Inference and Representation, Fall 2023

Problem Set 1: PCA & Maximum Entropy

Solution

Total: 100 points

1. *Non-negative Matrix Factorization (NMF).* [2, 3] is an alternative to PCA when data and factors can be cast as non-negative. We seek to factorize the $N \times p$ data matrix $\mathbf{X}$ as

$$\mathbf{X} \approx \mathbf{W}\ \mathbf{H}\ , \tag{1}$$

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

(a) 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)

$$\mathcal{L}(\mathbf{W}, \mathbf{H}) = \sum_{i,j} [x_{ij} \log((WH)_{ij}) - (WH)_{ij}]\ . \tag{2}$$

Solution. (5 pts)
The log-likelihood of a poisson distribution is

$$\log L\left(x_{ij} | (\mathbf{WH})_{ij}\right) = \log e^{-(\mathbf{WH})_{ij}} \frac{[(\mathbf{WH})_{ij}]^{x_{ij}}}{x_{ij}!} = x_{ij} \log((\mathbf{WH})_{ij}) - (\mathbf{WH})_{ij} + c.$$

So the log-likelihood of the model is (up to a constant)

$$\sum_{i,j} \log L\left(x_{ij} | (\mathbf{WH})_{ij}\right) = \sum_{i,j} [x_{ij} \log((\mathbf{WH})_{ij}) - (\mathbf{WH})_{ij}].$$

The following alternating algorithm (Lee, Seung, '01) converges to a local maximum of $\mathcal{L}(\mathbf{W}, \mathbf{H})$:

$$\begin{aligned} w_{ik} &\leftarrow w_{ik} \frac{\sum_j h_{kj} x_{ij} / (WH)_{ij}}{\sum_j h_{kj}}\ , & (3) \\ h_{kj} &\leftarrow h_{kj} \frac{\sum_i w_{ik} x_{ij} / (WH)_{ij}}{\sum_i w_{ik}}\ ,. & (4) \end{aligned}$$

We shall study this algorithm and prove its correctness.

A function $g(x, y)$ is said to minorize a function $f(x)$ if

$$\forall\ (x, y)\ ,\ g(x, y) \leq f(x)\ ,\ g(x, x) = f(x)\ .$$

(b) Show that under the update

$$x^{t+1} = \arg\max_x g(x, x^t)$$

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

Solution. (5 pts)
$f_{t+1} = g(x^{t+1}, x^{t+1}) \geq g(x^{t+1}, x^t) \geq g(x^t, x^t) = f_t$ where the first inequality is from the definition of minorizing function and the second inequality is due to the update rule.

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

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

Solution. (5 pts)
For a concave function function f, Jensen's inequality gives

$$f\left(\sum_i \alpha_i x_i\right) \geq \sum_i \alpha_i f(x_i), \quad \forall \alpha_i \in [0,1], \ \sum_i \alpha_i = 1.$$

Substituting $\alpha_i = c_i, x_i = y_i/c_i$ gives the answer generally. Also note that $\lim_{x \to 0} x \log x = 0$ completes cases of some $c_i = 0$.

(d) Deduce that $\forall i \in \{1, N\}$, $j \in \{1, p\}$,

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

where $c_{kij} = \frac{w^t_{ik} h^t_{kj}}{\sum_{k' \leq r} w^t_{ik'} h^t_{k'j}}$ and t is the current iteration.

Solution. (5 pts)
We have $\sum_k c_{kij} = 1$ and $c_{kij} \in [0,1]$ where both upper and lower bounds are from the non-negativity of $\mathbf{W}, \mathbf{H}$. So substituting $y_k = w_{ik} h_{kj}, c_k = c_{kij}$ gives the answer.

(e) Ignoring constants, show that

$$g(\mathbf{W}, \mathbf{H}; \mathbf{W}^t, \mathbf{H}^t) = \sum_{i,j,k} [x_{ij} c_{kij} (\log w_{ik} + \log h_{kj} - \log c_{kij}) - w_{ik} h_{kj}]$$

minorizes $\mathcal{L}(\mathbf{W}, \mathbf{H})$.

