

## $\Rightarrow$ Assignment ①

- ① Non-negative matrix Factorization is an alternative to PCA when data and factors can be cast as non-negative. We seek to factorize  $N \times P$  data matrix  $X$  as

$$X \approx WH$$

where  $W$  is  $N \times r$  and  $H$  is  $r \times P$ , with  $r \leq \min(N, P)$ ; and we assume that  $x_{ij}, w_{ik}, h_{kj} \geq 0$ .

- ② Suppose that  $x_{ij} \in \mathbb{N}$ . If we model each random variable  $x_{ij}$  as a (independent) Poisson random variable with mean  $(WH)_{ij}$ . Show that the log-likelihood of the model is (up to a constant)

$$\mathcal{L}(W, H) = \sum_{i,j} [x_{ij} \log(WH)_{ij} - (WH)_{ij}]$$

The following alternating algorithm converges to a local maximum of  $\mathcal{L}(W, H)$

$$w_{ik} \leftarrow w_{ik} \frac{\sum_j h_{kj} x_{ij}}{(\text{WH})_{ij}}$$

$$h_{kj} \leftarrow \frac{\sum_i w_{ik} x_{ij}}{\sum_i w_{ik}}$$

$$h_{kj} \leftarrow \frac{h_{kj} \sum_i w_{ik} x_{ij}}{\sum_i w_{ik}}$$

A function  $g(x, y)$  is said to minorize a function  $f(x, y)$  if  $\exists \gamma$  such that

$$f(x, y) \geq g(x, y) \leq f(x) + g(x, y) - f(y)$$

$\Rightarrow$  Solution

The likelihood of observing  $x_{ij}$  given the mean  $(\text{WH})_{ij}$  for a poisson distribution is given by

$$P(x_{ij} | (\text{WH})_{ij}) = \frac{e^{-(\text{WH})_{ij}} (\text{WH})_{ij}^{x_{ij}}}{x_{ij}!}$$

$\rightarrow$  log likelihood

$$\log P(x_{ij} | (WH)_{ij}) = \log \left( \frac{e^{-(WH)_{ij}} (WH)_{ij}^{x_{ij}}}{x_{ij}!} \right)$$

Expanding this gives.

$$\log P(x_{ij} | (WH)_{ij}) = -(WH)_{ij} +$$

$$x_{ij} \log((WH)_{ij}) - \log(x_{ij}!)$$

The term  $\log(x_{ij}!)$  is independent of W and H, so its a constant and can be ignored. Since the  $x_{ij}$  are

IID (Independent Identically Distributed), the log likelihood for all observations is just sum of all

observations.

$$L(W, H) = \sum_{ij} [-(WH)_{ij} + x_{ij} \log((WH)_{ij})]$$

Thus concluded.

(B) Show that under the update

$$x^{t+1} = \arg \max_{x_j} g(x_j; x^t)$$

The sequence  $f_t = f(x^t)$  is non-decreasing

→ Solution:

given that,

$$x^{t+1} = \arg \max_{x_j} g(x_j; x^t)$$

This implies

$$g(x^{t+1}; x^t) \geq g(x_j; x^t) \text{ for all } x_j$$

We assume that  $g$  is somehow related to the gradient or direction of steepest ascent for  $f$ , then maximizing  $g$  would lead to an increase in  $f$ .

→ Given the update rule and the relationship between  $g$  and  $f$ , we can infer that:

$$f^{t+1} = f(x^{t+1}) \geq f(x^t) = f_t$$

This shows  $f_t$  is non-decreasing.

(C) Using concavity of the logarithm, show that for any set of r values  $y_k > 0$  and  $0 \leq c_k \leq 1$  with  $\sum_{k \leq r} c_k = 1$

$$\log\left(\sum_{k \leq r} y_k\right) \geq \sum_{k \leq r} c_k \log\left(y_k/c_k\right)$$

(justified with diagram showing that  $c_k$  is a weight assigned to  $y_k$ )

→ The concavity of the logarithm function allows us to use Jensen's inequality, which states that for a concave function  $f$  and a convex combination of points.

$$f\left(\sum_i \lambda_i x_i\right) \geq \sum_i \lambda_i f(x_i)$$

where  $\lambda_i$  are non-negative weights such that  $\sum_i \lambda_i = 1$ .

→ Applying Jensen's Inequality.

$$\log \left( \sum_{k \in r} y_r \right) \geq \sum_{k \in r} c_k \log (y_k)$$

→ Introduce Multiplicative Identity.

