1 Problem

1.1 Solution

Optimal μ_k is obtained by finding first derivative of D w.r.t μ_k and setting it to zero.

$$D = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|\mathbf{x}_n - \mu_k\|_2^2$$

$$\frac{\partial D}{\partial \mu_k} = 0 \Rightarrow \sum_{n=1}^{N} r_{nk} 2(\mu_k - \mathbf{x}_n) = 0$$

$$\Rightarrow \mu_k = \frac{\sum_{n=1}^{N} r_{nk}\mathbf{x}_n}{\sum_{n=1}^{N} r_{nk}}$$

We can also see that the second derivative ($2\sum_{n=1}^{N} r_{nk}$) is positive , so we know that D will be minimized for the above μ_k value.

1.2 Solution

Let us define C_n^t as cluster center that is assigned to data point $\mathbf{x}_n$ at time t i.e.,

$$C_n^t = \mu_k \quad where \quad r_{nk}^t = 1$$

By this definition, our objective function can be written as:

$$D = \sum_{n=1}^{N} \|\mathbf{x}_n - C_n\|_2^2 \quad \text{...introducing time t in this expression}$$

$$D(\mu^t, \mathbf{R}^t) = \sum_{n=1}^{N} \left\|\mathbf{x}_n - C_n^t\right\|_2^2$$

Also from the way K-means algorithm is defined, we know that C_n^t is chosen such that it is minimum of distance from data point $\mathbf{x}_n$ to all cluster centers, so we know that:

$$\forall n \in [1, N], \forall k \in [1, K] : \left\|\mathbf{x}_n - C_n^t\right\|_2^2 \leq \|\mathbf{x}_n - \mu_k\|_2^2 \quad \text{...i.e.,}$$

$$\forall n \in [1, N] : \left\|\mathbf{x}_n - C_n^t\right\|_2^2 \leq \left\|\mathbf{x}_n - C_n^{t-1}\right\|_2^2$$

Combining the above two, we get:

$$D(\mu^t, \mathbf{R}^t) = \sum_{n=1}^{N} \left\|\mathbf{x}_n - C_n^t\right\|_2^2$$

$$\leq \sum_{n=1}^{N} \left\|\mathbf{x}_n - C_n^{t-1}\right\|_2^2 = D(\mu^{t-1}, \mathbf{R}^{t-1})$$

1.3 Solution

Again, we get the minimum by doing first derivative of objective function and setting it to zero:

$$D = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \mu_k\|_1$$

In the above, $\mu_k \in \mathbb{R}^d$ i.e., μ_k is a vector of dimension k. Let's consider the derivative of a given μ_k along one dimension value d:

$$\begin{aligned}
\frac{\partial D}{\partial \mu_{kd}} = 0 \Rightarrow & \sum_{n=1}^{N} r_{nk} \; sgn(x_{nd} - \mu_{kd}) = 0 \\
& \Rightarrow \sum \mathbb{I}(C_n = \mu_k)\mathbb{I}(x_{nd} > \mu_{kd}) - \sum \mathbb{I}(C_n = \mu_k)\mathbb{I}(x_{nd} \leq \mu_{kd}) = 0 \\
& \Rightarrow \sum \mathbb{I}(C_n = \mu_k)\mathbb{I}(x_{nd} > \mu_{kd}) = \sum \mathbb{I}(C_n = \mu_k)\mathbb{I}(x_{nd} \leq \mu_{kd})
\end{aligned}$$

where C_n is the cluster center assigned to data point $\mathbf{x}_n$ and sgn is the sign function. From the above, we can see that along dimension d, from the data points assigned to cluster k, the number of data points where $x_{nd} > \mu_{kd}$ should be equal to points where $x_{nd} \leq \mu_{kd}$ i.e., μ_{kd} is the median of values at dimension d of all data points assigned to cluster k. From this, it follows that μ_k that minimizes D is the element wise median of all data points assigned to cluster k.

1.4 Solution

Given $\tilde{\mu}_k$ is the center of cluster in feature space:

$$\tilde{\mu}_k = \frac{\sum_n r_{nk}\phi(\mathbf{x}_n)}{\sum_n r_{nk}} = \frac{\sum_n r_{nk}\phi(\mathbf{x}_n)}{N_k} \quad \text{...where } N_k = \sum_n r_{nk}$$

