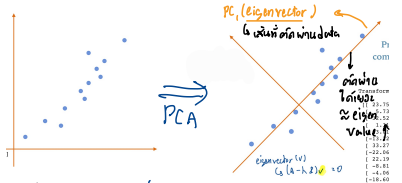
Reduced Dim

(1) PCA : หาแกนที่ทำให้ Var เยอะสุด. (Var มากสุดดี)



note eigenvalue (λ) & covar(x_i, x_j)

PCA step

(1) Standardize dataset (เพื่อให้แปรผลได้)

$$x_{new} = \frac{x - \mu}{\sigma}$$ (optional)

(2) Calculate the Covariance matrix

↳ Covar คือ บอกได้ว่าตัวแปรคู่ไหน มี trend ไปด้วยกัน (Relationship)

For Sample

$$\text{Cov}(x,y) = \frac{\Sigma (x_i - \bar{x}) * (y_i - \bar{y})}{(N-1)}$$

① Std Dataset. (optional)

f1	f2	f3	f4
-1	-0.63246	0	0.26062
0.33333	1.26491	1.73205	1.56374
-1	0.63246	-0.57735	-0.17375
0.33333	0	-0.57735	-1.04249
1.33333	-1.26491	-0.57735	-0.60812

② covar.

	f1	f2	f3	f4
f1	var(f1)	cov(f1,f2)	cov(f1,f3)	cov(f1,f4)
f2	cov(f2,f1)	var(f2)	cov(f2,f3)	cov(f2,f4)
f3	cov(f3,f1)	cov(f3,f2)	var(f3)	cov(f3,f4)
f4	cov(f4,f1)	cov(f4,f2)	cov(f4,f3)	var(f4)

	f1	f2	f3	f4
f1	0.8	-0.25298	0.03849	-0.14479
f2	-0.25298	0.8	0.51121	0.4945
f3	0.03849	0.51121	0.8	0.75236
f4	-0.14479	0.4945	0.75236	0.8

(3) eigenvalue (λ) & eigenvector (V)

① eigenvalue (λ)

$\det(A - \lambda I) = 0$; A = Matrix, I = Identity [1 0; 0 1]

② eigenvector (V) (ต่อ)

$(A - \lambda I)v = 0$ if $\lambda = 2.516$

$$\begin{pmatrix} 0.800000-\lambda & -0.252982 & 0.038490 & -0.144791 \\ -0.252982 & 0.800000-\lambda & 0.511208 & 0.494498 \\ 0.038490 & 0.511208 & 0.800000-\lambda & 0.752355 \\ -0.144791 & 0.494498 & 0.752355 & 0.800000-\lambda \end{pmatrix} \times \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix} = 0$$

Result

	λ=2.516 e1	λ=1.06 e2	e3	e4
v_1	0.161960	-0.917059	-0.307071	0.196162
v_2	-0.524048	0.206922	-0.817319	0.120610
v_3	-0.585896	-0.320539	0.188250	-0.720099
v_4	-0.596547	-0.115935	0.449733	0.654547

$\lambda = 2.51579324$, 1.0652885 , 0.39388704 , 0.02503121

(4) เรียงค่า λ มาก-น้อย & V ที่เกี่ยวข้อง.

(5) เลือก V จากเกณฑ์ในส่วน ตามจำนวน n_component

(6) Transform to Xpc

f1	f2	f3	f4
-1.000000	-0.632456	0.000000	0.260623
0.333333	1.264911	1.732051	1.563740
-1.000000	0.632456	-0.577350	-0.173749
0.333333	0.000000	-0.577350	-1.042493
1.333333	-1.264911	-0.577350	-0.608121
			(5,4)

×

e1	e2
0.161960	-0.917059
-0.524048	0.206922
-0.585896	-0.320539
-0.596547	-0.115935
	(4,2)

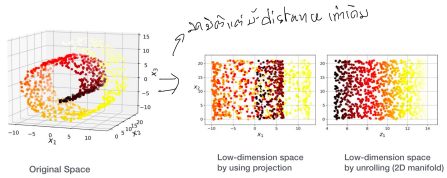
=

nf1	nf2
0.014003	0.755975
-2.556534	-0.780432
-0.051480	1.253135
1.014150	0.000239
1.579861	-1.228917
	(5,2)

n_component

Proportion of total variance explained by first r PCs: $f(r) = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_r}{\lambda_1 + \lambda_2 + \cdots + \lambda_q}$ (q = total)

Manifold learning



↳ ลดมิติข้อมูลแต่ยังคง distance เหมือนเดิม

2 type

- Linear manifold learning uses direct pairwise distances in original space.
  - ↳ number of pair = $\frac{n(n-1)}{2}$
- Non-linear manifold learning uses geodesic distance or gives more weights to closer neighbors in original space.