Solution. (10 pts)
First check $g(\mathbf{W}, \mathbf{H}; \mathbf{W}, \mathbf{H}) = \mathcal{L}(\mathbf{W}, \mathbf{H})$. It is

$$\begin{aligned} g(\mathbf{W}, \mathbf{H}; \mathbf{W}, \mathbf{H}) &= \sum_{i,j,k} \left[x_{ij} \frac{w_{ik} h_{kj}}{\sum_{k' \leq r} w_{ik'} h_{k'j}} \log \sum_{k' \leq r} w_{ik'} h_{k'j} - w_{ik} h_{kj} \right] \\ &= \sum_{i,j} \left[x_{ij} \log \sum_{k' \leq r} w_{ik'} h_{k'j} - \sum_{k} w_{ik} h_{kj} \right] \\ &= \sum_{i,j} [x_{ij} \log((\mathbf{WH})_{ij}) - (\mathbf{WH})_{ij}] = \mathcal{L}(\mathbf{W}, \mathbf{H}). \end{aligned}$$

Then we are to show, $\forall\ \mathbf{W}, \mathbf{H}, \mathbf{W}^t, \mathbf{H}^t$, it holds $g(\mathbf{W}, \mathbf{H}; \mathbf{W}^t, \mathbf{H}^t) \leq \mathcal{L}(\mathbf{W}, \mathbf{H})$. It follows

$$\mathcal{L}(\mathbf{W}, \mathbf{H}) \geq \sum_{i,j} [x_{ij} \sum_{k \leq r} c_{kij} \log(w_{ik} h_{kj} / c_{kij}) - \sum_{k} w_{ik} h_{kj}] = g(\mathbf{W}, \mathbf{H}; \mathbf{W}^t, \mathbf{H}^t),$$

where the inequality is from (d).
Therefore, $g(\mathbf{W}, \mathbf{H}; \mathbf{W}^t, \mathbf{H}^t)$ minorizes $\mathcal{L}(\mathbf{W}, \mathbf{H})$.

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

Solution. (10 pts)
Taking the partial derivatives of g w.r.t. w_{ik} as 0, we have

$$\frac{\partial}{\partial w_{ik}} g = \sum_{j} [x_{ij} c_{kij} \frac{1}{w_{ik}} - h_{kj}] = 0,$$

which gives

$$w_{ik} = \frac{\sum_j x_{ij} c_{kij}}{\sum_j h_{kj}} = \frac{\sum_j x_{ij} w_{ik}^t h_{kj}^t / (\mathbf{WH})_{ij}}{\sum_j h_{kj}} = w_{ik}^t \frac{\sum_j x_{ij} h_{kj}^t / (\mathbf{WH})_{ij}}{\sum_j h_{kj}}.$$

Similarly, taking the partial derivatives of g w.r.t. h_{kj} as 0, we have

$$\frac{\partial}{\partial h_{kj}} g = \sum_{i} [x_{ij} c_{kij} \frac{1}{h_{kj}} - w_{ik}] = 0,$$

which gives

$$h_{kj} = \frac{\sum_i x_{ij} c_{kij}}{\sum_i w_{ik}} = \frac{\sum_i x_{ij} w_{ik}^t h_{kj}^t / (\mathbf{WH})_{ij}}{\sum_i w_{ik}} = h_{kj}^t \frac{\sum_i x_{ij} w_{ik}^t / (\mathbf{WH})_{ij}}{\sum_i w_{ik}}.$$

2. *NMF vs PCA on images.* The following questions refer to the code and data `hw1.zip`. The MNIST dataset (in `mnist_all.mat`) contains images of handwritten digits labeled with

their associated numeric value. The file called `nmf.ipynb` has code performing the following tasks: (a) plot the singular vectors corresponding to the top 10 singular values of the data, and (b) project the training data and the test data (obtained using `load_test_data` in `mnist_tools.py`) onto the first $k = 8$ principal components and run nearest neighbors for each test image in this lower dimensional space. (The PCA code for this problem was adapted from Homework 1 in [1] .)

Solution. (10 pts)
The submitted jupyter notebook generates results with no explicit errors.

(a) Now apply the NMF algorithm with $r \in \{3, 6, 10\}$ and plot the rows of H (using `plot_image_grid` in `plot_tools.py`)

Solution. (10 pts)
The sample figures are as follows ($r = 3, 6, 10$). The results can be different but shall be in the same style (white major background with black pixels in the mid).

(b) Project the training data and the test data (obtained using `load_test_data` in `mnist_tools.py`) onto the r rows, and run nearest neighbors for each test image in this lower dimensional space. Include your choice for r, and the plots of your nearest neighbor results in your submitted homework document.

