DS-GA-1005 Inference and Representation
Assignment

Tao Li
taoli@nyu.edu

September 29, 2020

1. By the definition, we have

$$\log p_{W,H}((x_{ij})) = \log(\prod_{ij} \frac{(WH)_{ij}^{x_{ij}} e^{-(WH)_{ij}}}{(x_{ij})!}) = \sum_{ij}[x_{ij}\log(WH)_{ij} - (WH)_{ij}] - \sum_{ij}\log(x_{ij})!,$$

hence, we show that the log-likelihood is equal to (up to a constant)

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

We now develop the algorithm.

(a) We note that $f_t \leq f_{t+1}$ follows from the following inequalities

$$f(x^t) = g(x^t, x^t) \leq g(x^{t+1}, x^t) \leq f(x^{t+1}).$$

(b) Since logarithm is concave, we have

$$\log(\sum_{k\leq r} y_k) = \log(\sum_{k\leq r} c_k \frac{y_k}{c_k}) \geq \sum_{k\leq r} c_k \log(y_k/c_k).$$

(c) We notice that c_{kij} is non-negative and sum to 1, hence with the inequality shown above, we have

$$\log(\sum_{k\leq r} w_{ik}h_{kj}) = \log(\sum_{k\leq r} c_{kij}\frac{w_{ik}h_{kj}}{c_{kij}}) \geq \sum_{k\leq r} c_{kij}\log(w_{ik}h_{kj}/c_{kij}).$$

(d) We notice that

$$g(W,H;W^t,H^t) = \sum_{ij} x_{ij}(\sum_k (c_{kij}\log w_{ik}h_{kj}) - (WH)_{ij}),$$

hence to show that $g(W,H;W^t,H^t) \leq \mathcal{L}(W,H)$, it suffices to show that $\sum_k c_{kij}\log w_{ik}h_{kj} \leq \log(WH)_{ij}$. By the inequalities shown above, we have

$$\begin{aligned}\sum_k c_{kij}\log w_{ik}h_{kj} &= \sum_k c_{kij}\log(c_{kij}w_{ik}h_{kj}/c_{kij}) \\ &= \sum_k c_{kij}\log(w_{ik}h_{kj}/c_{kij}) + \sum_k c_{kij}\log c_{kij} \\ &\leq \log(\sum_k w_{ik}h_{kj}),\end{aligned}$$

where the last inequality follows from the fact that $0 \leq c_{kij} \leq 1$.

(e) Apparently, $g(W, H; W^t, H^t)$ is concave respect to W, H. Hence, the maximizer is given by the stationary point.

$$\frac{\partial g(W, H; W^t, H^t)}{\partial w_{ik}} = \sum_j x_{ij} c_{kij} w_{ik}^{-1} - \sum_j h_{kj} = 0$$

We solve the equation for w_{ik}, we then obtain

$$w_{ik}^{-1} \sum_j x_{ij} \frac{w_{ik}^t h_{kj}^t}{(W^t H^t)_{ij}} = \sum_j h_{kj},$$

where the superscript t highlights quantities related to the $t-$ iteration. Finally, the equality lead to the following update rule

$$w_{ik}^{t+1} = w_{ik}^t \frac{\sum_j x_{ij} h_{kj}^t / (W^t H^t)_{ij}}{\sum_j h_{kj}}$$

2. See the ipython file. To make the code more clear, we provide the following supplement. For the update

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

the numerator can be rewritten as

$$\sum_j h_{kj} \left(\frac{X.}{WH} \right)_{ij} = \left[\left(\frac{X.}{WH} \right) H^\mathsf{T} \right]_{ik},$$

where ./ denotes the pointwise division. As for the denominator $\sum_j h_{kj}$, we also view it as the $ik-$ component of the matrix $\mathbf{1}_N \mathbf{1}_p^\mathsf{T} H^\mathsf{T}$. Finally, we have the update rule as

$$W \leftarrow W. \times \frac{\left(\frac{X.}{WH} \right) H^\mathsf{T}.}{\mathbf{1}_N \mathbf{1}_p^\mathsf{T} H^\mathsf{T}},$$

where we use ./ and .× denote the pointwise operation. Similar rule is also applied to H update.

