DEEP LEARNING

Wk 05

Structured Predictn ⟷ Hidden Markov Model (Supervised Learng)
→ Generative Model.

Naive Bayes ⟷ (vs) Logistic Regressn

Discriminative Model for Structured Predictn

Nxt wk - Unsupervised Regressn (PCA)

Advanced Topics

$$P(y|x) = \frac{1}{1+\exp(-y(\vec{\theta}\cdot x+\theta_0))}$$ Logistic Regressn

↓

+1 or -1 tagging

↓

Now what if multiple classes

↓

$$P(y|x) = \frac{\exp\left(\sum_{k=1}^{K} \lambda_k \cdot f_k(x,y)\right)}{\sum_{y'} \exp\left(\sum_{k=1}^{K} \lambda_k \cdot f_k(x,y')\right)}$$ Softmax Regressn

Entropy - Degree of uncertainty in data sent

↳ In war, sent 1 bit → event occur in war

2 bit → 4 possible event occurred

A	0	0
B	0	1
C	1	0
D	1	1

3 bit → 8 events

$\log_2 K$ bit → 2^K possible event → Expected no of bit : $\sum_x P(x) \cdot \log_2 \frac{1}{P(x)}$

↳ Degree of uncertainty high if $P_1, P_2, P_3, P_4 = \frac{1}{4}$

↳ Entropy

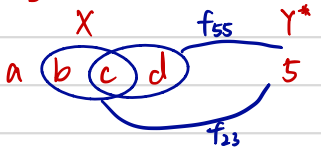
Constraint Optimizatn Problem

Given (X, Y^*)

$\max_{P(Y|X)} \sum_Y P(Y|X) \cdot \log P(Y|X)$ constraint by ① $\sum_Y P(Y|X) \cdot f_k(X,Y) = f_k(X,Y^*) \rightarrow \sum_Y P(Y|X) \cdot f_k(X,Y) - f_k(X,Y^*) = 0$

② $\sum_Y P(Y|X) = 1 \rightarrow \sum_Y P(Y|X) - 1 = 0$

Eg



Add Lagrangian multiplier

bad guys ↓ $(-\infty, \infty)$ ↗ can be +ve/-ve

$$\sum_Y P(Y|X) \cdot \log P(Y|X) - \sum_{k=1}^{K} \lambda_k \left[\sum_Y P(Y|X) \cdot f_k(x,y) - f_k(x,y^*)\right] - \lambda_0 \left[\sum_Y P(Y|X) - 1\right]$$

bad guys $(-\infty, \infty)$ can be +ve/-ve

$$\sum_Y P(Y|X)\cdot \log P(Y|X) - \sum_{k=1}^{k} \lambda_k \left[\sum_Y P(Y|X)\cdot f_k(x,y) - f_k(x,y^*) \right] - \lambda_0 \left[\sum_Y P(Y|X) - 1 \right]$$

$$\frac{\partial L}{\partial P(Y|X)} = \log P(Y|X) + 1 - \sum_{k=1}^{k} \lambda_k [f_k(x,y)] - \lambda_0 = 0 \quad - (1)$$

$$\log P(Y|X) = -1 + \sum_{k=1}^{k} \lambda_k [f_k(x,y)] + \lambda_0$$

$$P(Y|X) = \exp\left(-1 + \sum_{k=1}^{k} \lambda_k [f_k(x,y)] + \lambda_0\right)$$

$$= \exp\left(\sum_{k=1}^{k} \lambda_k f_k(x,y)\right) \cdot \exp(\lambda_0 - 1)$$

$$= \frac{\exp\left(\sum_{k=1}^{k} \lambda_k f_k(x,y)\right)}{Z}, \quad Z = \frac{1}{\exp(\lambda_0 - 1)}$$

$$= \frac{\exp\left(\sum_{k=1}^{k} \lambda_k f_k(x,y)\right)}{\sum_{y'} \exp\left(\sum_{k=1}^{k} \lambda_k \cdot f_k(x,y')\right)}$$

$P(y_1|X) = \frac{1}{Z}$
$P(y_2|X) = \frac{3}{Z}$
$P(y_3|X) = \frac{5}{Z}$

bad guys $(-\infty, \infty)$ can be +ve/-ve

$$\max_\lambda \sum_Y P(Y|X)\cdot \log P(Y|X) - \sum_{k=1}^{k} \lambda_k \left[\sum_Y P(Y|X)\cdot f_k(x,y) - f_k(x,y^*) \right] - \lambda_0 \left[\sum_Y P(Y|X) - 1 \right]$$

$$= \max_\lambda \sum_Y P(Y|X) \underbrace{\left[\log P(Y|X) - \sum_{k=1}^{k} \lambda_k - \lambda_0\right]}_{-1 \text{ from eqn (1)}} + \sum_{k=1}^{k} \lambda_k \cdot f_k(x,y^*) + \lambda_0$$

$$= \max_\lambda \underbrace{-\sum_Y P(y|x)}_{-1} + \sum_{k=1}^{k} \lambda_k f_k(x,y^*) + \lambda_0$$

$$= \max_\lambda \log P(y^*|X) \rightarrow \text{Learning Problem.}$$

Add a log infront.