Clustering

Basic

(1) Prepare data

- num → scaling $\frac{X_i - \min(X_i)}{\max(X_i) - \min(X_i)}$
- Ordinal → varible, i | value: Small, 1 → 1/6; Medium, 2 → 3/6; Large, 3 → 5/6. M = 3, $v = \frac{i - 1/2}{M}$
- normimal (one-hot)

Black	0	0	1
Blonde	0	1	0
Brown	1	0	0

(-) Similarity :

cosine = $\cos(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{\boldsymbol{x}_1 \cdot \boldsymbol{x}_2}{\|\boldsymbol{x}_1\| \|\boldsymbol{x}_2\|}$

$\text{Similarity}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{1}{1 + \text{d}(\boldsymbol{x}_1, \boldsymbol{x}_2)}$ or $\exp(-\text{d}(\boldsymbol{x}_1, \boldsymbol{x}_2))$ or $1 - \frac{\text{d}(\boldsymbol{x}_1, \boldsymbol{x}_2)}{\max \text{d}}$

$\text{Dissimilarity}(\boldsymbol{x}_1, \boldsymbol{x}_2) = 1 - \text{Similarity}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ (euclidian distance)

$x_1 = \{\text{Black}, \text{C1}\} \rightarrow \{(0,0,1), (0,0,1)\} = \{x_{11}, x_{12}\}$
$x_2 = \{\text{Brown}, \text{C2}\} \rightarrow \{(1,0,0), (1,0,0)\} = \{x_{21}, x_{22}\}$
$x_3 = \{\text{Brown}, \text{C3}\} \rightarrow \{(1,0,0), (1,0,0)\} = \{x_{31}, x_{32}\}$

จำนวน columns

$d(\mathbf{x}_1, \mathbf{x}_2) = m - \sum_{k=1}^{m} x_{1k} \cdot x_{2k} = 2 - 0 - 0 = 2$

$d(\mathbf{x}_1, \mathbf{x}_3) = 2 - 0 - 0 = 2$

$d(\mathbf{x}_2, \mathbf{x}_3) = 2 - 1 - 0 = 1$

nominal Distance.

(-) Distance of Mixtype

Numeric n/a, Ordinal 3/a, Nominal 2/a

	Age	Height	Weight	Gender					
Person 1	45	170	64	1					
Person 2	31	172	77	1					
Person 3	29	159	56	0					
Person 4	64	163	72	1					
..	..	..	..	..	..	..	..	..	..
..	..	..	..	..	..	..	..	..	..

↳ $D(P_1, P_2) = \frac{n/a \cdot D(P_1,P_2) + 3/a \cdot D(P_1,P_2) + 2/a \cdot D(P_1,P_2)}{q}$ (num, ordinal, nominal)

q ⇒ num of features

(-) Centroid → mode (Nominal), mean (Num, Ord)

(2) Kmean

SSE

$$\text{SSE}(\mathcal{C}) = \sum_{k=1}^{K} \sum_{\boldsymbol{x}_j \in C_k} \|\boldsymbol{x}_j - \boldsymbol{\mu}_k\|^2$$

(3) หาจำนวน k-clustering

① elbow Method : ดำนวน SSE ของแต่ละ K-cluster

↳ >> 0 ดี ; เลือกตาม SSE พอพอใจ ที่ k ตัวสุด.

② Silhouette coef (Avg cs)) : $\frac{1}{n} \sum_{i=1}^{n} \frac{b_i - a_i}{\max(a_i, b_i)}$

— Average distance to all other data points within the same cluster (cohesion). a_i

— Average distance to all data points in the nearest neighboring cluster (separation). b_i

③ Calinski = $S = \frac{B_K/(K-1)}{W_K/(N-K)}$

$B_K = \sum_{q=1}^{K} N_q \|\boldsymbol{c}_q - \boldsymbol{c}_G\|^2$ Between-cluster dispersion (global center, cluster center)

$W_K = \sum_{q=1}^{K} \sum_{\boldsymbol{x} \in C_q} \|\boldsymbol{x} - \boldsymbol{c}_q\|^2$ Within-cluster dispersion

Note

④ Hopskins : $H = \frac{\sum_{i=1}^{p} w_i}{\sum_{i=1}^{p} u_i + \sum_{i=1}^{p} w_i}$ (w_i = distance(x_i, nearest(x_{i1})), u_i = เลือกมาจาก Var หนูข้อมูลจริง.)