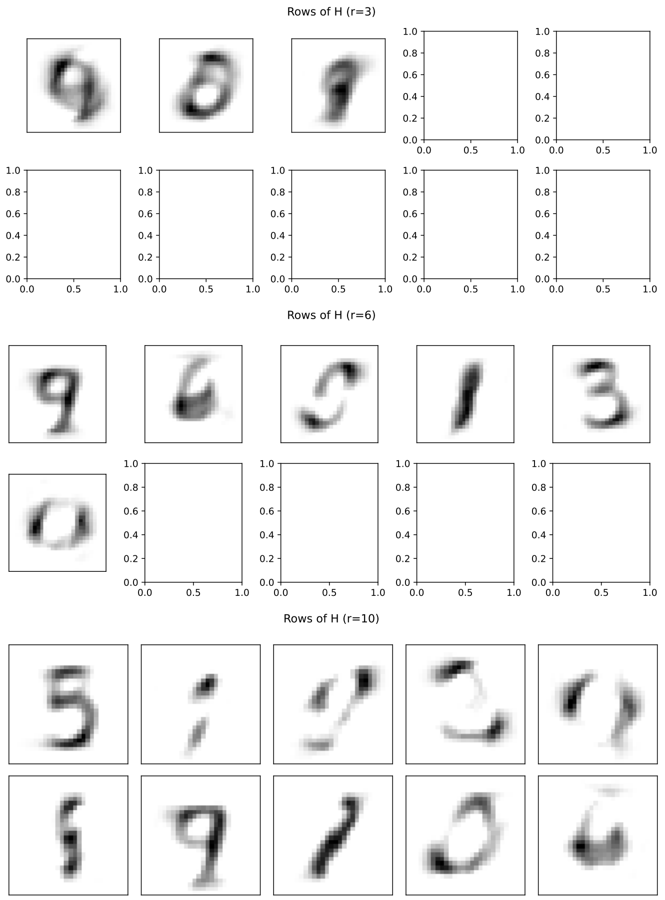
Solution. (5 pts)
The sample figure is as follows (for any pick of $r \in \{3, 6, 10\}$). The result can be different but shall be in the same style (five pairs of images). The digits can be incorrect.

(c) Comment on the differences between the PCA and NMF (To understand this better, plot the coefficients of a random image for each digit in terms of the first 10 principal components, on one plot and in terms of NMF for $r = 10$ on another plot).

Solution. (5 pts)
The sample figure is as follows (left column is NMF, and the right is PCA). Note that the plot should be 10 coefficients of 2 methods for 10 images, so there should be 200 data points in total. Any reasonable comment on the difference is good.

Include a .zip file with your online homework submission containing all of your source code.

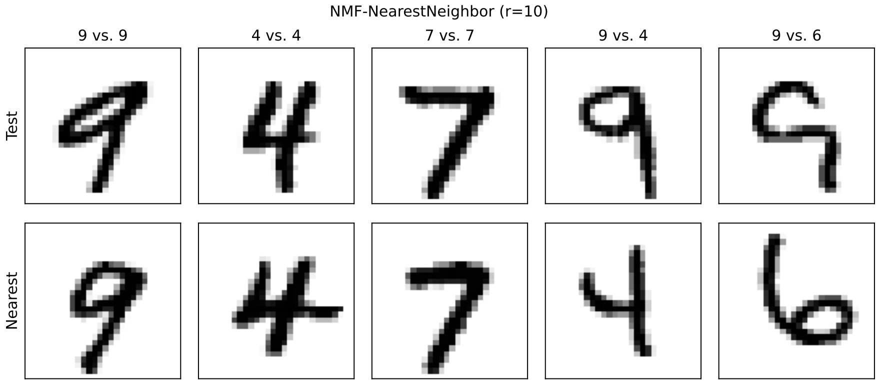
Keep in mind that the data points in the training and test data are given as rows.



Rows of H (r=3)

Rows of H (r=6)

Rows of H (r=10)



NMF-NearestNeighbor (r=10)

