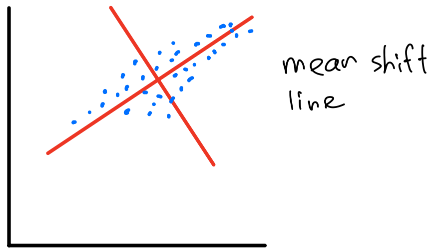
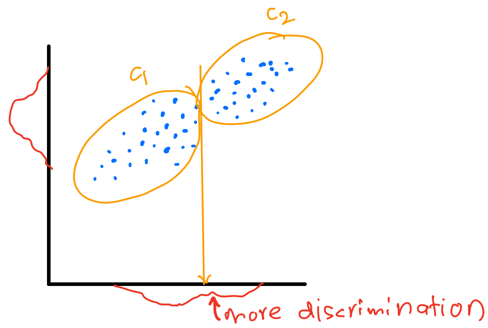
Linear Discriminant Analysis



PCA

(operates on data,
linear method,
unsupervised)

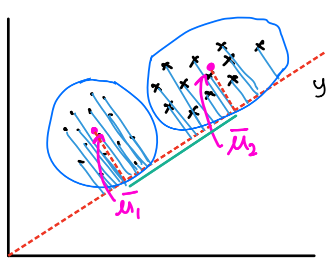
* give me top 3 principal components



LDA

(operates of feature subspace
linear method,
supervised)

- Data $\langle x_1, x_2 \ldots x_N \rangle$
- N_1 samples belong to class c_1
- N_2 samples belong to class c_2
- Find a line that maximize the class seperation.



$y = w^T x$ find weights

only w^T change to rotate the line

* maximize difference between averages

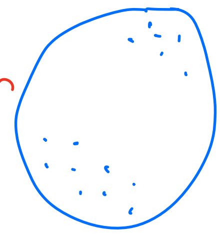
- Define a good seperation measure
- Measure $\mu_i = \dfrac{1}{N_i} \sum_{x \in c_i} x$

$$\tilde{\mu}_i = \frac{1}{N_i} \sum_{y \in c_i} y = \frac{1}{N_i} \sum_{x \in c_i} w^T x = w^T \mu_i$$

- Driving force for seperation

argmax over w objective $f(w) = |\tilde{\mu}_1 - \tilde{\mu}_2|$ L1 norm

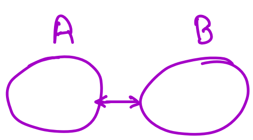
$= |w^T(\mu_1 - \mu_2)|$

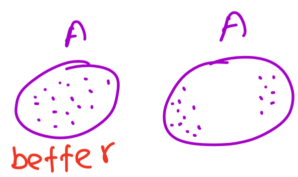


better class → variance inside class has to be minimum

Distance to other classes has to be maximum

But we are ignoring variability inside classes.





Fisher Approach

Normalize the distance(difference) between the means by intra-class scatter = variance

variance inside class $\tilde{s}_i^2 = \sum_{y \in c_i} (y - \tilde{\mu}_i)^2$

Intra class scatter $= \tilde{s}_1^2 + \tilde{s}_2^2$ (sum should be minimum)

Objective(w) = maximum inter class seperation

minimum intra class variability

Fisher Discriminant

$$\text{Objective}(w) = \frac{|\tilde{\mu}_1 - \tilde{\mu}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2} \leftarrow \text{L2 norm}$$

t - SNE (t- Distributed stochastic neighbor embedings)

* bring anything to 2D

→ non linear data visualizer

→ t-test | t-distribution (normal distribution)

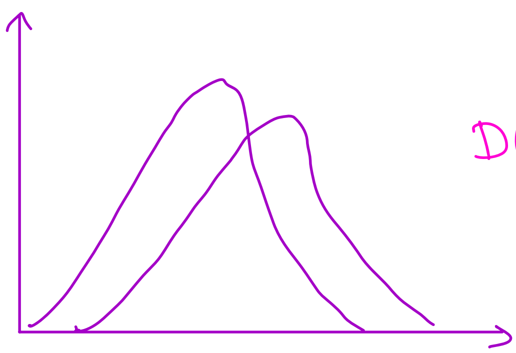
t-SNE doesnt use any norm (distance metric)

~~L1~~, ~~L2~~ .. ~~Lp~~

Kullback Leibler Divergence > distance metric

Given 2 probability distributions p, q the KL divergence measures the distance

$D(p\|q)$ How much does p distribution diverges from q



$$D(p\|q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$

universe of disourse (set of main events)

$$D(p\|q) \neq D(q\|p)$$

→ KL divergence is not a metric

ex:- Divergence D(observed || normal)

How far data is deviate from normal Distribution

$D(\text{observed} \,||\, \text{binomial})$

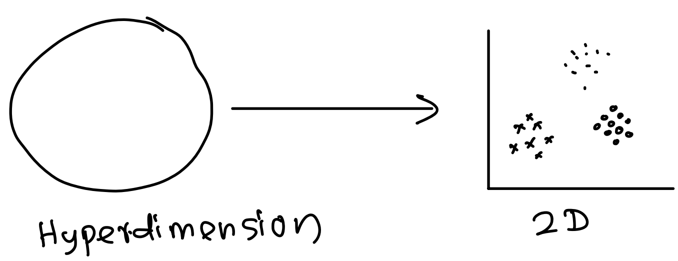
Relationship to entropy $H(X)$

$$H(X) = \sum_{x \in X} p(x) \log \frac{1}{p(x)}$$

$$= \log N - D(\underbrace{p(x)}_{\text{true distribution}} \,||\, \underbrace{P_U(x)}_{\text{Uniform distribution}})$$

The shannon entropy is the number of bits necessary to identify X from N equally likely possibilities less the KL divergence of the uniform distribution from the true distribution.

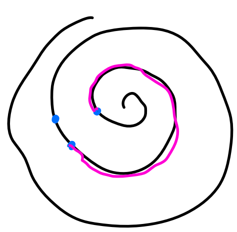
t-SNE idea



classifier ✗

visualizer ✓

Similarity in high dimensions corresponds to short distance in low dimensions.



perplexing !

t-SNE minimizes the sum of KL divergence over all data points using gradient decent method.

$$\text{Objective} = \sum_i D(P_i \| Q_i)$$

$$= \sum_i \sum_j P_{j|i} \log \frac{P_{j|i}}{Q_{j|i}}$$

$$\left. P_{j|i} = \frac{\exp\left(-\frac{\|x_i - x_j\|^2}{2\delta_i^2}\right)}{\sum_{k \neq i} \exp\left(-\frac{\|x_i - x_k\|^2}{2\delta_i^2}\right)} \right\} \text{high dimention}$$

$$\left. q_{j|i} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k \neq i} \exp\left(-\|y_i - y_k\|^2\right)} \quad \delta_i = \frac{1}{\sqrt{2}} \right\} \text{low dimension}$$

Dim Reduction and Visualization

* PCA
* LDA
* t-SNE