Machine Learning Homework 4

General Instructions

Homework must be submitted electronically on Canvas. Make sure to explain your reasoning or show your derivations. Except for answers that are especially straightforward, you will lose points for unjustified answers, even if they are correct.

General Instructions

You are allowed to work with at most one other student on the homework. With your partner, you will submit only one copy, and you will share the grade that your submission receives. You should set up your partnership on Canvas as a two-person group by joining one of the preset groups named "HW4 Group n" for some number n.

Submit your homework electronically on Canvas. We recommend using LaTeX, especially for the written problems. But you are welcome to use anything as long as it is neat and readable.

For the programming portion, you should only need to modify the Python files. You may modify the iPython notebooks, but you will not be submitting them, so your code must work with our provided iPython notebook.

Relatedly, cite all outside sources of information and ideas.

Written Problems

1. We will derive the expectation-maximization (EM) algorithm using *variational* analysis.

   Let $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ be a set of data vectors. Let $Z = \{z_1, \ldots, z_n\}$ be set of multinomial variables corresponding to which of K Gaussians generated each example. Let the Gaussian parameters be means $\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$ and covariance matrices $\{\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K\}$. Let $\Theta = \{\theta_1, \ldots, \theta_K\}$ be multinomial prior probabilities of which Gaussian generates each example.

   Each data point is generated by first sampling a Gaussian index from $p(z|\Theta)$, then sampling from the Gaussian $\mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$. The log likelihood of any observations given these mixture model parameters is

   $$L(X, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \Theta) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} p(z_i|\Theta)\, \mathcal{N}(x_i|\boldsymbol{\mu}_k, \Sigma_k) \right) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \theta_k\, \mathcal{N}(x_i|\boldsymbol{\mu}_k, \Sigma_k) \right) . \tag{1}$$

   Since we will never observe any z variables, these are considered *hidden* or *latent* variables. When computing the likelihood of observed variables, we sum over all possible states of the latent variables, weighted by the probability of those states.

   We start by doing something a little weird. We create an independent distribution q for the latent variables such that

   $$q(z_1, \ldots, z_n) := q(z_1)q(z_2) \ldots q(z_n). \tag{2}$$

   We then rewrite the log likelihood from Equation (1) so that each data point's likelihood is multiplied by the q distribution divided by itself. In other words, we multiply the terms in the innermost summation by $\frac{q(z_i=k)}{q(z_i=k)}$, resulting in the equivalent form of the likelihood:

   $$L(X, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \Theta, q) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \theta_k\, \mathcal{N}(x_i|\boldsymbol{\mu}_k, \Sigma_k)\, \underbrace{q(z_i = k)/q(z_i = k)}_{1} \right) . \tag{3}$$

Jensen's inequality guarantees that for any convex function φ and any distribution over random variable X,

$$\varphi\left(\mathbb{E}\left[X\right]\right) \leq \mathbb{E}\left[\varphi(X)\right] \ . \tag{4}$$

We use Jensen's inequality and the fact that $\log$ is a **concave** function (i.e., $-\log$ is a convex function) to form a lower bound on the log likelihood:

$$L(X, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \Theta, q) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \theta_k \, \mathcal{N}(x_i | \boldsymbol{\mu}_k, \Sigma_k) q(z_i = k) / q(z_i = k) \right) \tag{5}$$

$$\geq \sum_{i=1}^{n} \sum_{k=1}^{K} q(z_i = k) \log \left(\theta_k \, \mathcal{N}(x_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) / q(z_i = k) \right) \tag{6}$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} q(z_i = k) \log \left(\theta_k \, \mathcal{N}(x_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) - \sum_{i=1}^{n} \sum_{k=1}^{K} q(z_i = k) \log q(z_i = k). \tag{7}$$

(a) (5 points) You now have a lower bound objective function that depends on the Gaussian mixture model parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and Θ and the variational distribution q. If we hold q fixed, maximizing the Gaussian mixture model parameters gets us the Gaussian EM updates (the so-called "m-step"). You will prove this fact for one of the mixture parameters. (You are welcome to try the other parameters too for fun.) Show that maximizing the lower bound with respect to the cluster mixture probabilities Θ is exactly the EM update for these variables.

You will need to use a Lagrange multiplier to enforce that $1 = \sum_{k=1}^{K} \theta_k$. Then solve for the settings of the Lagrange multiplier and each parameter θ_k to find the optimum of the constrained optimization. In other words, if the Lagrange multiplier is ζ and the Lagrangian form of the objective function is $\tilde{L}$, then you can find the solution by solving the following equations:

$$\frac{\partial \, \tilde{L}}{\partial \, \theta_k} = 0, \quad \text{and} \quad \frac{\partial \, \tilde{L}}{\partial \, \zeta} = 0. \tag{8}$$

Solve each equation in order and plug in the result into the original Lagrangian objective. Each zero-derivative condition will tell you something about the original objective that allows you to simplify it. You should end up with a final formula for the optimal value of θ_k that is relatively compact; Terms should simplify significantly to result in a simple final expression.

