MDS) ③ Nonlinear DR

Curse of dimensionality: Cause overfitting and data sparsity

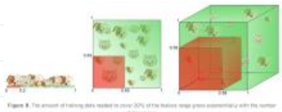
Increasing the dimensionality may lead to better result. But at the same time, it will make number of samples per unit smaller. (Data Sparsity ①)

Obtaining a better classifier by increasing dimensionality is equivalent to using a more complex nonlinear classifier in low-dimensional space



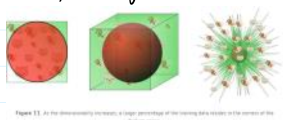
(Data Sparsity ②): As the dimensionality increase, more data is needed for each dimension to cover the proportion of space

This means that increasing the dimensionality also requires exponentially increasing the data



Data Sparsity ③: (Data is distributed differently in high-dimensional space) Assuming that the mean of the eigenvalues is take as the centroid of the space in two dimensions

Make a circle with the feature range as the diameter, and the samples outside the circle are distributed in the corners of the space.



The characteristics of the data in the corners are widely disparate. These data is more difficult to classify.

$V_{tesseract}^{(超立方体)} = 1 \quad V_{n\text{-}sphere}^{(超球体)} = k(0.5)^D \quad D\to\infty \quad \therefore V_{n\text{-}sphere}\to 0$ ∴ Almost all of the samples are distributed in the corners.

dist: Euclidean distance from the sample point to the center of the space. $\lim_{D\to\infty}\frac{dist_{max}-dist_{min}}{dist_{min}} = 0$ ∴ All samples in high-dimensional space are far from the center.

∴ No difference in maximum and minimum distance

∴ Euclidean, Manhattan, Mahalanobis distance and other methods gradually fail in high-dimensional space.

P23 Sample Mean & Variance

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X^i = \frac{1}{N}\underbrace{(x^1, x^2 \cdots x^N)}_{X^T}\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}_{N\times 1} = \frac{1}{N}X^T 1_N$$

$$S = \frac{1}{N}\sum_{i=1}^{N}(x^i-\bar{x})(x^i-\bar{x})^T \Rightarrow \sum_{i=1}^{N}(x^i-\bar{x}) = (x^1-\bar{x}\ \ x^2-\bar{x}\ \cdots\ x^N-\bar{x}) = X^T - \bar{x}1_N^T = X^T(I_N - \frac{1}{N}1_N 1_N^T)$$

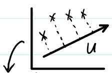
Let $I_N - \frac{1}{N}1_N 1_N^T = H_{(n\times n)}$ ⇒ Centering Matrix $\therefore S = \frac{1}{N}X^T H H^T X$ $\quad H = \begin{bmatrix} \frac{n-1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & \frac{n-1}{n} & & \vdots \\ \vdots & \vdots & \ddots & \\ -\frac{1}{n} & -\frac{1}{n} & \cdots & \frac{n-1}{n} \end{bmatrix}$

$\because H^T = H \quad H^n = H \quad \therefore S = \frac{1}{N}X^T H X$

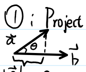
P24 PCA - Maximum Projection Variance

PCA main task: Reconstructing the original eigenspace

Two methods: ① Maximum Projection Varience ② Minimum Reconfiguration Cost



① : Project the data onto linearly independent basis and make the variance of the projected point set as small as possible.



$\vec{a}\cdot\vec{b} = |\vec{a}||\vec{b}|\cos\theta \quad \therefore \|u\| = 1 \quad a^T u = \|a\|\cos\theta$, which is the projection of a on u

First make X zero-centered ⇒ $x^i - \bar{x} \Rightarrow E(x-\bar{x}) = 0 \Rightarrow E[(x-\bar{x})^T u] = E[(x-\bar{x})^T]E[u] = 0$

$$J = D[(x-\bar{x})^T u] = E[((x-\bar{x})^T u)^2] - 0 = \frac{1}{N}\sum_{i=1}^{N}[(x^i-\bar{x})^T u]^2 \quad s.t.\ u^T u = 1$$

