Problem Set 4

Pu Rui 2016280148
Machine Learning

June 16, 2017

Problem 1.1:

For each observation there is a vector which supports $\sum_{k=1}^{K} x_k = 1$ (for N observations $\sum_{k=1}^{K} x_k = N$).

If we denote probability of $x_k = 1$ by parameter μ_i then the distribution of x is given:

$$p(X|\mu) = \prod_{k=1}^{K} \mu_k^{x_k}$$

where $\mu = (\mu_1, ..., \mu_K)^T$ and the parameters μ_k are between zero and one and $\sum_{k=1}^{K} \mu_k = 1$ because they represent probabilities. Also we can see that the distribution is normalized.

$$\sum_x p(x|\mu) = \sum_{k=1}^{K} \mu_k = 1$$

and that:

$$E[x|\mu] = \sum_x p(x|\mu)x = (\mu_1, ..., \mu_K)^T = \mu$$

Now consider a dataset X of n independent observations $x_1, ..., x_N$. The corresponding likelihood function takes the form:

$$p(X|\mu) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_n k} = \prod_{k=1}^{K} \mu_k^{\sum_{n=1}^{N} x_{nk}} = \prod_{k=1}^{K} \mu_k^{x_k} (1)$$

From (1) it takes the form below, which is known as the multinomial distribution and $\sum_{k=1}^{K} x_k = N$.

$$p(X|\mu) = \frac{N!}{\prod_k x_k!} \prod_{k=1}^{K} \mu_k^{k_i}$$

The normalization coefficient is the number of ways of partitioning n objects into k groups of size $x_1, ..., x_k$

We see that the likelihood function depends on the n observations through K quantities:

$$x_k = \sum_{n=1}^{N} x_{nk}$$

x_k represents the number of observations in cluster K

In order to find the maximum likelihood solution forμ, we need to maximize $\ln p(X|\mu)$ with respect to μ_k taking into consideration the constraint $\sum_k \mu_k = 1$, in order to do so we use Lagrange multiplier λ and maximizing:

$$\ln(p(X|\mu)) = \ln(N!) - \sum_{k=1}^{K} (x_k!) + \sum_{k=1}^{K} x_k \ln \mu_k + \lambda(\sum_{k=1}^{K} \mu_k - 1)$$
$$\frac{\partial ln(p(X|\mu))}{\partial \mu_k} = \frac{x_k}{\mu_k} + \lambda = 0 \Rightarrow \mu_k = \frac{-x_k}{\lambda}$$

After substituting this into $\sum_k \mu_k = 1 \Rightarrow \lambda = -n \Rightarrow \mu_k^{ML} = \frac{x_k}{N}$

Problem 1.2:

We introduce a latent variable c_d corresponding to each cluster.

Conditional distribution of the observed data set, given the latent variable is:

$$p(d|c_d = k, \mu) = \frac{n_d!}{\prod_w T_{dw}!} \prod_w \mu_{wk}^{T_{dw}}, n_d = \sum_w T_{dw}$$

And the distribution of latent variable is given by:

$$p(c_d = k) = \pi_k$$

And :

$$p(d) = \sum_{k=1}^{K} P(d|c_d = k)p(c_d = k) = \frac{n_d!}{\prod_w T_{dw}} \sum_{k=1}^{K} \pi_k \prod_w \mu_{wk}^{T_{dw}}$$

Responsibilities given by:

$$\gamma(c_d = k) = E[c_d = k] = \frac{\pi_k \prod_w \mu_{wk}^{T_{dw}}}{\sum_{i=1}^{K} \pi_i \prod_w \mu_{wi}^{T_{dw}}}$$

These represent E-step equations

To derive the M-step equation we add to the expected complete-data log likelihood function a set of Lagrange multiplier terms enforcing constraints $\sum_{k=1}^{K} \pi_k = 1$ as well as $\sum_{k=1}^{K} \mu_{wk} = 1$ and maximizing with respect to mixing coefficient π_k , after eliminating the Lagrange multiplier λ, we have:

$$\pi_k = \frac{\sum_{d=1}^{D} \gamma(c_d = k)}{D}$$

Then maximizing with respect toμ_{wk} and eliminating Lagrange multipliers we have:

$$\mu_{wk} = \frac{1}{n_k} \sum_{d=1}^{D} \gamma(c_d = k) T_{dw}$$

Where $n_k = \sum_{d=1}^{D} \gamma(c_d = k) n_d$, this shows value of μ_{wk} is given by fraction of those counts assigned to cluster k.