$$\begin{aligned}
\tilde{D} &= \sum_n \sum_k r_{nk}\|\phi(\mathbf{x}_n) - \tilde{\mu}_k\|_2^2 \\
&= \sum_k \sum_n r_{nk}(\phi(\mathbf{x}_n) - \tilde{\mu}_k)^T(\phi(\mathbf{x}_n) - \tilde{\mu}_k) \\
&= \sum_k \sum_n r_{nk}[\phi(\mathbf{x}_n)^T\phi(\mathbf{x}_n) - \phi(\mathbf{x}_n)^T\tilde{\mu}_k - \tilde{\mu}_k^T\phi(\mathbf{x}_n) + \tilde{\mu}_k^T\tilde{\mu}_k] \\
&= \sum_k \sum_n r_{nk}[f(\mathbf{x}_n, \mathbf{x}_n) - 2\phi(\mathbf{x}_n)^T\frac{\sum_m r_{mk}\phi(\mathbf{x}_m)}{N_k} + \frac{\sum_l r_{lk}\phi(\mathbf{x}_l)^T}{N_k}\frac{\sum_m r_{mk}\phi(\mathbf{x}_m)}{N_k}] \\
&= \sum_k \sum_n r_{nk}[f(\mathbf{x}_n, \mathbf{x}_n) - 2\frac{\sum_m r_{mk}f(\mathbf{x}_n, \mathbf{x}_m)}{N_k} + \frac{\sum_l \sum_m r_{lk}r_{mk}f(\mathbf{x}_l, \mathbf{x}_m)}{N_k^2}] \\
&= \sum_n f(\mathbf{x}_n, \mathbf{x}_n)\sum_k r_{nk} - \sum_k 2\frac{\sum_n \sum_m r_{nk}r_{mk}f(\mathbf{x}_n, \mathbf{x}_m)}{N_k} + \sum_k \frac{\sum_l \sum_m r_{lk}r_{mk}f(\mathbf{x}_l, \mathbf{x}_m)}{N_k^2}\sum_n r_{nk} \\
&= \sum_n f(\mathbf{x}_n, \mathbf{x}_n) - \sum_k \frac{\sum_l \sum_m r_{lk}r_{mk}f(\mathbf{x}_l, \mathbf{x}_m)}{N_k}
\end{aligned}$$

2 Problem

2.1 Solution

If k is the index where $z_k = 1$, the probability of X is given by distribution θ_k i.e.,

$$\begin{aligned}
P(X = \mathbf{x}|z_k = 1) &= \prod_{i=1}^{d} \theta_{ki}^{x_i}(1 - \theta_{ki})^{(1-x_i)} \quad \text{...can also be expressed as} \\
&= \prod_{i=1}^{d} \theta_{ki}^{z_k x_i}(1 - \theta_{ki})^{z_k(1-x_i)} \quad \text{...since } z_k = 1
\end{aligned}$$

Also, we can see that when $z_{k'} = 0$ (all other dimension values for $\mathbf{z}$), the term $\pi_{k'}^{z'_k} = 1$ and $\forall i \in [1, d], \theta_{k'i}^{z'_k x_i}(1 - \theta_{k'i})^{z'_k(1-x_i)} = 1$ as well. So multiplying any value with these terms

won't change the value. We will write the joint probability from this observation as:

$$P(X = \mathbf{x}, Z = \mathbf{z}|\Theta, \pi) = \prod_{k=1}^{K} \pi_k^{z_k} \prod_{i=1}^{d} \theta_{ki}^{z_k x_i} (1 - \theta_{ki})^{z_k(1-x_i)}$$

The marginal probability can be written as below...

$$P(X = \mathbf{x}|\Theta, \pi) = \sum_{k=1}^{K} \pi_k \prod_{i=1}^{d} \theta_{ki}^{x_i} (1 - \theta_{ki})^{(1-x_i)}$$

Note that if we know a specific k for which $z_k = 1$ in $\mathbf{z}$, the expression for $P(X = \mathbf{x}, Z = \mathbf{z}|\Theta, \pi)$ can be simplified greatly (which is the case for next solution).

2.2 Solution

$$\begin{aligned}\gamma_{nk} = P(z_k = 1|X = \mathbf{x}_n, \Theta, \pi) &= \frac{P(z_k = 1, X = \mathbf{x}_n|\Theta, \pi)}{P(X = \mathbf{x}_n|\Theta, \pi)} \\ &= \frac{\pi_k \prod_{i=1}^{d} \theta_{ki}^{x_{ni}} (1 - \theta_{ki})^{(1-x_{ni})}}{\sum_{k=1}^{K} \pi_k \prod_{i=1}^{d} \theta_{ki}^{x_{ni}} (1 - \theta_{ki})^{(1-x_{ni})}}\end{aligned}$$

2.3 Solution

The expression that we need to maximize w.r.t Θ and π is given by:

$$\begin{aligned}Q &= \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk}\ lnP(\mathbf{x}_n, z_k = 1|\Theta, \pi) \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} ln[\pi_k \prod_{i=1}^{d} \theta_{ki}^{x_{ni}} (1 - \theta_{ki})^{(1-x_{ni})}] \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk}\ ln(\pi_k) + \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \sum_{i=1}^{d} [x_{ni}\ ln(\theta_{ki}) + (1 - x_{ni})\ ln(1 - \theta_{ki})]\end{aligned}$$

We can consider only the first term for deriving updates for π and second term for deriving updates for Θ since other terms would be constants in each derivation.