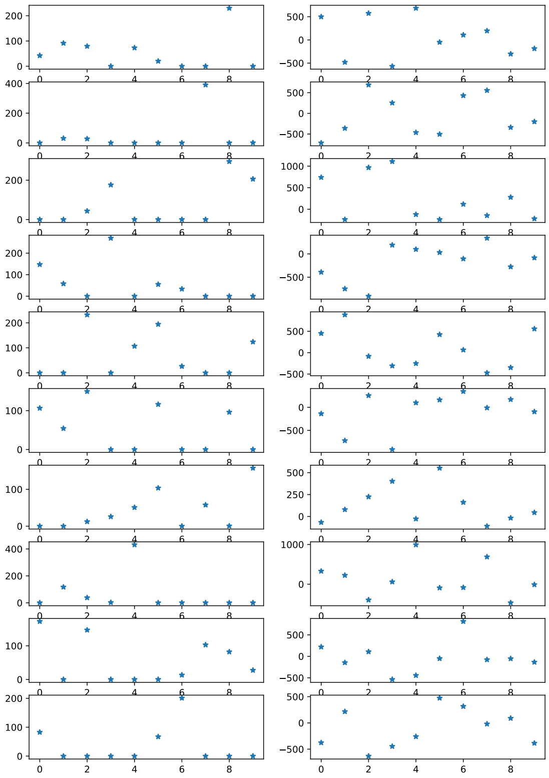
9 vs. 9

4 vs. 4

7 vs. 7

9 vs. 4

9 vs. 6


200
100
0
500
0
−500
400
200
0
500
0
−500
200
0
1000
500
0
200
100
0
0
−500
200
100
0
500
0
−500
100
0
0
−500
100
0
500
250
0
400
200
0
1000
0
100
0
500
0
−500
200
100
0
500
0
−500
0
2
4
6
8

3. *Maximum Entropy Distributions.* In this exercise, we will study Maximum Entropy Distributions.

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

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

   By identifying $p \in \mathcal{P}(\mathcal{X})$ with a point $\tilde{p} = (p(x_1), \dots, p(x_N))$ in the N-dimensional simplex $\Delta_N := \{y \in \mathbb{R}^N; y_i \geq 0, \sum_i y_i = 1\}$, show that H is a concave function in Δ_N.

Solution. (5 pts)
$\forall\ p_a, p_b \in \Delta_N, \forall\ \alpha_1, \alpha_2 \in [0,1]$ with $\alpha_1 + \alpha_2 = 1$, we have

$$H(\alpha_1 p_a + \alpha_2 p_b) - (\alpha_1 H(p_a) + \alpha_2 H(p_b)) = -\sum_{i=1}^{n} (\alpha_1 p_{a,i} + \alpha_2 p_{b,i}) \log(\alpha_1 p_{a,i} + \alpha_2 p_{b,i})$$

$$+ \sum_{i=1}^{n} \alpha_1 p_{a,i} \log(p_{a,i}) + \sum_{i=1}^{n} \alpha_2 p_{b,i} \log(p_{b,i})$$

$$= -\sum_{i=1}^{n} [\alpha_1 p_{a,i} \log(\alpha_1 + \alpha_2 \frac{p_{b,i}}{p_{a,i}}) + \alpha_2 p_{b,i} \log(\alpha_1 \frac{p_{a,i}}{p_{b,i}} + \alpha_2)]$$

$$\geq -\log\left(\sum_{i=1}^{n} \alpha_1 p_{a,i} \left(\alpha_1 + \alpha_2 \frac{p_{b,i}}{p_{a,i}}\right) + \alpha_2 p_{b,i} \left((\alpha_1 \frac{p_{a,i}}{p_{b,i}} + \alpha_2)\right)\right)$$

$$= -\log\left(\sum_{i=1}^{n} p_{a,i}\alpha_1^2 + \alpha_1\alpha_2 p_{b,i} + \alpha_2\alpha_1 p_{a,i} + p_{b,i}\alpha_2^2\right) = -\log((\alpha_1 + \alpha_2)^2) = 0$$

The inequality is from the concavity of the log function (note that $\sum_i \alpha_1 p_{a,i} + \alpha_2 p_{b,i} = 1$). Then the last two equalities hold due to $\sum_i p_{a,i} = 1$ and $\sum_i p_{b,i} = 1$, and since $\alpha_1 + alpha_2 = 1$.
Also we have to check that $\alpha_1 p_a + \alpha_2 p_b \in \Delta_N$, which means Δ_N is a convex set.
In conclusion, we have H is a concave function in Δ_N.

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

Solution. (5 pts)
Since H is concave in the simplex, $p_i \geq 0, \forall\ i$ and $\sum_{i=1}^{n} p_i = 1$, we have

$$H(p) \geq \sum_{i=1}^{n} p_i H(e_i),$$

where e_i is a distribution that the whole probability 1 goes to the i-th coordinates.
Since $H(e_i) = 0$ for any i, we have $H(p) \geq 0$.

We would like to find the maximum of $H(p)$ subjected to $\sum_{i=1}^{n} p_i = 1$. Using Lagrangian multipliers, we have

$$L(p, \alpha) = H(p) + \alpha(\sum_{i=1}^{n} p_i - 1),$$
$$\frac{\partial}{\partial p_i} L = -1 - \log p_i + \alpha.$$

Setting $\frac{\partial}{\partial p_i} L = 0$ for any i gives that p_i's are equal to each other, which means $p_i = 1/N$, for any i. Hence, $(1/N, 1/N, \ldots, 1/N)$ achieves the maximum entropy as $-N * (1/N)\log(1/N) = \log N$.