↳ H > 0.5 มีcluster

↳ H < 0.5 ไม่มีcluster

↳ Uniform data. ไม่มี cluster เพราะ ทุกจุดมีระยะห่างเท่ากัน.

Advance

limitation of KMean → สมมุติว่าแต่ละ cluster เป็นทรงกลม และขนาดเท่ากัน

(1) GMM → กลุ่ม KMean โดยคลายแน่จาก distributions

↳ 1 cluster ⇒ 1 gaussian dist (bell shape)

① Univariate → single feature (1 D-Dim)

Input: n data points $\{x_1, x_2, \ldots, x_n\}$ and the number of clusters K
Output: K model parameters $\{(\mu_k, \sigma_k)\}$, $1 \le k \le K$

Initialization
Set $\hat{\mu}_k$ to be randomly chosen from dataset.
$\hat{\sigma}_k^2 \leftarrow \frac{1}{n}\sum_i (x_i - \bar{x})^2$
$w_k \leftarrow 1/K$

$\theta_i = (\mu_i, \sigma_i)$
$f(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

repeat
  Expectation Step
  Compute for all data points x_i, $w_{ki} \leftarrow \frac{w_k f(x_i|\theta_k)}{\sum_{j=1}^{K} w_j f(x_i|\theta_j)}$, $\forall k$
  $w_k \leftarrow (1/n)\sum_i w_{ki}$, $\forall k$

  Maximization Step
  $\hat{\mu}_k \leftarrow \frac{1}{n}\sum_i \frac{w_{ki}}{w_k} x_i$
  $\hat{\sigma}_k^2 \leftarrow \frac{1}{n}\sum_i \frac{w_{ki}}{w_k}(x_i - \hat{\mu}_k)^2$
until $\hat{\mu}_k$ converges
return

$L(\Theta, w|X) = \prod_{i=1}^{n} \sum_{k=1}^{K} w_k f(x_i; \theta_k)$

② Multivariate : (n feature, n Dim)

Input: n data points $X = \{x_1, x_2, \ldots, x_n\}$, $x_i \in \mathbb{R}^d$, $X \in \mathbb{R}^{d \times n}$, and the number of clusters K
Output: K model parameters $\{(\mu_k, \Sigma_k)\}$, $1 \le k \le K$, $\mu_k \in \mathbb{R}^d$, $\Sigma_k \in \mathbb{R}^{d \times d}$

Initialization
Set $\hat{\mu}_k$ to be randomly chosen from dataset.
$\hat{\Sigma}_k \leftarrow \frac{1}{n}(X - \bar{x})(X - \bar{x})^T$
$w_k \leftarrow 1/K$
$\theta_k \leftarrow (\hat{\mu}_k, \hat{\Sigma}_k)$

repeat
  Expectation Step
  Compute for all data points x_i, $w_{ki} \leftarrow \frac{w_k f(x_i|\theta_k)}{\sum_{j=1}^{K} w_j f(x_i|\theta_j)}$
  $w_k \leftarrow (1/n)\sum_i w_{ki}$

  Maximization Step
  $\hat{\mu}_k \leftarrow \frac{1}{n}\sum_i \frac{w_{ki}}{w_k} x_i$
  $\hat{\Sigma}_k \leftarrow \frac{1}{n}\frac{\sum_{i=1}^{n} w_{ki}(x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T}{w_k}$
until $\hat{\mu}_k$ converges
return μ_k, cover, w_k

parameter in each K
↳ μ_k = n
↳ covar = n, $\frac{n(n-1)}{2}$
weight in each k
w_k = K
K cluster

num params = $k(\mu_k + 6) + K$

(2) Hierchical Clustering

- MIN (Single linkage)
- MAX (Complete linkage)
- Group Average (Average linkage)
- Ward's Method

① find distance Pair & ② find min

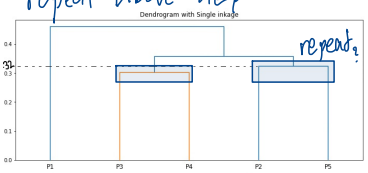
	P1	P2	P3	P4	P5
P1	0				
P2	1.04139	0			
P3	0.59304	0.77369	0		
P4	0.46098	0.61612	0.30232	0	
P5	0.81841	0.32388	0.45222	0.35847	0