We also note that NMF only guarantees local optimum and different results may show up every time the kernel restarts, as we choose random initialization.

3. (a) Since $X = VY + \mu + \epsilon$ and $Y \sim \mathcal{N}(0, I_J), \epsilon \sim \mathcal{N}(0, \text{diag}(\beta))$, the covariance of X is $\Sigma = VV^\mathsf{T} + \text{diag}(\beta)$ and its mean is μ. Hence the likelihood of the data X is given by

$$\mathcal{L}(\mu, \Sigma | X) = \prod_{x \in X} \frac{1}{\sqrt{2\pi}^d |\Sigma|^{\frac{1}{2}}} e^{-\frac{(x-\mu)^\mathsf{T} \Sigma^{-1} (x-\mu)}{2}}.$$

(b) Though both serves as a linear method to break the curse of dimensionality, they are actually fundamentally different. One simple way to explain this is to focus on the prior. For PCA, we do not assume any model for the observations, we start from the sample variances and try to identify those influential components. By PCA, we reduce the correlated observed variables to smaller set of independent composite variables, produced by projection. On the other hand, factor analysis is built upon a prior model with predefined latent variables, which in fact reveals the causal relationship between latent variables and the observed data.

In fact, the case given in (c) is just a good example where PCA coincides with factor analysis, i.e., where noise components ϵ_i are assumed of equal variance $\mathrm{Var}(\epsilon) = \beta I$. However, if this is not the case, e.g., $\mathrm{Var}(\epsilon) = \mathrm{diag}(\beta_1, \cdots, \beta_d)$, then the two are different.

(c) We notice that the log-likelihood function is given by

$$l(\mu, \Sigma \mid X) = -\frac{N}{2} \log |\Sigma| - \frac{1}{2} \sum_{i=1}^{N} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu),$$

where N denotes the number of data points, i.e., $|X|$. Taking the derivative with respect to μ is straightforward

$$\frac{\partial l}{\partial \mu} = \sum_{i=1}^{N} (x_i - \mu)^T \Sigma^{-1},$$

which implies the MLE for μ is the sample mean

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i.$$

MLE of Σ is the sample variance $\hat{v}$ and we reformulate the problem as

$$\inf_{V \in \mathbb{R}^{d \times J}, \beta \in \mathbb{R}} \|\hat{v} - VV^{\mathsf{T}} - \beta I\|_2$$

We note that this is simply a low-rank approximation problem, where the residual is further approximated by the identity matrix. If we let $\Lambda \in \mathbb{R}^{J \times J}$ be a diagonal matrix whose diagonal entries are J sorted eigenvalues (descending order), then we have the following procedure for seeking V and β.

- step 1: Compute the matrix P whose columns are the J leading eigenvectors of $\hat{v}$
- step 2: Update V by $V = P\Lambda^{1/2}$
- step 3: Update β by $\beta = \frac{1}{d} \operatorname{trace}(\hat{v} - VV^{\mathsf{T}})$

(d) See the ipython file.

4. (a) For $p, q \in \Delta_N$ and $\lambda \in [0,1]$, we have

$$H(\lambda p + (1-\lambda) q) = -\sum_{i=1}^{N} (\lambda p_i + (1-\lambda) q_i) \log(\lambda p_i + (1-\lambda) q_i).$$

On the other hand, $f(x) = -x \log x$ is strictly concave on $[0,1]$, since its second order derivative $f''(x) = -\frac{1}{x} < 0$ on $[0,1]$. Hence, we have

$$-(\lambda p_i + (1-\lambda) q_i) \log(\lambda p_i + (1-\lambda) q_i) \geq -\lambda p_i \log p_i - (1-\lambda) q_i \log q_i,$$

which leads to

$$\begin{aligned} H(\lambda p + (1-\lambda) q) &\geq -\lambda \sum_{i=1}^{N} p_i \log p_i - (1-\lambda) \sum_{i=1}^{N} q_i \log q_i \\ &\geq \lambda H(p) + (1-\lambda) H(q) \end{aligned}$$

(b) We notice that for $p \in \Delta_N$, $p_i \in [0,1]$, which means $\log(p_i)$ is non-positive. Hence, $H(p) = -\sum_{i=1} p_i \log(p_i)$ is non-negative. As for proving the maximum entropy is achieved at the uniform distribution, we shall use contradiction, which we find extremely helpful for explaining why uniform distribution is the maximizer. We shall adopt the optimization viewpoint in (c), as introduced in class.