For (X, y^*)

$$L = \log P(y^*|X) = \sum_{k=1}^{k} \lambda_k f_k(x,y^*) - \log \underbrace{\sum_Y}_{\text{general } Y.} \underbrace{\exp\left(\sum_{k=1}^{k} \lambda_k f_k(x,y)\right)}_{Z}$$

X, Y Next Pg

$$\frac{\partial L}{\partial \lambda_j} = f_j(X,Y^*) - \frac{1}{Z} \sum_Y \left[\exp\left(\sum_{k=1}^{k} \lambda_k f_k(x,y)\right) \cdot f_j(x,y)\right]$$

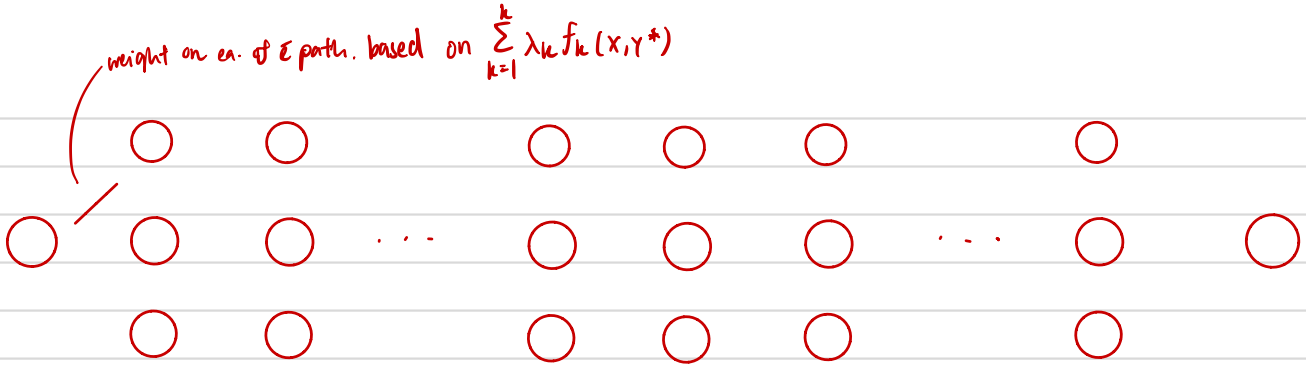
$$= f_j(X,Y^*) - \sum_Y \frac{\exp\left(\sum_{k=1}^{k} \lambda_k f_k(x,y)\right)}{\sum_{y'} \exp\left(\sum_{k=1}^{k} \lambda_k f_k(x,y')\right)} \cdot f_j(x,y)$$

$$= f_j(x,y^*) - \sum_Y P(y|x) \cdot f(x,Y)$$

$f_j(x,y^*)$: Actual time you observe

$\sum_Y P(y|x) \cdot f(x,Y)$: Expected time you observe y. → calculated using forward-backward algo
↳ due to lack of closed form soln

Eg. Fortn is a fine inventn
N V D A N
f_{12}: feature 12. f_{23}



For $\{-\log \sum_{Y} \exp[\sum_{k=1}^{k} \lambda_k f_k(x,y)]\}$

Compare to $\log \sum_{a \in S} \exp(a) = 21.000\ldots$ $\max_{a \in S}(a) = 21$ where $S = \{2, 5, 21\}$

Hence $l = \log P(y^*|x) = \sum_{k=1}^{k} \lambda_k f_k(x, y^*) - \max_{y} (\sum_{k=1}^{k} \lambda_k f_k(x, y))$ $+\Delta$

$= 0$ then training set is linearly separable (Perceptron Loss) and $y = +1$ & -1

score of true path is $\bar{e}$ (1)x $+\Delta$ if not 0 and obtain hinge loss

score of best path = structured SVM.

Introduction of Latent Variable.

Faith - Noun - Proper Noun? or Improper noun. (ie Latent variable) learn from data-set.