③ merge & recompute distance

	P1	P2	P3,P4	P5
P1	0			
P2	1.04139	0		
P3,P4	0.46098	0.61612	0	
P5	0.81841	0.32388	0.35847	0

min(D(P3,P1), D(P4,P1))
min(D(P3,P2), D(P4,P2))
min(D(P3,P5), D(P4,P5))

④ repeat above step



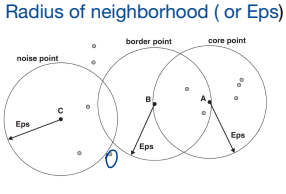
(-) Ward's Method (Centroid linkage)

Let be the SS error of every points from the centroid of cluster .
Determine all combinations of clusters to join such that is smallest.
Repeat until only one cluster is left.

For K clusters, the total SS is $\text{ESS}_K = \sum_{i=1}^{K} \text{ESS}_i$

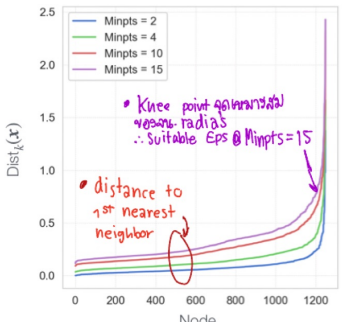
(3) Density-Based Clustering

- Density threshold of dense region (MinPts)
- Radius of neighborhood (or Eps)



Choice of MinPts & Eps

1. เลือก ค่า MinPts → K = MinPts
2. หาระยะ x ของ distance k^{th} เพื่อนบ้านใกล้สุด
   - Small values of $dist_k(x)$ จุดใน cluster & จุดผิดปกติมีค่าใหญ่
   - Knee point of $dist_k(x)$ บอกค่าเหมาะสม Eps สำหรับ MinPts = K



Factor Analysis.

merge feature? feature?

↳ advance PCA ; บอกได้ว่า 1st PC

Step 1 Std dataset

Step 2 factor loading(L) = eigenvector's PCA (n_component)

Mathematic Behind.

$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$ $\mathbf{x}_p$

	wheel-base	length	width	height	curb-weight	engine-size	compression-ratio	city-mpg	highway-mpg
0	88.6	168.8	64.1	48.8	2548	130	9.0	21	27
1	88.6	168.8	64.1	48.8	2548	130	9.0	21	27
2	94.5	171.2	65.5	52.4	2823	152	9.0	19	26
3	99.8	176.6	66.2	54.3	2337	109	10.0	24	30
4	99.4	176.6	66.4	54.3	2824	136	8.0	18	22
...	...	...	...	...	...	...	...	...	...
200	109.1	188.8	68.9	55.5	2952	141	9.5	23	28
201	109.1	188.8	68.8	55.5	3049	141	8.7	19	25
202	109.1	188.8	68.9	55.5	3012	173	8.8	18	23
203	109.1	188.8	68.9	55.5	3217	145	23.0	26	27
204	109.1	188.8	68.9	55.5	3062	141	9.5	19	25

$$\hat{\rho}_{12} \approx \frac{1}{n}\mathbf{x}_1 \cdot \mathbf{x}_2 = \frac{1}{205}(88.6 \cdot 168.8 + 88.6 \cdot 16 + 94.5 \cdot 171.2 + \cdots + 109.1 \cdot 188.8)$$

(1.) Factor Analysis : eigenvector ของ 1st PC

$$\frac{X_1-\mu_1}{\sigma_1} = X_1^{(s)} = \ell_{11}F_1 + \ell_{12}F_2 + \cdots + \ell_{1m}F_m + \varepsilon_1$$
$$\frac{X_2-\mu_2}{\sigma_2} = X_2^{(s)} = \ell_{21}F_1 + \ell_{22}F_2 + \cdots + \ell_{2m}F_m + \varepsilon_2$$
$$\vdots \qquad \vdots$$
$$\frac{X_p-\mu_p}{\sigma_p} = X_p^{(s)} = \ell_{p1}F_1 + \ell_{p2}F_2 + \cdots + \ell_{pm}F_m + \varepsilon_p$$

$$\underset{(p\times 1)}{\mathbf{X}} = \underset{(p\times m)}{\mathbf{L}} \cdot \underset{(m\times 1)}{\mathbf{F}} + \underset{(p\times 1)}{\boldsymbol{\epsilon}}$$

