STAT545 HW4

Yi Yang

10/3/2018

1. Problem 1: Exponential family distributions

1. Consider a random varibale x that can take D values and that is distributed according to the discrete distribution with parameters $\overrightarrow{\pi}$. We will wirte this as $p(x|\overrightarrow{\pi})$, with $p(x = c|\overrightarrow{\pi}) = \pi_c$ for $c \in \{1, \dots, D\}$.

(a) Write $p(x|\overrightarrow{\pi})$ as an exponential family distribution and give the natural parameters $\overrightarrow{\eta}$ 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 $\pi_D = 1 - \sum_{i=1}^{D-1} \pi_i$).

(b) Write $E[\phi(x)]$, the expectation of the feature vector ϕ as a function of the natural parameters $\overrightarrow{\eta}$. 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 π and setting η so that the population average and the empirical average match).

Solution:

1. For this discrete distribution, the exponential form is given as follows

(a)

$$
\begin{aligned}
p(x|\overrightarrow{\pi}) &= \prod_{k=1}^{D} \pi_k^{\delta(x=k)} \\
&= \exp(\sum_{k=1}^{D} \delta(x=k) \cdot \log(\pi_k))
\end{aligned}
$$

The feature vector is given by $\phi(x) = [\delta(x = 1), \delta(x = 2), \cdots, \delta(x = D)]^\top$. The natural parameter is given by $\eta = [\log(\pi_1), \log(\pi_2), \cdots, \log(\pi_D)]^\top$. The minial exponential form is derived as follows, the minial feature vector is given by $\phi(x) = [\delta(x = 1), \delta(x = 2), \cdots, \delta(x = D - 1)]^\top$.

$$
\begin{aligned}
p(x|\overrightarrow{\pi}) &= \exp(\sum_{k=1}^{D} \delta(x=k) \cdot \log(\pi_k)) \\
&= \exp(\sum_{k=1}^{D-1} \delta(x=k) \log(\frac{\pi_k}{\pi_D}) + \sum_{k=1}^{D-1} \delta(x=k) \log(\pi_D) + \delta(x=D) \log(\pi_D)) \\
&= \pi_D \exp(\sum_{k=1}^{D-1} \delta(x=k) \log(\frac{\pi_k}{\pi_D}))
\end{aligned}
$$

(b) The expectation of the feature vector is given by

$$
\mathbb{E}[\phi(x)] = \begin{bmatrix} \mathbb{E}[\delta(x=1)] \\ \vdots \\ \mathbb{E}[\delta(x=D)] \end{bmatrix} = \begin{bmatrix} \pi_1 \\ \vdots \\ \pi_D \end{bmatrix} = \exp(\eta)
$$

2. Let x be Poisson distributed with mean λ. Repeat parts (a), (b).

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).

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).

Solution:

2. The Poisson distribution can be expressed in exponential form as follows

$$\begin{aligned} p(x|\lambda) &= \frac{\lambda^x \exp(-\lambda)}{x!} \\ &= \exp(-\lambda)\frac{1}{x!}\exp(\log(\lambda)x) \end{aligned}$$

The feature vector and the minimal feature vector are both $\phi(x) = x$, the natural parameter is $\eta = \log(\lambda)$. The expectation of the featrue vector is derived as follows

$$\mathbb{E}[\phi(x)] = \mathbb{E}[x] = \sum_{x=1}^{\infty} \frac{\lambda^x \exp(-\lambda)}{(x-1)!} = \lambda = \exp(\eta)$$

3. The 1-D normal distribution can be expressed in exponential form as follows,

$$\begin{aligned} p(x|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \\ &= \frac{\exp\left(-\frac{\mu^2}{2\sigma^2}\right)}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x\right) \end{aligned}$$

The feature vector and the minimal feature vector are both $\phi(x) = [x^2, x]^\top$, the natural parameter is $\eta = [-\frac{1}{2\sigma^2}, \frac{\mu}{\sigma^2}]^\top$. The expectation of the featrue vector is given by

$$\mathbb{E}[\phi(x)] = \begin{bmatrix} \mathbb{E}[x^2] \\ \mathbb{E}[x] \end{bmatrix} = \begin{bmatrix} \mu^2 + \sigma^2 \\ \mu \end{bmatrix} = \begin{bmatrix} -\frac{1}{2\eta_1} + \frac{\eta_2^2}{4\eta_1^2} \\ -\frac{1}{2\eta_1} \end{bmatrix}$$

4. The geometric distribution can be expressed in exponential form as follows,

$$\begin{aligned} Pr(X=k) &= (1-p)^X p \\ &= \exp(X\log(1-p) + \log(p)) \end{aligned}$$

The feature vector and the minimal feature vector are both x, and the natural parameter is $\log(1-p)$. The expectation of the feature vector is given by

$$\mathbb{E}[\phi(x)] = \mathbb{E}[x] = \sum_{k=0}^{\infty} k(1-p)^k p = \frac{1-p}{p} = \frac{\exp(\eta)}{1-\exp(\eta)}$$

2. Problem 2: EM for mixture of Bernoulli vectors

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.

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 kth 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 ith pixel to 1 with probability μ_i^k for i from 1 to 784.

2. Given N observations $X = (x_1, \ldots, x_N)$ and their cluster assignments $C = (c_1, \ldots, c_N)$, write down the log joint-probability $\log p(X, C|\pi, \overrightarrow{\mu})$.

3. If we observed both X and C, what are the maximum likelihood estimates of π and the μ^ks?
4. Explain why $p(C|X, \pi, \overrightarrow{\mu}) = \prod_{i=1}^{N} p(c_i|x_i, \pi, \overrightarrow{\mu})$.. Write down $p(c_i|x_i, \pi, \overrightarrow{\mu})$. This is the q of the EM algorithm.
5. Write down the variational lower bound $\mathcal{F}(q, \pi, \overrightarrow{\mu})$ for the EM algorithm. Use the first expression in the slides involving the entropy $H(q)$.
6. For a given q, what are the π and $\overrightarrow{\mu}$ that maximize this? These expressions should be a simple relaxation of part (3).
7. Write an EM algorithm that maximizes $\mathcal{F}$ by alternately maximizing w.r.t. q (step 4) and $(\pi, \overrightarrow{\mu})$ (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.
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?
9. The entropy of a distribution is a measure of how 'random' it is. For $K = 2$, calculate the entropy of the final $q(c_i|x_i, \overrightarrow{\mu}, \pi)$ of each digit, and plot the digit with the largest entropy. This is the digit with largest ambiguity about its correct cluster.