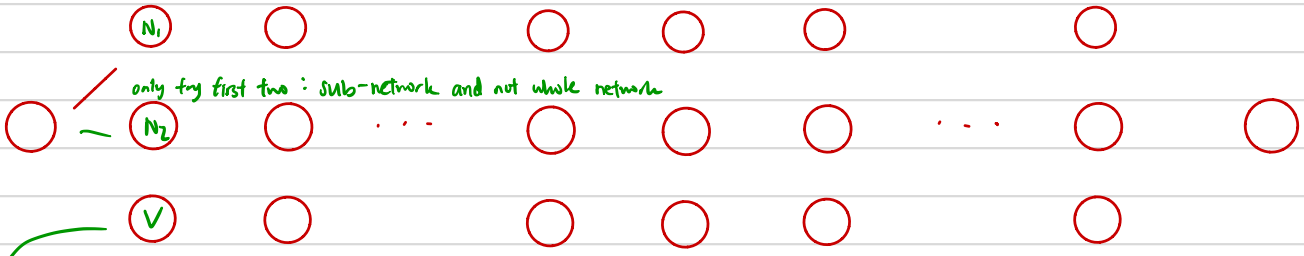
$$\sum_{h} p(y^*, h|x) = \sum_{h} \frac{\exp(\sum_{k=1}^{k} \lambda_k f_k(x, h, y^*))}{\sum_{h', y'} \exp(\sum_{k=1}^{k} \lambda_k f_k(x, h', y'))}$$

$$l = \log \sum_{h} \exp(\sum_{k=1}^{k} \lambda_k f_k(x, h, y^*)) - \log \sum_{h', y'} \exp(\sum_{k=1}^{k} \lambda_k f_k(x, h', y'))$$

$$\frac{\partial l}{\partial \lambda_j} = \sum_{h} p(h|x, y^*) \cdot f_j(x, y^*, h) - \sum_{h', y} p(h', y|x) \cdot f_j(x, y, h')$$

↓ Given X, Y^* ↳ Obtain Feature Count.

↓ Given X and with else ↳ Obtain Feature Count



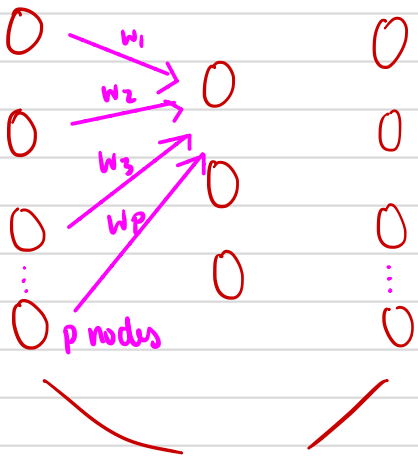
λ : tuned for optimal path in sub-network. and max eqn

↳ Tendency to go verb is low.

(X, Y) can be diff from (X, y^*)

W2P5

Autoencoder



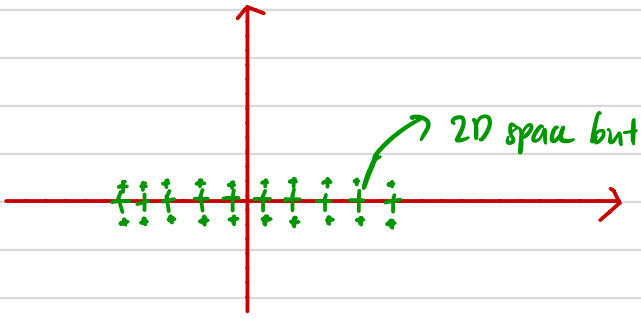
Unsupervised Regression
(Left out from Normal Class)

output layer similar to input layer

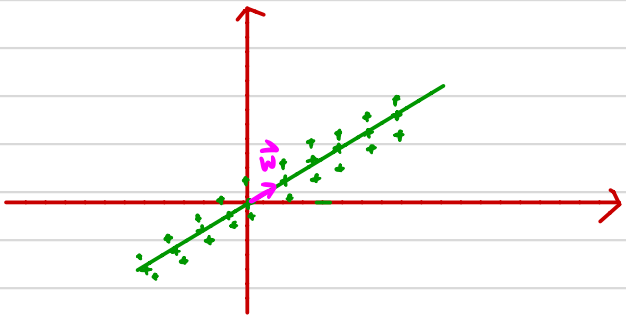
movies

users					
	2		5		
	1			4	
		2	1		5
	1	2		5	2
	5	2	5		1
	3		1		3

← Ratg for movie



2D space but simplified to 1D space due to nature of space



Approach 1

$\max_{\vec{w}} \frac{1}{n} \sum_{i=1}^{n} (\vec{x}_i \cdot \vec{w})^2$

└ variance max projectn for $x_1, \ldots, x_n$ on $\vec{w}$

$\|\vec{w}\|^2 = 1$

Approach 2

$\min_{\vec{w}} \frac{1}{n} \sum_{i=1}^{n} \| \vec{x}_i - (x_i \cdot \vec{w}) \cdot \vec{w} \|^2$