แล้วมาดู relationship ของแต่ละ feature จะได้รู้ว่าควร merge ไหน

(Correlation Matrix) $\mathrm{Corr}\{\mathbf{X}\} \triangleq \mathbf{R} = \mathbf{L}\mathbf{L}^T + \boldsymbol{\Psi}$

Factor loadings

$$\mathbf{L}\mathbf{L}^T + \boldsymbol{\Psi} = \begin{pmatrix} .969 & -.231 \\ .519 & .807 \\ .785 & -.587 \\ .971 & -.210 \\ .704 & .667 \end{pmatrix}\begin{pmatrix} .969 & .519 & .785 & .971 & .704 \\ -.231 & .807 & -.587 & -.210 & .667 \end{pmatrix} + \begin{pmatrix} .007 & 0 & 0 & 0 & 0 \\ 0 & .079 & 0 & 0 & 0 \\ 0 & 0 & .040 & 0 & 0 \\ 0 & 0 & 0 & .013 & 0 \\ 0 & 0 & 0 & 0 & .060 \end{pmatrix} = \begin{pmatrix} 1.000 & .317 & .896 & .990 & .528 \\ .317 & 1.000 & -.066 & .335 & .904 \\ .896 & -.066 & 1.000 & .885 & .161 \\ .990 & .335 & .885 & 1.000 & .543 \\ .528 & .904 & .161 & .543 & 1.000 \end{pmatrix}$$

Specific variances

L L^T

$X_1 = .969F_1 - .231F_2$
$X_2 = .519F_1 + .806F_2$
$X_3 = .784F_1 - .587F_2$
$X_4 = .971F_1 - .210F_2$
$X_5 = .704F_1 + .667F_2$

dataset = kind int happy lik just

```
[[1,5,5,1,1],
 [8,9,7,9,8],
 [9,8,9,9,8],
 [9,9,9,9,9],
 [1,9,1,1,9],
 [9,7,7,9,9],
 [9,7,9,9,7]]
```

Result Factor Analysis

Observed variables		Loadings F1	Loadings F2	Communalities	Specific var.
Kind	l_1	0.9695	-0.2311	0.9933	0.0067
Intelligent	l_2	0.5194	0.8069	0.9209	0.0791
Happy	l_3	0.7845	-0.5872	0.9603	0.0397
Likable	l_4	0.9709	-0.2099	0.9867	0.0133
Just	l_5	0.7040	0.6669	0.9404	0.0596

Variance $F_i = \sum_{i}^{n} l_i^2$ | Prop. of Var $= \frac{\sum l^2}{p}$

ผลรวมของค่าลักษณะเฉพาะทั้งหมดมีค่าเท่ากับ p (จำนวน Observed Variables) = p

Choosing the number of factor

1. เลือก factor = จน ค่าลักษณะเฉพาะ (eigenvalue) มีค่ามากกว่า 1
2. ดู scree Plot ของ Eigenvalue
3. สัดส่วนเฉพาะค่าลักษณะเฉพาะ (Cumulative var proportion)
   # อธิบายผล แปรปรวนให้มากสุด (อย่างน้อย 50%) โดยใช้จน-ปัจจัยไม่มากเกินไป จนไม่สามารถอธิบายความได้ | Factor Interpretation = ค่าปัจจัยใกล้กันเข้าด้วย

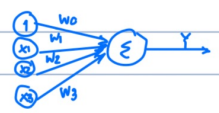
Factor Rotation มาช่วยให้ความให้ค่า loadings เด่นขึ้นมา

Observed variables	Loadings F1	Loadings F2	Communalities	Specific var.
Kind	0.9506	0.2994	0.9933	0.0067
Intelligent	0.0314	0.9591	0.9209	0.0791
Happy	0.9747	-0.1011	0.9603	0.0397
Likable	0.9409	0.3184	0.9867	0.0133
Just	0.2616	0.9338	0.9404	0.0596

maximise variance F loadings in each column

Group 1 : King, Happy, Likeable
Group 2 : Intelligent, Just
Group 1 : ด้านมิตรภาพ Amiability
G.2 = Logical ด้านใน

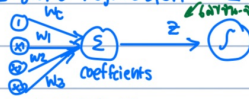
Lec 11 Neural Networks & Deep Learning



Neural networks basics

- Activation : ① Linear Regression Model $Y = W_0 + W_1X_1 + W_2X_2 + W_3X_3$

② Logistic Regression $Z = W_0 + W_1X_1 + W_2X_2 + W_3X_3$ $Y = \frac{1}{1+e^{-z}}$



