Homework #5

Instructions

Notes on notation:

- Unless stated otherwise, scalars are denoted by small letter in normal font, vectors are denoted by small letters in bold font and matrices are denoted by capital letters in bold font.

- $\|.\|$ means L2-norm unless specified otherwise i.e. $\|.\| = \|.\|_2$

Problem 1 Optimization over the simplex (10 points)

Many machine learning problems, especially clustering problems, involve optimization over a **simplex**. Thus, in this exercise, you will prove two optimization results over the simplex that we used multiple times in the lectures. Recall that a $K - 1$ dimensional simplex Δ is defined as

$$\Delta = \{\mathbf{q} \in \mathbb{R}^K \mid q_k \geq 0, \ \forall k \text{ and } \sum_{k=1}^{K} q_k = 1\},$$

which means that a $K - 1$ dimensional simplex has K non-negative entries, and the sum of all K entries is 1. The property that a simplex sums to one coincides with a probability distribution of a discrete random variable of K possible outcomes, and is thus frequently used in clustering problems.

1.1 Given $a_1, \ldots, a_K \in \mathcal{R}^+$, solve the following optimization over simplex (find the optimal value of q_k.) **(5 points)**

$$\arg\max_{\mathbf{q}} \sum_{k=1}^{K} a_k \ln q_k,$$
$$\text{s.t. } q_k \geq 0,$$
$$\sum_{k=1}^{K} q_k = 1.$$

The stationary condition states that for each k, **(2 points)**

$$\frac{a_k}{q_k^*} + \lambda + \lambda_k = 0 \quad (\text{with } \lambda_k \geq 0)$$

$\frac{a_k}{q_k^*} + \lambda - \lambda_k = 0$ is also acceptable if we constrain $\lambda_k \leq 0$.
and thus

$$q_k^* = -\frac{a_k}{\lambda + \lambda_k} \neq 0.$$

The complementary slackness condition implies that $\lambda_k q_k^* = 0$ and thus $\lambda_k = 0$. **(2 points)**
Finally, feasibility implies

$$\sum_{k=1}^{K} q_k^* = -\sum_{k=1}^{K} \frac{a_k}{\lambda} = 1.$$

Solving for λ and plugging it back gives the solution $q_k^* = \frac{a_k}{\sum_{k'} a_{k'}}$. **(1 points)**

1.2 Given $b_1, \ldots, b_K \in \mathcal{R}$, solve the following optimization problem **(5 points)**

$$\arg\max_{\mathbf{q} \in \Delta} \sum_{k=1}^{K} (q_k b_k - q_k \ln q_k),$$
$$\text{s.t. } q_k \geq 0,$$
$$\sum_{k=1}^{K} q_k = 1.$$

The Lagrangian of this problem is

$$L(\mathbf{q}, \lambda, \lambda_1, \ldots, \lambda_K) = \sum_{k=1}^{K} (q_k b_k - q_k \ln q_k) + \lambda \left(\sum_{k=1}^{K} q_k - 1 \right) + \sum_{k=1}^{K} \lambda_k q_k, \quad (\text{with } \lambda_k \geq 0)$$

$\sum_{k=1}^{K} (q_k b_k - q_k \ln q_k) + \lambda \left(\sum_{k=1}^{K} q_k - 1 \right) - \sum_{k=1}^{K} \lambda_k q_k$ is also acceptable if we constrain $\lambda_k \leq 0$.
The stationary condition implies that for each k **(2 points)**

$$b_k - 1 - \ln q_k + \lambda + \lambda_k = 0,$$

and thus
$$q_k = \exp(b_k - 1 + \lambda + \lambda_k) \propto e^{b_k + \lambda_k} \neq 0.$$

Complementary slackness then implies $\lambda_k = 0$ and thus $q_k \propto e^{b_k}$. **(2 points)**
Therefore $q_k = \frac{e^{b_k}}{\sum_{k'} e^{b_{k'}}}$. **(1 points)**

Problem 2   Gaussian Mixture Model and EM   (15 points)

2.1 In the lecture we applied EM to learn Gaussian Mixture Models (GMMs) and showed the M-Step without a proof. Now it is time that you prove it.