Using the fact that  $\sum_{k \in r} c_k = 1$ , we

can multiply each term  $y_k$  by  $\frac{c_k}{c_k}$

$$\begin{aligned} \log \left( \sum_{k \in r} y_k \right) &= \log \left( \sum_{k \in r} y_k \cdot \frac{c_k}{c_k} \right) \\ &\geq \log \left( \sum_{k \in r} y_k \cdot c_k \right) \end{aligned}$$

Using Jensen's inequality to modify

$$\log \left( \sum_{k \in r} \frac{y_k}{c_k} \cdot c_k \right) \geq \sum_{k \in r} c_k \log \left( \frac{y_k}{c_k} \right)$$

$$\Rightarrow \log \left( \sum_{k \in r} y_k \right) \geq \sum_{k \in r} c_k \log \left( \frac{y_k}{c_k} \right)$$

(Q) Deduce that  $\forall i \in \{1, N\}, j \in \{1, P\}$

$$\log \left( \sum_{k \leq r} w_{ik} h_{kj} \right) \geq \sum_{k \leq r} c_{kij} \log \left( \frac{w_{ik} h_{kj}}{c_{kij}} \right)$$

where  $c_{kij} = \frac{w_{ik} h_{kj}}{\sum_{l \leq r} w_{il} h_{lj}}$

$$\sum_{k \leq r} w_{ik} h_{kj}$$

and  $t$  is the current iteration.

$$\log \left( \sum_{k \leq r} y_k \right) \geq \sum_{k \leq r} c_k \log \left( \frac{y_k}{c_k} \right)$$

also substituting with (ii)

$$y_k \rightarrow w_{ik} h_{kj} \quad \text{and} \quad c_k \rightarrow c_{kij}$$

$$c_k \rightarrow \frac{w_{ik} h_{kj}}{\sum_{l \leq r} w_{il} h_{lj}}$$

$$\log \left( \sum_{k \leq r} w_{ik} h_{kj} \right) \geq \sum_{k \leq r} c_{kij} \log \left( \frac{w_{ik} h_{kj}}{c_{kij}} \right)$$

(e) Ignoring constants, show that

$$g(w, h; w^t, h^t) = \sum_{ijk} [$$

$$x_{ij} c_{kij} (\log w_{ik} + \log h_{kj} - \log c_{kij}) \\ - w_{ik} h_{kj}] \text{ minorizes } L(w, h)$$

→ Solution

To show that the function  $g$  minorizes  $L$ , we need to ensure that:

①  $g(w, h; w^t, h^t) \leq L(w, h)$  for all  $w, h$ .

②  $g(w, h; w, h) = L(w, h)$

→ Apply Jensen's Inequality.

rearranging

$$x_{ij} c_{kij} (\log w_{ik} + \log h_{kj} - \log c_{kij}) \\ - w_{ik} h_{kj}$$

$$x_{ij} c_{kj} \log\left(\frac{w_{ik} h_{kj}}{c_{kj}}\right) - w_{ik} h_{kj}$$

$$\text{Now, } c_{kj} = \frac{w_{ik} h_{kj}}{\sum_{k \in r} w_{ik}^t h_{kj}^t}$$

$$\therefore \text{we get } g(w, h; w^t, h^t) = \sum_{i, j, k}$$

$$g(w, h; w^t, h^t) = \sum_{i, j, k}$$

$$x_{ij} \frac{w_{ik} h_{kj}^t}{\sum_{k \in r} w_{ik}^t h_{kj}^t} \log\left(\frac{w_{ik} h_{kj}}{\sum_{k \in r} w_{ik}^t h_{kj}^t}\right) - w_{ik} h_{kj}$$

Simplifying:

$$g(w, h; w^t, h^t) = \sum_{i,j,k} \left[ x_{ij} \frac{w_{ik}^t h_{kj}^t}{\sum_{k \leq r} w_{ik}^t h_{kj}^t} \right]$$

$$\log \left( \frac{w_{ik} h_{kj} \sum_{k \leq r} w_{ik}^t h_{kj}^t}{w_{ik}^t h_{kj}^t} \right) - w_{ik} h_{kj}$$

$\rightarrow$  compare it with  $L(w, h)$

$$L(w, h) = \sum_{i,j} \left[ x_{ij} \log \left( \sum_{k \leq r} w_{ik} h_{kj} \right) - \sum_{k \leq r} w_{ik} h_{kj} \right]$$

$$f(\sum_i d_i x_i) \geq \sum_i d_i f(x_i)$$

from Jensen's Inequality

$$\log \left( \sum_{k \leq r} w_{ik} h_{kj} \right) \geq$$

$$\frac{\sum_{k \leq r} w_{ik} h_{kj}^t}{\sum_{k \leq r} w_{ik}^t h_{kj}^t} \log \left( \frac{w_{ik} h_{kj} \sum_{k \leq r} w_{ik}^t h_{kj}^t}{w_{ik}^t h_{kj}^t} \right)$$

From this we conclude that each term of  $g$  is  $\leq$  to  $l$  thus proving minorization.