Solution: Dropping all terms that aren't affected by Θ and adding a Lagrange penalty, we have the objective function

$$\tilde{L} = \sum_{k=1}^{K} \log \theta_k \sum_{i=1}^{n} q(z_i = k) + \zeta \left(1 - \sum_{k=1}^{K} \theta_k \right).$$

Taking the derivative with respect to θ_k and setting to zero, we have

$$\partial \tilde{L} / \partial \theta_k = \frac{1}{\theta_k} \sum_{i=1}^{n} q(z_i = k) - \zeta = 0$$

$$\theta_k = \frac{1}{\zeta} \sum_{i=1}^{n} q(z_i = k)$$

Plugging this back in, we get

$$\tilde{L} = \sum_{k=1}^{K} \left(\log \left(\sum_{i=1}^{n} q(z_i = k) \right) - \log \zeta \right) \sum_{i=1}^{n} q(z_i = k) + \zeta - \sum_{k=1}^{K} \sum_{i=1}^{n} q(z_i = k).$$

Taking derivatives with respect to ζ and setting to zero, we get

$$\partial \tilde{L}/\partial \zeta = -\frac{1}{\zeta}\sum_{k=1}^{K}\sum_{i=1}^{n} q(z_i = k) + 1 = 0\ ,$$
$$\zeta = \sum_{k=1}^{K}\sum_{i=1}^{n} q(z_i = k) = n\ .$$

So $\theta_k = \frac{1}{n}\sum_{i=1}^{n} q(z_i = k)$.

Alternatively, we can just directly take the partial derivative of the original Lagrangian $\tilde{L}$ for ζ and set that to zero.

$$0 = 1 - \sum_{k=1}^{K} \theta_k.$$

Plugging in the formula for θ_k from its zero-gradient condition, we get

$$\begin{aligned}
0 &= 1 - \sum_{k=1}^{K}\frac{1}{\zeta}\sum_{i=1}^{n} q(z_i = k) \\
1 &= \frac{1}{\zeta}\sum_{i=1}^{n}\sum_{k=1}^{K} q(z_i = k)1 \qquad = \frac{1}{\zeta}\sum_{i=1}^{n} 1 \\
1 &= \frac{n}{\zeta} \\
\zeta &= n.
\end{aligned}$$

In the third step, we use the fact that valid q distributions should sum to 1. Plugging this back into the zero-gradient condition for θ_k again gets us to the update formula.

(b) (5 points) This second part is a bit more involved, but follows a similar line of reasoning. Show how to find the q parameters for each data point that maximize the lower bound.

You'll need to use Lagrange multipliers to enforce that $1 = \sum_k q(z_i = k)$ for each i, and you should be able to consider each data point's q distribution independently. Then solve for the settings of the Lagrange multiplier and each parameter $q(z_i = k)$ to find the optimum of the constrained optimization. In other words, if the Lagrange multiplier for the i'th variable is ζ_i and the Lagrangian form of the objective function is $\tilde{L}$, then you can find the solution by solving the following equations:

$$\frac{\partial\, \tilde{L}}{\partial\, q(z_i = k)} = 0, \quad \text{and} \quad \frac{\partial\, \tilde{L}}{\partial\, \zeta_i} = 0, \tag{9}$$

where we abuse notation to refer to the multinomial probability $q(z_i = k)$ as a variable.

You should again end up with a final formula for the optimal value of $q(z_i = k)$ that is relatively compact; Terms should simplify significantly to result in a simple final expression.

Solution: First, we write a Lagrangian with the simplex constraint:

$$\begin{aligned}
\tilde{L} = &\sum_{i=1}^{n}\sum_{k=1}^{K} q(z_i = k)\log\left(\theta_k\, \mathcal{N}(x_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) - \sum_{i=1}^{n}\sum_{k=1}^{K} q(z_i = k)\log q(z_i = k) \\
&+ \sum_{i=1}^{n}\zeta_i\left(1 - \sum_{k=1}^{K} q(z_i = k)\right)\ .
\end{aligned}$$

Taking the derivative with respect to $q(z_i = k)$, we get

$$\begin{aligned}
\partial \tilde{L}/\partial q(z_i = k) &= \log\left(\theta_k\, \mathcal{N}(x_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) - \log q(z_i = k) - 1 - \zeta_i = 0\ , \\
\log q(z_i = k) &= \log\left(\theta_k\, \mathcal{N}(x_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) - 1 - \zeta_i\ .
\end{aligned}$$