Consider a GMM with the following PDF of $\mathbf{x}_n$:

$$p(\mathbf{x}_n) = \sum_{k=1}^{K} \omega_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \Sigma_k) = \sum_{k=1}^{K} \frac{\omega_k}{\left(\sqrt{2\pi}\right)^D |\Sigma_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \Sigma_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)\right)$$

where $K \in \mathbb{N}$ is the number of Gaussian components, $D \in \mathbb{N}$ is dimension of a data point $\mathbf{x}_n$. This GMM has K tuples of model parameters $(\boldsymbol{\mu}_k, \Sigma_k, \omega_k)$, which standards for the mean vector, covariance matrix, and component weight of the k-th Gaussian component. $|\Sigma|$ denotes the determinant of matrix Σ.

For simplicity, we further assume that all components are isotropic Gaussian, i.e., $\Sigma_k = \sigma_k^2 I$. Find the MLE of *the expected complete log-likelihood*. Equivalently, find the optimal solution to the following optimization problem.

$$\underset{\omega_k, \boldsymbol{\mu}_k, \Sigma_k}{\arg\max} \sum_n \sum_k \gamma_{nk} \ln \omega_k + \sum_n \sum_k \gamma_{nk} \ln \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \Sigma_k)$$

$$\text{s.t. } \omega_k \geq 0$$

$$\sum_{k=1}^{K} \omega_k = 1$$

where γ_{nk} is the posterior of latent variables computed from the E-Step. Hint: you can make use of the result from Problem 1.1.   **(10 points)**

To find $\omega_1, \ldots, \omega_K$, we simply solve

$$\underset{\omega}{\arg\max} \sum_n \sum_k \gamma_{nk} \ln \omega_k.$$

$$\text{s.t. } \omega_k \geq 0$$

$$\sum_{k=1}^{K} \omega_k = 1 \qquad \textbf{(2 points)}$$

According to Problem 1.1 with $a_k = \sum_n \gamma_{nk}$, the solution is

$$\omega_k^* = \frac{\sum_n \gamma_{nk}}{\sum_k \sum_n \gamma_{nk}} = \frac{\sum_n \gamma_{nk}}{\sum_n 1} = \frac{\sum_n \gamma_{nk}}{N}. \qquad \textbf{(2 points)}$$

To find $\boldsymbol{\mu}_k$ and σ_k, we solve for each k

$$\underset{\boldsymbol{\mu}_k, \Sigma_k}{\arg\max} \sum_n \gamma_{nk} \ln \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \Sigma_k) = \underset{\boldsymbol{\mu}_k, \Sigma_k}{\arg\max} \sum_n \gamma_{nk} \ln \left[\frac{1}{(\sqrt{2\pi}\sigma_k)^D} \exp\left(-\frac{1}{2\sigma_k^2} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2\right) \right]$$

$$= \underset{\boldsymbol{\mu}_k, \Sigma_k}{\arg\max} \sum_n \gamma_{nk} \left(-D \ln \sigma_k - \frac{\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2}{2\sigma_k^2}\right),$$

where D is the length of $\mathbf{x}_n$.   **(2 points)**

First we set the derivative w.r.t. $\boldsymbol{\mu}_k$ to 0:

$$\frac{1}{\sigma_k^2} \sum_n \gamma_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k) = 0,$$

which gives

$$\mu_k^* = \frac{\sum_n \gamma_{nk} \mathbf{x}_n}{\sum_n \gamma_{nk}} \qquad \textbf{(2 points)}$$

Next we set the derivative w.r.t. σ_k to 0:

$$\sum_n \gamma_{nk} \left(-\frac{D}{\sigma_k} + \frac{\|\mathbf{x}_n - \mu_k\|^2}{\sigma_k^3} \right) = 0.$$

Solving for σ_k gives

$$(\sigma_k^*)^2 = \frac{\sum_n \gamma_{nk} \|\mathbf{x}_n - \mu_k^*\|^2}{D \sum_n \gamma_{nk}}. \qquad \textbf{(2 points)}$$

2.2 In the lecture we derived EM through a lower bound of the log-likelihood function. Specifically, we find the tightest lower bound by solving

$$\arg\max_{\mathbf{q}_n \in \Delta} \mathbb{E}_{z_n \sim \mathbf{q}_n} \left[\ln p(\mathbf{x}_n, z_n \,; \theta^{(t)}) \right] - \mathbb{E}_{z_n \sim \mathbf{q}_n} \left[\ln \mathbf{q}_n \right].$$

Find the optimal solution to this optimization problem using the results from Problem 1.2. (The solution was already given in the lecture, but here we require you to derive it.)