(F) Finally derive update steps (3,4) by setting to zero the partial derivatives of  $g$ .

$$g(\omega, H; \omega^t, H^t) = \sum_{i,j,k} [$$

$$x_{ij} c_{kij} (\log w_{ik} + \log h_{kj} - \log l_{kj}) -$$

$$- \log(w_{ik} h_{kj})]$$

Partial derivation wrt  $w_{ik}$

$$\frac{\partial g}{\partial w_{ik}} = \sum_j \left[ x_{ij} c_{kij} \frac{1}{w_{ik}} - h_{kj} \right] = 0$$

$$\frac{\partial g}{\partial h_{kj}} = \sum_i \left[ x_{ij} c_{kij} \frac{1}{h_{kj}} - w_{ik} \right] = 0$$

Now we have to find  $w_{ik}$  and  $h_{kj}$

$$\sum_j [x_{ij} c_{kij} \frac{1}{w_{ik}}] = \sum_j h_{kj}$$

and

$$\sum_j [x_{ij} c_{kij} \frac{1}{h_{kj}}] = \sum_j w_{ik}$$

$$\therefore w_{ik} = \frac{\sum_j x_{ij} c_{kij}}{\sum_j h_{kj}}$$

$$h_{kj} = \frac{\sum_i x_{ij} c_{kij}}{\sum_i w_{ik}}$$

Hence derived.

(Q3)

### Maximum Entropy distributions.

(a) Let  $X = \{x_1, \dots, x_N\}$  be a discrete set and  $\mathcal{P} \in \mathcal{P}(X)$  be the space of probability distributions defined over  $X$ . Define the entropy

$$H(\mathbf{p}) := - \sum_{i=1}^N p_i \log(p_i)$$

By identifying  $\mathbf{p} \in \mathcal{P}(X)$  with a point  $\tilde{\mathbf{p}} = (p(x_1), \dots, p(x_N))$  in the  $N$ -dimensional simplex

$$\Delta_N := \{ \mathbf{y} \in \mathbb{R}^N : y_i \geq 0, \sum_i y_i = 1 \}$$

Show that  $H$  is a concave function in  $\Delta_N$ .

$\Rightarrow$  Solution

$$H(\mathbf{p}) = - \sum_{i=1}^N p_i \log(p_i)$$

Taking a second derivative of  $H$ ,

$$\frac{\partial^2 H(\mathbf{p})}{\partial p_i \partial p_j} \Rightarrow \begin{cases} -\frac{1}{p_i} & \text{if } i=j \\ 0 & \text{if } i \neq j \end{cases}$$

So, the Hessian matrix  $H''(P)$  is diagonal element -  $\frac{1}{p_i}$  which are all non-positive since  $p_i > 0$ . Therefore  $H$  is concave.

Q.E.D.

- (b) Show that  $H$  is non-negative in the simplex, and that its maximum is attained at the uniform distribution ( $\frac{1}{N}, \dots, \frac{1}{N}$ ) with maximum entropy  $\log(N)$ .

$$H(P) = -\sum_{i=1}^N p_i \log(p_i)$$

each term  $p_i \log(p_i)$  is non-positive because  $\log(p_i)$  is negative for  $0 < p_i < 1$ , and  $p_i \log(p_i) = 0$  for  $p_i = 0$ . Thus,  $H(p)$  is non-negative.

To find the maximum of  $H$ , we take the derivative and set it to zero:

$$\frac{\partial H(P)}{\partial p_i} = -\log(p_i) - 1 = 0 \Rightarrow p_i = \frac{1}{e}$$

Since  $\sum_{i=1}^N p_i = 1$ , we have

$H(p)$  with  $N+1$  states  $\rightarrow p_i = \frac{1}{N+1}$   
 $\therefore H(p) = \log(N+1)$

Substituting back into  $H(p)$  gives,

$$H_{\max} = - \sum_{i=1}^N \frac{1}{N} \log\left(\frac{1}{N}\right) = \log(N)$$

Now let us move from a discrete to a continuous domain by setting  $X = [0, R]$ . Let  $p(x)$  now denote the space of probability distributions admitting a density  $p(x)$ ,  $x \in X$ , and define the Shannon entropy as

$$H(p) = - \int_X p(x) \log p(x) dx$$

Show that the maximum entropy distribution in  $p(x)$  is the uniform measure in  $X$ , with entropy  $\log R$ .

→ We need to maximize  $H(p)$  subject to two constraints

① The probability density function  $p(x)$  must be normalized  $\int_0^R p(x) dx = 1$

②  $p(x)$  should be non-negative for all  $x \in [0, R]$ .