③ Artificial Neuron = a single-layer neural network

$Z = W_0 + W_1X_1 + W_2X_2 + W_3X_3$ | $Y = f(Z) = h(X)$ | Sigmoid Logistic $f(z) = \frac{1}{1+e^{-z}}$

Sigmoind คือ transform ให้อยู่ช่วง output [0, 1]

- Loss/cost : MSE หาค่า weight ที่ดีที่สุด fit train data ที่สุดเพื่อ $h = f(\Sigma_j w_j x_j)$ ให้ได้ best-fit weight (x1, x2) $L(w_1, w_2) = \frac{1}{2N}\sum_i (w_1x_1^{(i)} + w_2x_2^{(i)} - y^{(i)})^2$ [ค่า Error]
- Loss Function landscape = Error ∴ หาค่า lost ต่ำสุด
- Gradient Desent Algorithm ปรับ weight เพื่อลงเข้าหา gradiance มากสุด ?

stop criteria $\|\nabla L(w_1, w_2)\| < \varepsilon$ → tolerance ↳ ค่าคลาดเคลื่อนของ loss ที่ยอมรับได้

Sum of squared elements
ค่า gradian น้อยมาก

- ปรับ weight : $w_{ji} \Leftarrow w_{ji} - \alpha \cdot \frac{\partial\ \text{Loss}}{\partial w_{ji}}$

3) แก้ Overfit ③.1 L1 & L2 Regularization
③.2 Dropout Layer (∵ ค่า weight บางตัวไม่เข้าเลย)

Feature Importance using Permutation สลับค่าใน col → predict → ดู Accuracy ?
แล้วทำซ้ำ แต่ละ feture → ถ้าทำแล้ว feature ไหน acc ลดเยอะสุด = สำคัญ

• Classification Loss/Cost : Cross-Entropy : ค่าผิดจากค่าทำนายต่างกัน

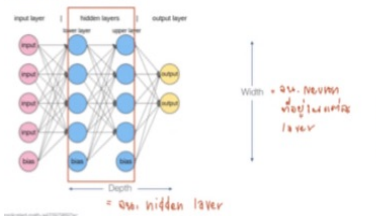
$$\text{Error}_i \text{ or } L_i(\boldsymbol{w}) = -(y_i \log_2(h(x_i;\boldsymbol{w})) + (1-y_i)\log_2(1-h(x_i;\boldsymbol{w})))$$

True value — Predict prob.

ค่าผิดจาก h	ค่าทำนาย y	Squared errror	Cross-entropy
0.8	1	0.04	0.10
0.1	1	0.81	1.00
0.1	0	0.01	0.05

ผลผิด/ทำนาย : ต่างกันน้อย → A: ยังมี error ให้ gradian ปรับได้อยู่

- Neural Network (NN) perceptron ซ่อนอยู่ มักเพิ่ม hidden layers ความสัมพันธ์ input↔output



Network 1: มี 2 Layer, มี 3 unit

Network 2: มี neuron = 7 unit, มี weight = 9 (เส้น), 3 layer

(QUIZ) ให้ลอง single-layer perceptron network มี 3 input with no bias unit และมี 1 sigmoid-activation output. มีค่าน้ำหนัก W1 = -1.5, W2 = 1, W3 = 1. For a dataset {(X1,X2,X3,y)} with 3 data points {(1,0,0,0.3), (1,1,0,0.5), (1,0,1,0.6)} what's the total squared errors of the network?

sol^n

$z = (X_1 \times W_1) + (X_2 \times W_2) + (X_3 \times W_3)$ weight sum data point

$\delta(z) = \frac{1}{1+\exp(-z)}$ sigmoid activation function

$(\delta(z) - y)^2$ Square error of each data point

Data Point 1 (1,0,0,0.3)

$z = (1 \times 1.5) + (0 \times 1) + (0 \times 1) = -1.5$

$\delta(z) = \frac{1}{1+e^{(-(-1.5))}} \approx 0.182$

Square error1 = $(0.182 - 0.3)^2 \approx 0.01386$

Data Point 2 (1,1,0,0.5)

$z = (1 \times 1.5) + (1 \times 1) + (0 \times 1) = -0.5$

$\delta(z) = \frac{1}{1+e^{(-(-0.5))}} \approx 0.377$

Square error2 = $(0.377 - 0.5)^2 \approx 0.0151$

Data Point 3 (1,0,1,0.6)