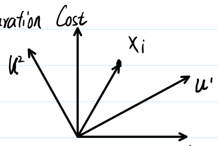
$$= u^T \sum_{i=1}^{N}\left[\frac{1}{N}(x^i-\bar{x})(x^i-\bar{x})^T\right]u = u^T S u$$

$$\therefore \begin{cases}\hat{u} = \arg\max_u u^T S u \\ s.t.\ u^T u = 1\end{cases} \qquad \mathcal{L}(u,\lambda) = u^T S u + \lambda(1-u^T u) \qquad \frac{\partial \mathcal{L}}{\partial u} = 2Su - 2\lambda u = 0$$

$\therefore S\hat{u} = \lambda\hat{u}$ ∴ PCA is to calculate the eigenvetors of S and sort them by eigenvalues.

And select the first q as required.

P25 Minimum Reconfiguration Cost



$$x^i = (\underbrace{x^{iT}u^1}_{\text{The coordinates of } x^i \text{ on } u^1})u^1 + (x^{iT}u^2)u^2 \qquad u^{1T}u^1 = 1$$

Original P dimensions: $x^i = \sum_{k=1}^{p}[(x^i-\bar{x})^T u^k]u^k$ Left q dimensions: $\hat{x}^i = \sum_{k=1}^{q}[(x^i-\bar{x})^T u^k]u^k$

② Reconfiguration cost: $J = \frac{1}{N}\sum_{i=1}^{N}\|x^i-\hat{x}^i\|^2 = \frac{1}{N}\sum_{i=1}^{N}\left\|\sum_{k=q+1}^{p}[(x^i-\bar{x})u^k]u^k\right\|^2 = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=q+1}^{p}[(x^i-\bar{x})u^k]^2$

$$= \sum_{k=q+1}^{p} u^{kT} S u^k$$

$\begin{cases} u^k = \arg\min_{u_k}\sum_{k=q+1}^{p} u^{kT} S u^k \\ s.t.\ u^{kT}u^k = 1\end{cases}$ ∵ Linearly independent among $u^{kT}Su^k$ ∴ Solve optimazation problems one by one.

① find the q principal components, while ② find the p-q principal components to be deleted

P26. Singular Value Decomposition (SVD) and PCA & PCoA

① $H_N = I_N - 1_N 1_N^T$ $\quad HX$ (centralized) $= U\Sigma V^T \Rightarrow$ SVD $\begin{cases} U^T U = I \\ V^T V = VV^T = I \\ \Sigma \text{ diagonal matrix}\end{cases}$ (V is eigenvetor)

② $S_{p\times p} = \frac{1}{N}X^T H X = \frac{1}{N}X^T H^T H X = \frac{1}{N}V\Sigma U^T U\Sigma V = \frac{1}{N}V\Sigma^2 V^T$ ∴ S can be calculated from the SVD of the centralized data.

③ let $T = HXX^TH^T = U\Sigma V^T V\Sigma U^T = U\Sigma^2 U^T$

① $H_N = I_N - \frac{1}{N} 1_N 1_N^T$ HX (centralized) $= U \Sigma V^T \Rightarrow$ SVD $\begin{cases} V^T V = V V^T = I \\ \Sigma \text{ diagonal matrix} \end{cases}$ (V is eigenvetor)

② $S_{p\times p} = \frac{1}{N} X^T H X = \frac{1}{N} X^T H^T H X = \frac{1}{N} V \Sigma U^T U \Sigma V = \frac{1}{N} V \Sigma^2 V^T$ $\therefore S$ can be calculated from the SVD of the centralized data.