(c) Now let us move from a discrete to a continuous domain, by setting $\mathcal{X} = [0, R]$. Let $\mathcal{P}(\mathcal{X})$ now denote the space of probability distributions admitting a density $p(x)$, $x \in \mathcal{X}$ [1], and define the Shannon entropy as

$$H(p) = -\int_{\mathcal{X}} p(x) \log p(x) dx \ .$$

Show that the maximum entropy distribution in $\mathcal{P}(\mathcal{X})$ is the uniform measure in $\mathcal{X}$, with entropy $\log R$.

Solution. (5 pts)
We are to find the maximum of $H(p)$ subjected to $\int_{\mathcal{X}} p(x)dx = 1$. Using Lagrangian multipliers, we have

$$L(p, \alpha) = H(p) + \alpha(\int_{\mathcal{X}} p(x)dx - 1),$$
$$\frac{\partial}{\partial p} L = -\log p(x) - 1 + \alpha.$$

Setting $\frac{\partial}{\partial p} L = 0$ for any $x \in \mathcal{X}$ gives $p(x)$ is the same value across all x. So the maximum entropy distribution is the uniform distribution over $\mathcal{X}$, with entropy $-\int_{\mathcal{X}} \frac{1}{R} \log \frac{1}{R} dx = \log R$.

(d) We now attempt to understand maximum entropy distributions defined over $\mathcal{X} = \mathbb{R}$. For $p \in \mathcal{P}(\mathbb{R})$, denote the *spread* $\Delta_p := \sup_{p(x) \neq 0} x - \inf_{p(x) \neq 0} x$ as the smallest interval containing the support of p (where we abuse notation and identify a probability distribution in $\mathcal{P}(\mathbb{R})$ with its density). Prove that the maximum entropy distribution in $\mathcal{P}(\mathbb{R})$ with given mean c and spread $\Delta < \infty$ is the uniform distribution over the interval $(c - \Delta/2, c + \Delta/2)$, with entropy $\log \Delta$.

Solution. (5 pts)

[1] the technical condition is that the distribution ν is *absolutely continuous* with respect to the uniform Lebesgue measure μ, and we write $\nu \ll \mu$, and the density p is the *Radon-Nykodym* derivative of ν with respect to μ

From (c), we know that the maximum entropy distribution over an interval of finite length Δ is the uniform distribution with density $p \equiv \frac{1}{\Delta}$. It is easy to verify that the uniform distribution with mean c and spread Δ is exactly the uniform distribution over the interval $(c - \Delta/2, c + \Delta/2)$, with entropy $\log \Delta$.

(e) Conclude that the maximum entropy distribution over $\mathcal{P}(\mathbb{R})$ with given mean does not exist. How does the answer change if now we consider distributions over $\mathcal{P}(\mathbb{R}_+)$?

Solution. (10 pts)
The maximum entropy distribution over $\mathcal{P}(\mathbb{R})$ with given mean does not exist, because $\log \Delta$ from (d) with $\Delta \to \infty$ is unbounded.
For distribution over $\mathcal{P}(\mathbb{R}_+)$, we would like to optimize $H(P)$ subjected to $\int_{\mathcal{X}} p(x)dx = 1$ and $\int_{\mathcal{X}} x \cdot p(x)dx = c$. Using Lagrangian multiplier, we have

$$L(p, \alpha, \beta) = H(p) + \alpha(\int_{\mathcal{X}} p(x)dx - 1) + \beta(\int_{\mathcal{X}} x \cdot p(x)dx - c),$$
$$\frac{\partial}{\partial p} = -\log p(x) - 1 + \alpha + \beta x,$$

which means $p(x) = \exp(-1+\alpha+\beta x)$. Further computation of $\frac{\partial}{\partial \alpha} L = 0$, $\frac{\partial}{\partial \beta} L = 0$ gives the exact value of α and β, both of which are not zero. So the maximum entropy distribution over $\mathcal{P}(\mathbb{R}_+)$ is exponential distribution.

References

[1] NYU Center for Data Science. Ds-ga 1013 course, optimization-based data analysis. *Available online at* `http://www.cims.nyu.edu/~cfgranda/pages/OBDA_fall17/index.html`, Fall 2017.

[2] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by nonnegative matrix factorization. *Nature*, 401:788–791, 1999.

[3] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *In NIPS*, pages 556–562. MIT Press, 2000.