$z = (1 \times -1.5) + (0 \times 1) + (1 \times 1) = -0.5$

$\delta(z) = \frac{1}{1+e^{(-(-0.5))}} \approx 0.377$

Square error 3 = $(0.377 - 0.6)^2 \approx 0.04968$

∴ Total Squared Error = 0.01386 + 0.0151 + 0.04968 ≈ 0.07864 ≈ 0.08 # Ans

(QUIZ) ให้ลองหา Feed-forward neural network with one hidden layer with 4 nodes + bias unit. มี output layer 3 nodes และมี input layer 2 nodes + bias unit. How many weight parameters are there?

sol^n Input Layers : 2 nodes + 1 bias = 3 nodes (input → Hidden → output)

Output : 3 nodes

hidden Layers : 4 nodes + 1 bias = 5 nodes ถือเป็นมี bias ของ hidden.

[1] Input Layer to Hidden Layer = 3x4 = 12 weights ไม่มี bias

[2] Hidden Layer to Output Layer = 3x5 = 15 weights include bias

Total Weight Parameter = 12+15 = 27 total weights # Ans

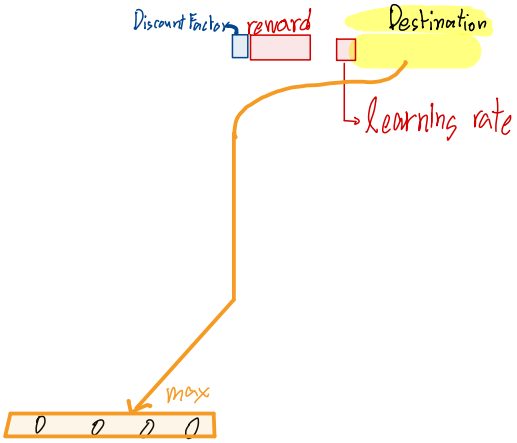
- Deep Neural Network(DNN) : Neural networks with 2 ชั้นขึ้นไป hiden layer
- "shallow neural network" Python → #1 sklearn, #2 Tensorflow, #3 keras
- Neural Network Regression with Keras ดูว่า overfit/underfit
  - → ถ้าค่า lost > training lost = overfit!

Recomendation Tuning

Hyper-parameter Tuning Recommendations
- 1 - 5 hidden layers (with same # hidden states/units in layer)
- Cone-like topology : (layer ลดลงเรื่อยๆ ถึง 0)
- Batch size ≤ 32
- Make it deeper tends to be more helpful than wider. (deep ดีกว่า wide)
- Learning rate -- Start large and reduce till converge

- ข้อควรระวัง Training Deep Neural Net
  - 1) Unstable (weight ตอนอัพ update)
  - 2) Speed-up training time
  - 3) ป้องกัน overfit (อาจลด hidden)

- Fixing Unstable Gradients 1) Unstable
  - 1.1) Weight initialization
  - 1.2) Non-Saturating Activation
  - 1.3) Batch Normalization : Scale ค่า x ของ output ออกมาแต่ละ layer ลดปัญหา Unstable gradient
- 2) หรือใช้ซ้ำกันได้ : Transference Learning



Quiz 1) PCA

(original) x_1 x_2

QUIZ 1. ให้ (x_1, x_2) มีค่าเป็น $\begin{bmatrix} 5 & 4 \\ 1 & -2 \\ -1 & 1 \\ -3 & 1 \end{bmatrix}$ ค่า Sample variance ของข้อมูล 2 ตัวแปรมีค่าเป็นเท่าใด

$\frac{1352.0}{12} =$

np.cov(X, rowvar=False)

$\bar{X}_1 = \frac{8}{4} = 2$, $\bar{X}_2 = \frac{4}{4} = 1$ → $\begin{bmatrix} 3 & 3 \\ -1 & -3 \\ -3 & 0 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 9 \\ 3 \\ 0 \\ 0 \end{bmatrix}$ ∴ $\text{var}(x_1, x_2) = \frac{9+3+0+0}{n-1} = \frac{12}{3} = 4$

