Inference and Representation, Fall 2020

Problem Set 1: PCA & Maximum Entropy

Due: Tuesday, Sep 29, 2020 (as a PDF and .zip files uploaded to NYU Classes)

---

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 Poisson random variable with mean $(WH)_{ij}$, show that the log-likelihood of the model is (up to a constant)

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

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

$$w_{ik} \quad \leftarrow \quad w_{ik} \frac{\sum_j h_{kj} x_{ij}/(WH)_{ij}}{\sum_j h_{kj}}\ , \tag{3}$$

$$h_{kj} \quad \leftarrow \quad h_{kj} \frac{\sum_i w_{ik} x_{ij}/(WH)_{ij}}{\sum_j w_{kj}}\ ,. \tag{4}$$

We shall study this algorithm and prove its correctness.

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

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

   (a) 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.

   (b) 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 $\sum_{k \leq r} c_k = 1$,

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

   (c) Deduce that

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

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

(d) Ignoring constants, show that

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

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

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

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] .)

   (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`)

   (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.

   (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).

   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.

3. *Factor Analysis and Principal Component Analysis.* Suppose X is a d-dimensional observation coming from an underlying unknown distribution. *Factor Analysis* (FA) is a generative model for the data of the form

$$X_i = \sum_{j=1}^{J} v_{i,j} Y_j + \mu_i + \epsilon_i \ , \ i = 1 \dots d \ ,$$

where $Y_1 \dots Y_J$ are uncorrelated random variables with $J < d$, and $\epsilon_1 \dots \epsilon_d$ are zero-mean, uncorrelated from the rest, and μ is a deterministic bias vector. The variables Y are interpreted as latent, common factors of variation across features, while ϵ explains the remaining variability.

   (a) Consider a Gaussian prior for the latent variables of the form $Y \sim \mathcal{N}(0, I_J)$ and $\epsilon \sim \mathcal{N}(0, \mathrm{diag}(\beta))$. Derive the joint likelihood of the data X as a function of the parameters of the model $V \in \mathbb{R}^{d \times J}$, $\beta \in \mathbb{R}^d$ and $\mu \in \mathbb{R}^d$.

   (b) Discuss the differences between Factor Analysis and Principal Component Analysis (with J principal components), by giving examples where both algorithms will return the same solution, and examples where FA and PCA differ.

(c) If one supposes that $\beta_i = \beta_0$ for $i = 1 \dots d$, give an algorithm to estimate the MLE parameters in that case.

(d) *Factor Analysis vs PCA on personality data.* The following questions also refer to the code and data `hw1.zip`. Complete the function `FactorAnalysis` in the file `fa.ipynb` to perform Factor Analysis of the psychological data. Include the plots generated by `FactorAnalysis.ipynb` in your submission. Briefly describe the differences between the PCA and Factor Analysis results.

4. *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.

(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 $\log N$.

(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 $\log R$.

(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 $\log \delta$.

(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 $\mathcal{P}(\mathbb{R}_+)$?

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.

[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

[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.