Plugging this back in to the $\log q$ term, we get

$$\begin{aligned}
\tilde{L} =& \sum_{i=1}^{n}\sum_{k=1}^{K} q(z_i = k) \log\left(\theta_k \mathcal{N}(x_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) - \sum_{i=1}^{n}\sum_{k=1}^{K} q(z_i = k)(\log\left(\theta_k \mathcal{N}(x_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) - 1 - \zeta_i) \\
&+ \sum_{i=1}^{n} \zeta_i \left(1 - \sum_{k=1}^{K} q(z_i = k)\right) \\
=& \sum_{i=1}^{n}\sum_{k=1}^{K} q(z_i = k) \log\left(\theta_k \mathcal{N}(x_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) - \sum_{i=1}^{n}\sum_{k=1}^{K} q(z_i = k) \log\left(\theta_k \mathcal{N}(x_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) \\
&+ \sum_{i=1}^{n}\sum_{k=1}^{K} q(z_i = k)(1 + \zeta_i) + \sum_{i=1}^{n} \zeta_i \left(1 - \sum_{k=1}^{K} q(z_i = k)\right) \\
=& \sum_{i=1}^{n}\sum_{k=1}^{K} q(z_i = k) + \sum_{i=1}^{n} \zeta_i \sum_{k=1}^{K} q(x_i = k) + \sum_{i=1}^{n} \zeta_i - \sum_{i=1}^{n} \zeta_i \sum_{k=1}^{K} q(z_i = k) \\
=& \sum_{i=1}^{n}\sum_{k=1}^{K} q(z_i = k) + \sum_{i=1}^{n} \zeta_i \,.
\end{aligned}$$

Now that we've canceled a lot of big terms out, plugging back in for the last piece, we get

$$\tilde{L} = \sum_{i=1}^{n}\sum_{k=1}^{K} \exp\left(\log\left(\theta_k \mathcal{N}(x_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) - 1 - \zeta_i\right) + \sum_{i=1}^{n} \zeta_i \,.$$

Taking the derivative and setting to zero, we get

$$\begin{aligned}
\partial\tilde{L}/\partial\zeta_i =& -\sum_{k=1}^{K} \exp\left(\log\left(\theta_k \mathcal{N}(x_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) - 1 - \zeta_i\right) + 1 = 0 \,, \\
1 =& \exp(-\zeta_i) \sum_{k=1}^{K} \exp\left(\log\left(\theta_k \mathcal{N}(x_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) - 1\right) \,, \\
\exp(\zeta_i) =& \sum_{k=1}^{K} \exp\left(\log\left(\theta_k \mathcal{N}(x_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) - 1\right) \,, \\
\zeta_i =& \log\left(\sum_{k=1}^{K} \exp\left(\log\left(\theta_k \mathcal{N}(x_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) - 1\right)\right) \,.
\end{aligned}$$

Finally, plugging this back into the formula for q, we get

$$\begin{aligned}
q(z_i = k) &= \exp(\log\left(\theta_k \mathcal{N}(x_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) - 1 - \zeta_i) \\
&= \exp\left(\log\left(\theta_k \mathcal{N}(x_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) - 1 - \log\left(\sum_{k'=1}^{k'} \exp\left(\log\left(\theta_{k'} \mathcal{N}(x_i|\boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'})\right) - 1\right)\right)\right) \\
&= \frac{\exp\left(\log\left(\theta_k \mathcal{N}(x_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) - 1\right)}{\sum_{k'=1}^{k'} \exp\left(\log\left(\theta_{k'} \mathcal{N}(x_i|\boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'})\right) - 1\right)} \\
&= \frac{\exp\left(\log\left(\theta_k \mathcal{N}(x_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)\right) \exp\left(-1\right)}{\sum_{k'=1}^{k'} \exp\left(\log\left(\theta_{k'} \mathcal{N}(x_i|\boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'})\right)\right) \exp\left(-1\right)} \\
&= \frac{\theta_k \mathcal{N}(x_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k'=1}^{k'} \theta_{k'} \mathcal{N}(x_i|\boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'})} \,.
\end{aligned}$$

Alternatively, we can again directly solve for the zero-gradient condition of the Lagrange multipliers. First, recall that we found that

$$\log q(z_i = k) = \log\left(\theta_k \mathcal{N}(x_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) - 1 - \zeta_i$$

Setting derivative to zero for each Lagrange multiplier gives us

$$\sum_{k=1}^{K} q(z_i = k) = 1.$$

Exponentiating and plugging in the formula for q, we get

$$\sum_{k=1}^{K} \exp\left(\log\left(\theta_k \mathcal{N}(x_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) - 1 - \zeta_i\right) = 1$$

$$\sum_{k=1}^{K} \exp\left(\log\left(\theta_k \mathcal{N}(x_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) - 1\right)\exp(-\zeta_i) = 1.$$

$$\frac{1}{\exp(\zeta_i)} \sum_{k=1}^{K} \exp\left(\log\left(\theta_k \mathcal{N}(x_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)\right) - 1) = 1.$$

$$\sum_{k=1}^{K} \exp\left(\log\left(\theta_k \mathcal{N}(x_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)\right) - 1) = \exp(\zeta_i).$$

Plugging this back into the formula for q, again we can simplify as on the last page.

2. (5 points) Project proposal. See instructions on project homepage.