(2 points)

This is exactly in the same form of the problem in 1.2 with $b_k = \ln p(\mathbf{x}_n, z_n = k \,; \theta^{(t)})$. So the solution is

$$q_{nk}^* \propto p(\mathbf{x}_n, z_n = k \,; \theta^{(t)}),$$

or in other words,

$$q_{nk}^* = \frac{p(\mathbf{x}_n, z_n = k \,; \theta^{(t)})}{\sum_{k=1}^K p(\mathbf{x}_n, z_n = k \,; \theta^{(t)})} = \frac{p(\mathbf{x}_n, z_n = k \,; \theta^{(t)})}{p(\mathbf{x}_n \,; \theta^{(t)})} = p(z_n = k \mid \mathbf{x}_n \,; \theta^{(t)}).$$

2.3 The posterior probability of z in GMM can be seen as a *soft* assignment to the clusters; In contrast, K-means assign each data point to one cluster at each iteration (*hard* assignment). Describe how to set the model parameters such that the GMM in the previous question reduces to a K-means; please also derive $p(z_n = k | x_n)$ in this case.

(3 points)

Set all $\sigma_k = \sigma \to 0$ and $\omega_k = \frac{1}{K}$, we have

$$p(x_n, z_n = k) \propto \exp -\frac{1}{2\sigma_k^2} \|x_n - \mu_k\|^2,$$

where constant terms are ignored. **(2 points)**

The posterior then becomes

$$p(z_n = k | x_n) = \lim_{\sigma \to 0} \frac{\exp\{-\frac{1}{2\sigma^2} \|x_n - \mu_k\|^2\}}{\sum_j \exp\{-\frac{1}{2\sigma^2} \|x_n - \mu_j\|^2\}} \to \begin{cases} 1, & \text{if } k = \arg_j \min \|x_n - \mu_j\|^2 \\ 0, & \text{otherwise.} \end{cases}$$

(1 points)

Problem 3 Hidden Markov Model   (15 points)

Recall that an HMM is parameterized as follows:

- initial state distribution $P(X_1 = s) = \pi_s$

- transition distribution $P(X_{t+1} = s' \mid X_t = s) = a_{s,s'}$

- emission distribution $P(O_t = o \mid X_t = s) = b_{s,o}$

3.1 Suppose we observe a sequence of outcomes $o_1, \ldots, o_T$ and wish to predict the next state X_{T+1}

$$P(X_{T+1} = s \mid O_{1:T} = o_{1:T}).$$

Denote the forward message as

$$\alpha_s(T) = P(X_T = s, O_{1:T} = o_{1:T}).$$

Please derive how $P(X_{T+1} = s \mid O_{1:T} = o_{1:T})$ can be represented using $\alpha_s(T)$.   **(7 points)**

$$P(X_{T+1} = s \mid O_{1:T} = o_{1:T}) = \frac{P(X_{T+1} = s, O_{1:T} = o_{1:T})}{P(O_{1:T} = o_{1:T})}$$

$$\text{(marginalization)} = \frac{\sum_{s'} P(X_{T+1} = s, X_T = s', O_{1:T} = o_{1:T})}{\sum_{s''} P(X_T = s'', O_{1:T} = o_{1:T})} \quad \textbf{(2 points)}$$

$$\text{(factorization)} = \frac{\sum_{s'} P(X_T = s', O_{1:T} = o_{1:T})P(X_{T+1} = s \mid X_T = s', O_{1:T} = o_{1:T})}{\sum_{s''} P(X_T = s'', O_{1:T} = o_{1:T})} \textbf{(2 points)}$$

$$\text{(Markov property)} = \frac{\sum_{s'} P(X_T = s', O_{1:T} = o_{1:T})P(X_{T+1} = s \mid X_T = s')}{\sum_{s''} P(X_T = s'', O_{1:T} = o_{1:T})} \quad \textbf{(2 points)}$$

$$= \frac{\sum_{s'} \alpha_{s'}(T)a_{s',s}}{\sum_{s''} \alpha_{s''}(T)} \quad \textbf{(1 points)}$$

3.2 Describe how to set the model parameters such that an HMM reduces to the GMM described in Problem 2. In addition, derive the posterior probability $P(X_2 = s \mid O_1 = o_1, O_2 = o_2)$ using the parameters you set to show that the HMM really reduces to GMM. Hint: compare your result with the posterior $p(z|x)$ of GMM in Problem 2.2.   **(8 points)**

- Set the initial state distribution $P(X_1 = s) = \omega_s$.