$= \frac{1}{n} \sum_{i=1}^{n} \|\vec{x}_i\|^2 - 2 \vec{x}_i (\vec{x}_i \cdot \vec{w}) \vec{w} + \|(\vec{x}_i \cdot \vec{w}) \vec{w}\|^2$

Approach 2

$$\min_{\vec{w}} \frac{1}{n} \sum_{i=1}^{n} \| \vec{x}_i - (x_i \cdot \vec{w}) \cdot \vec{w} \|^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} \|\vec{x}^i\|^2 - 2 \vec{x}_i (\vec{x}_i \cdot \vec{w}) \vec{w} + \|(\vec{x}_i \cdot \vec{w}) \vec{w}\|^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} \|\vec{x}^i\|^2 - 2(\vec{x}_i \cdot \vec{w})^2 + (\vec{x}_i \cdot \vec{w})^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} \|\vec{x}^i\|^2 - (\vec{x}_i \cdot \vec{w})^2 + (\vec{x}_i \cdot \vec{w})^2$$

Approach 1

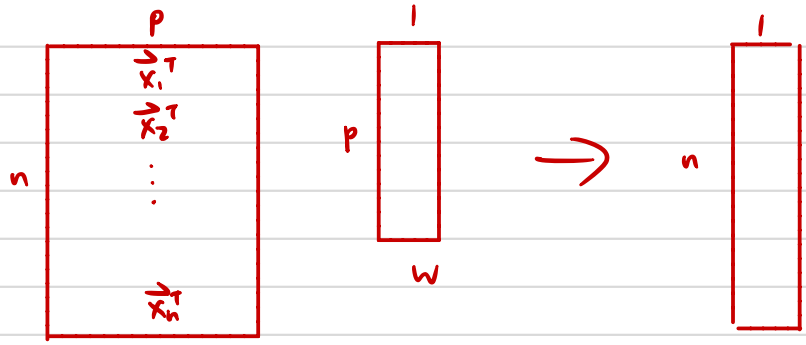
$$\max_{\vec{w}} \frac{1}{n} \sum_{i=1}^{n} (\vec{x}_i \cdot \vec{w})^2$$

$$= \frac{1}{n} (Xw)^T Xw$$

$$= \frac{1}{n} w^T X^T Xw$$

$$= w^T \left(\frac{X^T X}{n} \right) w$$ — tve semi definite

$$= w^T A w = w^T \lambda w = \lambda w^T w = \lambda$$



↳ Langrangian Multiplier

$$\mathcal{L} = w^T A w - \lambda (w^T w - 1), \quad w^T w = 1$$

$$\frac{d\mathcal{L}}{dw} = 2Aw - 2\lambda w = 0 \rightarrow A\vec{w} = \lambda \vec{w}$$