③ Let $T_{N\times N} = HXX^T H^T = U \Sigma V^T V \Sigma U^T = U \Sigma^2 U^T$

S: Calculate the direction (principal component V), and then calculate the coordinates under the new stage b $\overset{N\times P}{HX} \cdot \overset{P\times P}{V} \Leftarrow$ PCA

T: Find the coordinates directly $\Rightarrow HX \cdot V = U \Sigma V^T V = U \Sigma \Leftarrow$ (principle coordinate analysis PCoA)

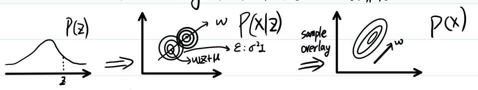
$\because T \underbrace{U\Sigma}_{\text{eigenvector}} = U \Sigma^2 U^T U \Sigma = U \Sigma \underbrace{\Sigma^2}_{\text{eigenvalue}}$ $\therefore U\Sigma$ is coordinate matrix

When the data dimension is too large ($P\times P$ is too large), PCoA (T) is better

P27 Probabilistic Perspective PCA (P-PCA)

$X \in \mathbb{R}^P$ $Z \in \mathbb{R}^q$ $q < P$ Let $Z \sim N(0_q, I_q)$ $X = \overset{(P\times q)}{W} Z + \mu + \varepsilon$

$\varepsilon \sim N(0, \sigma^2 I_P)$ P-PCA $\begin{cases} \text{Inference}: P(Z|X) \\ \text{Learning}: W, \mu, \sigma = \underset{W,\mu,\sigma}{\text{argmax}} \log P(Z|X) \end{cases}$ No calculation here



Each z determines a Gaussion distribution with $wz+\mu$ as expectation and $\varepsilon(\sigma^2 I)$ as variance.

$E(X|Z) = E(WZ + \mu + \varepsilon)$ $\because$ find $X|Z$ $\therefore Z$ is constant now $\therefore E(WZ+\mu+\varepsilon) = WZ + \mu$

$Var(X|Z) = Var(WZ + \mu + \varepsilon) = \sigma^2 I$ $X|Z \sim N(WZ+\mu, \sigma^2 I)$

$E(X) = E(WZ+\mu+\varepsilon) = \mu$ $Var(X) = Var(WZ+\mu+\varepsilon) = WIW^T + \sigma^2 I = WW^T + \sigma^2 I$

Review P7: $X = \begin{pmatrix} X_a \\ X_b \end{pmatrix}$ $X \sim N\left(\begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}, \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}\right)$ $\begin{cases} X_{b\cdot a} = X_b - \Sigma_{ba}\Sigma_{aa}^{-1} X_a \\ \mu_{b\cdot a} = \mu_b - \Sigma_{ba}\Sigma_{aa}^{-1}\mu_a \\ \Sigma_{bb\cdot a} = \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab} \end{cases}$

$\therefore X_{b\cdot a} \sim N(\mu_{b\cdot a}, \Sigma_{bb\cdot a})$

$X_b = X_{b\cdot a} + \Sigma_{ba}\Sigma_{aa}^{-1} X_a$ $E(X_b|X_a) = \mu_{b\cdot a} + \Sigma_{ba}\Sigma_{aa}^{-1} X_a = \mu_b + \Sigma_{ba}\Sigma_{aa}^{-1}(X_a - \mu_a)$ it means X_a is constant now

$Var(X_b|X_a) = \Sigma_{bb\cdot a}$ $\therefore X_b|X_a \sim N(\mu_b + \Sigma_{ba}\Sigma_{aa}^{-1}(X_a-\mu_a), \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab})$

$cov(X, Z) = E[(X-\mu)(Z-0)^T] = E[(WZ+\varepsilon)Z^T] = E(WZZ^T) + E(\varepsilon Z^T) = W Var(Z) = WI = W$

$\therefore P(Z|X)$ is found from the above equation and the calculated $E(X)$ $E(Z)$ $Var(X)$ $Var(Z)$ $cov(X,Z)$

Difference with GMM $\Rightarrow$ GMM: $Z \rightarrow$ discrete: $1, 2 \cdots k$ P-PCA: $Z \rightarrow$ continuous: $Z \sim N$

X $X|Z \sim N$ X $X|Z \sim N$