

Stats 545: Homework 4

Due before midnight on Sunday, Oct 21.
All plots should have labelled axes and titles.

Important: Rcode, tables and figures should be part of a single .pdf or .html files from R Markdown and knitr. See the class reading lists for a short tutorial. Any derivations can also be in Markdown, in Latex or neatly written on paper which you can give to me.

1 Problem 1: Exponential family distributions [20]

1. Consider a random variable x that can take D values and that is distributed according to the discrete distribution with parameters $\vec{\pi}$. We will write this as $p(x|\vec{\pi})$, with $p(x = c|\vec{\pi}) = \pi_c$ for $c \in \{1, \dots, D\}$.
 - (a) Write $p(x|\vec{\pi})$ as an exponential family distribution and give the natural parameters $\vec{\eta}$ as a function of π (note this means you can also write π as a function of η though you don't have to). Also write a *minimal* feature vector ϕ (note $\pi_D = 1 - \sum_{i=1}^{D-1} \pi_i$). [2 pts]
 - (b) Write $E[\phi(x)]$, the expectation of the feature vector ϕ as a function of the natural parameters $\vec{\eta}$. Recall that given some data $X = (x_1, \dots, x_N)$, maximum likelihood estimation (MLE) of η (and thus π) is moment matching (i.e. calculating the empirical average of ϕ and setting η so that the population average and the empirical averages match). [3 pts]
2. Let x be Poisson distributed with mean λ . Repeat parts (a), (b). [10 pts]
3. Let x be a 1-dimensional Gaussian with mean μ and variance σ^2 . Repeat parts (a), (b) (Note: both μ and σ^2 are parameters). [10 pts]
4. Let x follow a geometric distribution with success probability p : $\Pr(X = k) = (1 - p)^k p$ for $k = 0, 1, 2, \dots$. Repeat parts (a), (b). [10 pts]

2 Problem 2: EM for mixture of Bernoulli vectors [80]

1. We looked at the MNIST dataset last assignment. Write code to create a new dataset of only twos and threes using the information in `labels`. Each pixel can take values from 1 to 256: now threshold the images to be binary (0 or 1). Use a threshold between 1 to 5 (whatever you think is best). Do not use a for loop. [3]

We will model these binary images as a mixture of K Bernoulli vectors. Thus, we have K clusters, each of which is parametrized by a 784-dimensional vector with each component lying between 0 and 1. Call the k th cluster parameter μ^k , with $\mu^k \in [0, 1]^{784}$. The probability over clusters is a k -component probability vector π . Thus, to generate an observation, we first sample a cluster c from π , and then generate a random binary image x by setting the i th pixel to 1 with probability μ_i^c for i from 1 to 784.

2. Given N observations $X = (x_1, \dots, x_N)$ and their cluster assignments $C = (c_1, \dots, c_N)$, write down the log joint-probability $\log p(X, C|\pi, \vec{\mu})$. [4]
3. If we observed both X and C , what are the maximum likelihood estimates of π and the μ^k s? [4]

4. Explain why $p(C|X, \pi, \vec{\mu}) = \prod_{i=1}^N p(c_i|x_i, \pi, \vec{\mu})$. Write down $p(c_i|x_i, \pi, \vec{\mu})$. This is the q of the EM algorithm. [5]
5. Write down the variational lower bound $\mathcal{F}(q, \pi, \vec{\mu})$ for the EM algorithm. Use the first expression in the slides involving the entropy $H(q)$. [4]
6. For a given q , what are the π and $\vec{\mu}$ that maximize this? These expressions should be a simple relaxation of part (3). [5]
7. Write an EM algorithm that maximizes \mathcal{F} by alternately maximizing w.r.t. q (step 4) and $(\pi, \vec{\mu})$ (step 6). Although the algorithm doesn't require you to evaluate \mathcal{F} , your code should do this after each update. This is a useful diagnostic for debugging since \mathcal{F} should never decrease. Your stopping criteria should be when the value of \mathcal{F} stabilizes. [15]
8. Run the EM algorithm on the binary digits data set for $K = 2$ and 3. Plot the cluster parameters using `show_digit`. Also plot the trace of the evolution of \mathcal{F} . Write down the final value of π and \mathcal{F} . What are the units of the latter? [15]
9. The entropy of a distribution is a measure of how 'random' it is. For $K = 2$, calculate the entropy of the final $q(c_i|x_i, \vec{\mu}, \pi)$ of each digit, and plot the digit with the largest entropy. This is the digit with largest ambiguity about its correct cluster. [5]