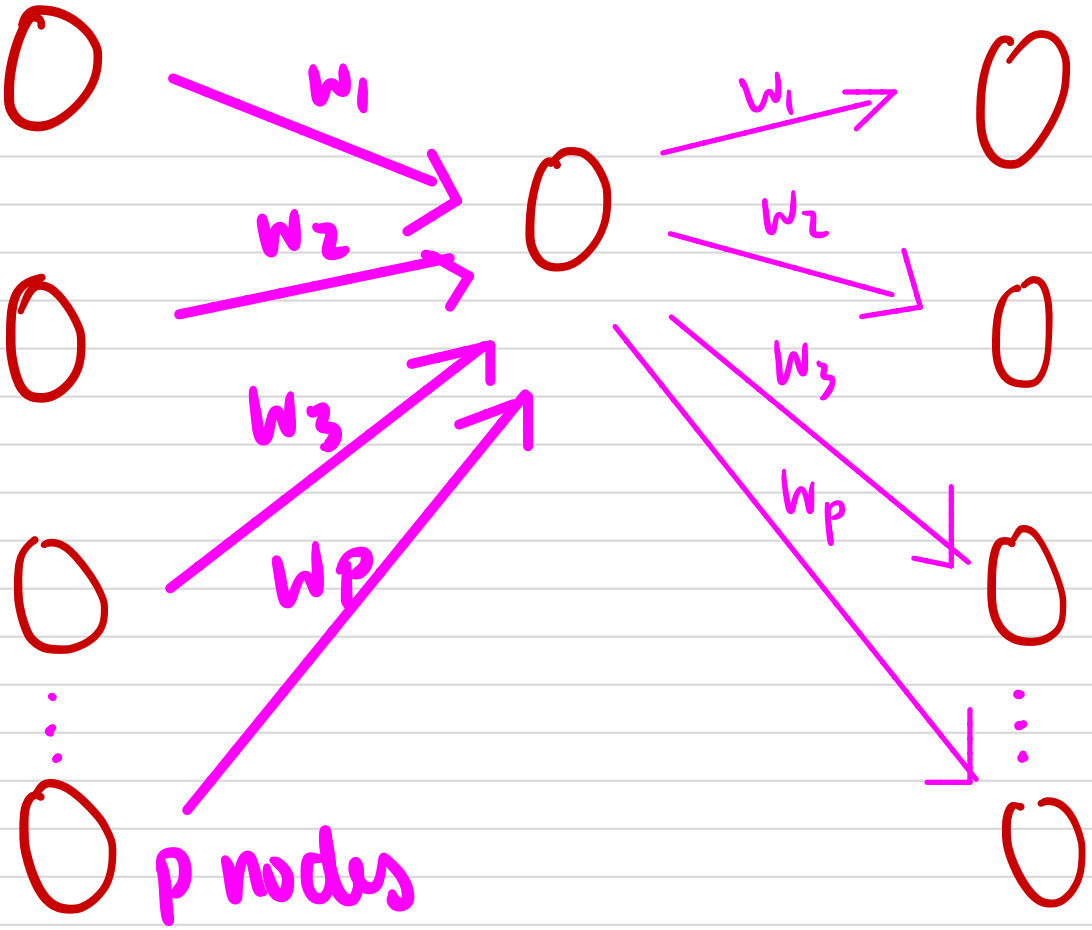
↳ Eigenvalue (no imaginary value)
↳ Eigenvector

Soln : Use SVD (Single Value Decomposition) & $V^T V = V V^T = I$

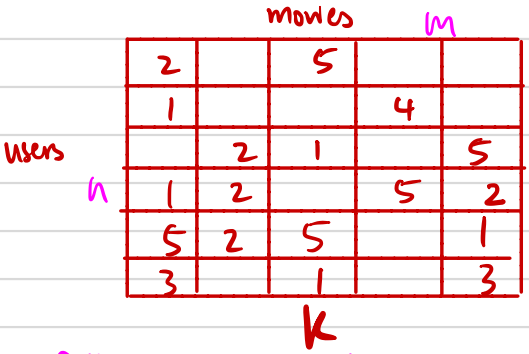
$$X = U \Sigma V^T$$

$$A = \frac{1}{n} V \Sigma U^T U \Sigma V^T$$

$$= V \frac{\Sigma^2}{n} V^T$$

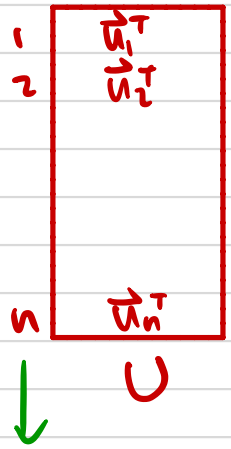


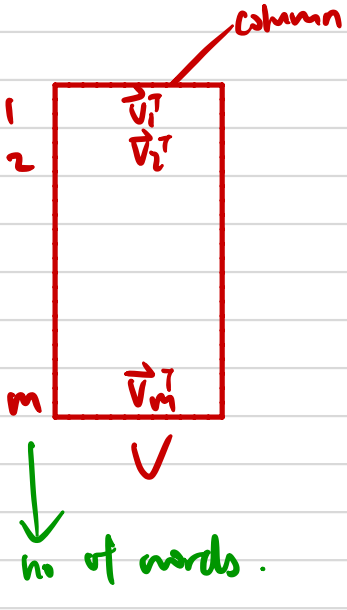
Principal Component Analysis.



Collaborative Filtering.

Aim: $\vec{u}_3 \cdot \vec{v}_2 \simeq 2$





Regularized Term

$$\arg\min_{U,V} \sum_{(i,j)\in D} (\vec{u}_i \cdot \vec{v}_j - Y_{ij})^2 + \lambda \sum (\vec{u}_i)^2 + \lambda \sum_{j=1}^{m} (\vec{v}_j)^2$$

↓

Find similar scores & input into empty movies rating scores.

↳ upgrades Linear → Ridge Regression.

values it don't veer too off

Latent Semantic Analysis