- Set the transition distribution $P(X_{t+1} = s' \mid X_t = s) = P(X_{t+1} = s') = \omega_{s'}$.
  The deprives HMM of its Markov property, and the HMM thus becomes independent across temporal/sequential order (as we will see when deriving the posterior).

- Set the emission distribution $P(O_t = o \mid X_t = s) = \mathcal{N}(o|\boldsymbol{\mu}_s, \sigma_s^2 \mathrm{I})$

The *state* in HMM becomes the *cluster* in GMM.   **(4 points)**

Now, to derive $P(X_2 = s' \mid O_1 = o_1, O_2 = o_2)$, we start from the definition of HMM factorization:

$$P(X_2 = s' \mid O_1 = o_1, O_2 = o_2) = \sum_s P(X_1 = s, X_2 = s' \mid O_1 = o_1, O_2 = o_2),$$

$$P(X_1 = s, X_2 = s' \mid O_1 = o_1, O_2 = o_2) = \frac{P(X_1 = s, X_2 = s', O_1 = o_1, O_2 = o_2)}{\sum_s \sum_{s'} P(X_1 = s, X_2 = s', O_1 = o_1, O_2 = o_2)},$$

$$P(X_1 = s, X_2 = s', O_1 = o_1, O_2 = o_2) = P(X_1 = s)P(X_2 = s' \mid X_1 = s)P(O_1 = o_1|X_1 = s)P(O_2 = o_2|X_2 = s').$$

Replacing the four terms in the joint PDF with the model parameters we set earlier, we have

$$P(X_1 = s, X_2 = s', O_1 = o_1, O_2 = o_2) = \omega_s \omega_{s'} \mathcal{N}(o_1 | \boldsymbol{\mu}_s, \sigma_s^2 I) \mathcal{N}(o_2 | \boldsymbol{\mu}_{s'}, \sigma_{s'}^2 I)$$

$$= \omega_s \mathcal{N}(o_1 | \boldsymbol{\mu}_s, \sigma_s^2 I) \, \omega_{s'} \mathcal{N}(o_2 | \boldsymbol{\mu}_{s'}, \sigma_{s'}^2 I)$$

$$= P(X_1 = s \mid O_1 = o_1) \, P(X_2 = s' \mid O_2 = o_2).$$

(This shows that by our model setting, the HMM has lost its temporal/sequential property; the data is now independent across time.) **(2 points)**

Thus,

$$P(X_2 = s' \mid O_1 = o_1, O_2 = o_2) = \frac{\sum_s \omega_s \omega_{s'} \mathcal{N}(o_1 | \boldsymbol{\mu}_s, \sigma_s^2 I) \mathcal{N}(o_2 | \boldsymbol{\mu}_{s'}, \sigma_{s'}^2 I)}{\sum_s \sum_{s'} \omega_s \omega_{s'} \mathcal{N}(o_1 | \boldsymbol{\mu}_s, \sigma_s^2 I) \mathcal{N}(o_2 | \boldsymbol{\mu}_{s'}, \sigma_{s'}^2 I)}$$

$$= \frac{\omega_{s'} \mathcal{N}(o_2 | \boldsymbol{\mu}_{s'}, \sigma_{s'}^2 I) \left(\sum_s \omega_s \mathcal{N}(o_1 | \boldsymbol{\mu}_s, \sigma_s^2 I) \right)}{\left(\sum_s \omega_s \mathcal{N}(o_1 | \boldsymbol{\mu}_s, \sigma_s^2 I) \right) \left(\sum_{s'} \omega_{s'} \mathcal{N}(o_2 | \boldsymbol{\mu}_{s'}, \sigma_{s'}^2 I) \right)}$$

$$= \frac{\omega_{s'} \mathcal{N}(o_2 | \boldsymbol{\mu}_{s'}, \sigma_{s'}^2 I)}{\sum_{s'} \omega_{s'} \mathcal{N}(o_2 | \boldsymbol{\mu}_{s'}, \sigma_{s'}^2 I)}$$

The posterior does not involve O_1 and X_1 and has the same form as that of a GMM that we have seen in Problem 2.2. **(2 points)**