Update for π is solved using the lagrangian with constraints and applying KKT conditions:

$$\text{Constraints: } \pi_k \geq 0 \ \forall k; \sum_{k=1}^{K} \pi_k = 1$$

$$L = \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma_{nk} \ ln(\pi_k) + \mu(1 - \sum_{k=1}^{K} \pi_k) - \sum_{k=1}^{K} \lambda_k \pi_k$$

$$\lambda_k = 0 \ \forall k \qquad \text{...from complimentary slackness}$$

$$\frac{\partial L}{\partial \pi_k} = 0 \Rightarrow \sum_{n=1}^{N} \frac{\gamma_{nk}}{\pi_k} - \mu - \lambda_k = 0$$

$$\Rightarrow \pi_k = \frac{\sum_{n=1}^{N} \gamma_{nk}}{\mu} \qquad \text{...putting this back in constraint}$$

$$\sum_{k=1}^{K} \frac{\sum_{n=1}^{N} \gamma_{nk}}{\mu} = 1 \Rightarrow \mu = \sum_{k=1}^{K}\sum_{n=1}^{N} \gamma_{nk} = \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma_{nk} = \sum_{n=1}^{N} 1 = N$$

Putting this back in expression for π_k, we have:

$$\pi_k = \frac{\sum_{n=1}^{N} \gamma_{nk}}{N}$$

Update for each θ_{ki} can be obtained by equating the first derivative of corresponding term to zero:

$$\frac{\partial Q}{\partial \theta_{ki}} = 0 \Rightarrow \sum_{n=1}^{N} \gamma_{nk} [\frac{x_{ni}}{\theta_{ki}} - \frac{1 - x_{ni}}{1 - \theta_{ki}}] = 0$$

$$\Rightarrow \frac{1 - \theta_{ki}}{\theta_{ki}} = \frac{\sum_{n=1}^{N} \gamma_{nk}(1 - x_{ni})}{\sum_{n=1}^{N} \gamma_{nk} x_{ni}}$$

$$\Rightarrow \frac{1}{\theta_{ki}} = 1 + \frac{\sum_{n=1}^{N} \gamma_{nk}(1 - x_{ni})}{\sum_{n=1}^{N} \gamma_{nk} x_{ni}} = \frac{\sum_{n=1}^{N} \gamma_{nk}}{\sum_{n=1}^{N} \gamma_{nk} x_{ni}}$$

From the above, we can write the expression for θ_{ki} as:

$$\theta_{ki} = \frac{\sum_{n=1}^{N} \gamma_{nk} x_{ni}}{\sum_{n=1}^{N} \gamma_{nk}}$$

Same thing can be expressed in vector form as:

$$\theta_k = \frac{\sum_{n=1}^{N} \gamma_{nk} \mathbf{x}_n}{\sum_{n=1}^{N} \gamma_{nk}}$$

3 Problem

3.1 Solution

From the class (not showing the proof again), we know the value for sequence probability:

$$P(X_{1:T} = x_{1:T}) = \sum_{s'} \alpha_{s'}(T) \quad \text{...so}$$

$$\begin{aligned}
P(Z_{T+1} = s | X_{1:T} = x_{1:T}) &= \sum_{s'} P(Z_{T+1} = s, Z_T = s' | X_{1:T} = x_{1:T}) && \text{...from marginal probability} \\
&= \frac{\sum_{s'} P(Z_{T+1} = s, Z_T = s', X_{1:T} = x_{1:T})}{P(X_{1:T} = x_{1:T})} && \text{...from conditional probability} \\
&= \frac{\sum_{s'} P(Z_{T+1} = s | Z_T = s', X_{1:T} = x_{1:T}) P(Z_T = s', X_{1:T} = x_{1:T})}{P(X_{1:T} = x_{1:T})} && \text{...chain rule} \\
&= \frac{\sum_{s'} P(Z_{T+1} = s | Z_T = s') P(Z_T = s', X_{1:T} = x_{1:T})}{P(X_{1:T} = x_{1:T})} && \text{...markov property} \\
&= \frac{\sum_{s'} a_{s',s} \alpha_{s'}(T)}{\sum_{s'} \alpha_{s'}(T)} && \text{...substituting values}
\end{aligned}$$

3.2 Solution

$$\begin{aligned}
P(Z_{T+k} = s | X_{1:T} = x_{1:T}) &= \sum_{s'} P(Z_{T+k} = s, Z_{T+k-1} = s' | X_{1:T} = x_{1:T}) && \text{...marginal probability} \\
&= \sum_{s'} P(Z_{T+k} = s | Z_{T+k-1} = s', X_{1:T} = x_{1:T}) P(Z_{T+k-1} = s' | X_{1:T} = x_{1:T}) && \text{...chain rule} \\
&= \sum_{s'} P(Z_{T+k} = s | Z_{T+k-1} = s') P(Z_{T+k-1} = s' | X_{1:T} = x_{1:T}) && \text{...markov property} \\
&= \sum_{s'} a_{s',s} P(Z_{T+k-1} = s' | X_{1:T} = x_{1:T}) && \text{...substituting value}
\end{aligned}$$