Apparently, as a strictly concave function defined on a convex set, there exists a unique maximizer. We assume that the maximizer is p, which is not a uniform distribution, namely, there exists i, j such that $p_j > p_i \geq 0$, where $p_i = p(x_i)$. Then we consider the following function $p(t) = p + t(e_i - e_j), t \in [\max(-p_i, p_j - 1), \min(p_j, 1 - p_i)]$, where e_i denotes the indicator corresponding to x_i. Straightforward computation gives

$$\frac{d}{dt} H(p(t)) = \nabla H \cdot \nabla P(t) = \log \frac{p_j(t)}{p_i(t)},$$

which implies $\frac{d}{dt} H(p(t))|_{t=0} > 0$, showing that p is not optimal. In fact, by choosing a larger t, i.e., decreasing p_j and increasing p_i, we can achieve a greater entropy. This contradicts with our assumption, hence the unique maximizer is the uniform distribution.

(c) We consider the following optimization problem

$$\begin{array}{ll} \text{minimize} & \int_{\mathcal{X}} p(x) \log p(x) dx \\ \text{subject to} & \int_{\mathcal{X}} p(x) dx = 1 \end{array}$$

The lagrangian is given by $L(\lambda, p) = \int_{\mathcal{X}} p(x) \log p(x) dx - \lambda(\int_{\mathcal{X}} p(x) dx - 1)$, and the first order variation gives

$$\log p(x) + 1 - \lambda = 0,$$

which implies that the density is a constant. Hence, the maximizer $p(x)$ is the uniform measure in $\mathcal{X}$.

(d) We denote I the smallest interval that contains the support of p, and by the argument in (c), we have that $p(x)$ is a constant over I, i.e., $p(x) = \frac{1}{|I|}$ for $x \in I$. On the other hand, with the mean c, we have

$$\frac{1}{|I|} \int_I x dx = c,$$

which implies $\frac{b^2 - a^2}{2|I|} = c$, where $a < b$ are the endpoints of the interval, and $|I| = b - a = \delta$. We obtain two equalities and solve for a, b. We then obtain $a = c - \frac{\delta}{2}, b = c + \frac{\delta}{2}$. Also, the entropy by definition (an integral of the constant $\frac{\log \delta}{\delta}$) is $H(p) = \log \delta$.

(e) A straightforward way to explain that such a distribution does not exist is that we view the support of the distribution is infinite, i.e, $\delta = \infty$, which leads to $p(x) = \frac{1}{\delta} = 0$ for all x. Apparently, it is impossible, hence such distribution does not exist.

As for $\mathcal{P}(\mathbb{R}_+)$ case, we consider the Lagrangian

$$\mathcal{L} = \int_0^\infty p(x) \log p(x) dx - \lambda_0 \left(\int_0^\infty p(x) dx - 1 \right) - \lambda \left(\int_0^\infty x p(x) dx - c \right),$$

and the fist order variation gives $1 + \log p(x) - \lambda_0 - \lambda_1 x = 0$, which implies $p(x) = e^{\lambda_0 - 1 + \lambda_1 x}$. For the normalization constraint, we have

$$\int_0^\infty p(x) = \int_0^\infty e^{\lambda_0 - 1 + \lambda_1 x} dx = 1,$$

which leads to $\lambda_1 = -e^{\lambda_0 - 1}$. For the mean constraint, we have

$$\int_0^\infty xp(x)dx = \int_0^\infty xe^{\lambda_0 - 1}e^{\lambda_1 x}dx = c,$$

and by substituting $\lambda_1 = e^{-\lambda_0 + 1}$ into above, we have

$$\int_0^\infty x * (-\lambda_1)e^{\lambda_1 x}dx = c,$$

which implies that $p(x)$ is the pdf of the exponential distribution, i.e.,

$$p(x) = \begin{cases} \frac{1}{c}e^{-\frac{x}{c}} & x \geq 0 \\ 0 & x < 0 \end{cases}$$