→ Apply Lagrange multipliers.

$$L(p, \lambda) = - \int_0^R p(x) \log(p(x)) dx + \lambda \left( 1 - \int_0^R p(x) dx \right)$$

$$\frac{\partial L}{\partial p(x)} = -\log(p(x)) - 1 - \lambda = 0.$$

Step 3:- Solve for the optimal  $p(x)$

$$\log(p(x)) = -1 - \lambda \Rightarrow p(x) = e^{-1-\lambda}$$

Applying the normalization constraint:

$$\int_0^R e^{-1-\lambda} dx = 1 \Rightarrow e^{-1-\lambda} = \frac{1}{R}$$

So maximum entropy distribution is

$$p(x) = \frac{1}{R}$$

$$H_{\max} = \pi \int_0^R \frac{1}{R} \log\left(\frac{1}{R}\right) dx \\ = \log(R).$$

We showed that MFD on the interval  $[0, R]$  is the Uniform distribution with density  $1/R$ , and its Entropy is  $\log R$ .

(d) We now attempt to understand

maximum entropy distributions defined over  $X \in \mathbb{R}$ . For  $p \in P(\mathbb{R})$ , denote the spread  $\Delta_p := \sup_{p(x) > 0} x - \inf_{p(x) > 0} x$  as the smallest interval containing the support of  $p$  (where we abuse notation and identify a probability distribution in  $P(\mathbb{R})$  with given mean  $c$  and Spread  $\Delta \leq \infty$  is the uniform distribution over the interval  $(c - \Delta/2, c + \Delta/2)$ , with entropy  $\log \Delta$ )

→ We are given that the spread  $\Delta P := \sup p(x) - \inf p(x)$  is finite, and we know the mean  $C$ . We want to maximize the entropy.

$$H(P) = - \int_{-\infty}^{\infty} p(x) \log(p(x)) dx.$$

→ To maximize  $H(P)$

① The PDF  $p(x)$  must be normalized

$$\int_{-\infty}^{\infty} p(x) dx = 1$$

$$② \text{The mean } C: \int_{-\infty}^{\infty} x \cdot p(x) dx = C.$$

→ Lagrangian is now:

$$L(P, \lambda_1, \lambda_2) = \int_{-\infty}^{\infty} p(x) \log(p(x)) dx$$

$$+ \lambda_1 \left( 1 - \int_{-\infty}^{\infty} p(x) dx \right) +$$

$$\lambda_2 \left( C - \int_{-\infty}^{\infty} x \cdot p(x) dx \right)$$

From question 1, we have

$$H(p) = -\sum_{i=1}^n p_i \log(p_i) \geq -\lambda_1 - \lambda_2$$

Since  $\Delta$  is finite, the support of  $p(\omega)$  must be finite. With given mean capture, the optimal distribution

is unique. Since the support of  $p(\omega)$  is finite, the support of  $p(\omega)$  must be finite. With given mean capture, the optimal distribution

is unique. Since the support of  $p(\omega)$  is finite, the support of  $p(\omega)$  must be finite. With given mean capture, the optimal distribution

$$p(\omega) = \begin{cases} \frac{1}{\Delta}, & \text{for } c - \Delta/2 \leq \omega \leq c + \Delta/2 \\ 0, & \text{otherwise} \end{cases}$$

Substituting  $p(\omega)$  into the formula for  $H$ ,

$$H = \max_{\Delta} \min_{c} \sum_{\omega=c-\Delta/2}^{c+\Delta/2} \frac{1}{\Delta} \log\left(\frac{1}{\Delta}\right)$$

Since  $\Delta$  is finite, the support of  $p(\omega)$  is not

- (e) Conclude that the maximum entropy distribution over  $P(R)$  with given mean does not exist. How does the answer change if now we consider distribution over  $P(R)$ ?

Since the  $\Delta$  is finite, the support of the probability is bounded namely  $\Delta$  & there is no probability distribution

with maximum entropy defined over entire  $R$ .

However, if we consider  $P(R^+)$ , the support can be unbounded to the right allowing for a distribution with a given mean ~~not~~ to exist.

A common example is the exponential distribution, which has maximum entropy among all distributions defined over  $R^+$  with given mean.

### ⇒ Conclusion

for a given mean  $C$  and finite spread  $\Delta$ , the uniform distribution over the interval  $(C - \Delta/2, C + \Delta/2)$  has maximum equal to  $\log(N)$

No maximum entropy distribution exists over the entire real line  $R$

with a given mean.

For Distribution defined over  $\mathbb{R}^+$  with a given mean, maximum entropy distributions do exist such as the exponential distribution.