2. ให้หาค่า Total Sample Variance ในชุดข้อมูล

$\begin{bmatrix} 6.67 & 4 \\ 4 & 6 \end{bmatrix}$

$\text{Var}(X_1) = \frac{(3)^2+(-1)^2+(-3)^2+(1)}{3} = 6.67$

$\text{Var}(X_2) = \frac{(3)^2+(-3)^2+(0)^2+(0)^2}{3} = 6$

total var = 6.67 + 6 = 12.67

Clustering

1. From a one-dimensional data set {2,4,10,12,3,20}, if we apply K-means clustering algorithm using K=2, and that we randomly pick the initial centroid as (mu1, mu2) = (1, 6). What is the centroid in the next iteration ?

centroid → cluster

$\mu_1 = 1$ → {2, 3}

$\mu_2 = 6$ → {4, 10, 12, 20}

update

new → $\mu_1 = \frac{2+3}{2} = 2.5$

→ $\mu_2 = \frac{4+10+12+20}{4} = 11.5$

2. ถ้าเรามี cluster ของข้อมูลเป็น {(5,4), (3,1), (1,1)} ค่า SSE ของ cluster มีค่าเท่าใด *

centroid $\mu_1 = \frac{5+3+1}{3} = 3$, $\mu_2 = \frac{4+1+1}{3} = 2$

$$SSE = \sqrt{\left(\sqrt{(5-3)^2+(4-2)^2}\right)^2 + \left(\sqrt{(3-3)^2+(1-2)^2}\right)^2 + \left(\sqrt{(1-3)^2+(1-2)^2}\right)^2}$$

$= 8 + 1 + 5 = 14$

Consider the likelihood function of a mixture of three bivariate Gaussian distributions below. How many model parameters to estimate in this mixture model ?

$$L(\Theta, w|X) = \prod_{i=1}^{n}\sum_{k=1}^{3} w_k f(x_i;\theta_k)$$

$$f(x;\theta_k) = \frac{\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)}{\sqrt{(2\pi)^2|\Sigma|}}, \quad \theta_k = (\mu_k, \Sigma_k)$$

→ Variable of each Gaussian dist.

- $\mu = \mu_1, \mu_2$
- $\text{cov} = \sigma_1, \sigma_2, \text{cov}(X_1, X_2)$
- total = 5 ตัว

→ K = 3 cluster

→ Weight = w_1, w_2, w_3 = 3 ตัว

numbers of param = 3(5) + 3 = 18 ตัว

From the current proximity matrix, which of the two clusters/points are to be merged next by using a single linkage?

Current proximity matrix

	1	23	4	5	6	7	8
1	0	1.414	4.472	5.657	6.708	5.099	7.280
23		0	3.162	4.243	5.000	3.162	5.385
4			0	2.000	4.123	4.243	5.385
5				0	2.236	3.162	3.606
6					0	2.236	1.414
7						0	2.236
8							0

Pairwise distance matrix

	1	2	3	4	5	6	7	8
1	0	1.414	2.000	4.472	5.657	6.708	5.099	7.280
2		0	1.414	3.162	4.243	5.385	4.000	6.083
3			0	4.000	4.472	5.000	3.162	5.385
4				0	2.000	4.123	4.243	5.385
5					0	2.236	3.162	3.606
6						0	2.236	1.414
7							0	2.236
8								0

single linkage = 1 กับ {2,3} และ 6,8

From each of the possible merges in the previous question, what is the distance of the closest point to the merged cluster ?

$\min(D(5,6), D(5,8)) = \min(2.23, 3.606) = 2.23$

$\min(D(7,6), D(7,8)) = \min(2.23, 2.23) = 2.23$

Distance 2.23

Factor Analysis

Observed variables	Loadings F1	Loadings F2	Communalities	Specific var.
Kind	0.9695	-0.2311	0.9933	0.0067
Intelligent	0.5194	0.8069	0.9209	0.0791
Happy	0.7845	-0.5872	0.9603	0.0397
Likable	0.9709	-0.2099	0.9867	0.0133
Just	0.7040	0.6669	0.9404	0.0596

What is the amount of variance in the data explained by factor F1 ? (Choose the nearest value)

$\text{var of } F_1 = (0.96)^2 + (0.52)^2 + (0.78)^2 + (0.97)^2 + (0.71)^2$

$= 3.261$

What is the proportion of variance in the data explained by factor F1 ? (Choose the nearest value)

$\text{prop} = \frac{\text{var } F_1}{p} = \frac{3.261}{5} = 65\%$

What is the proportion of variance in the data explained by factor F2? (Choose the nearest value)

$\text{prop} = \frac{(-0.23)^2 + (0.8)^2 + (-0.59)^2 + (-0.21)^2 + (0.67)^2}{5}$

$= 30\%$