University of Illinois Fall 2018

ECE 544NA: Pattern Recognition
Lecture 18: October 25

Lecturer: Alexander Schwing Scribe: Wei Ren

1 Introduction

One of the goals in unsupervised learning is to learn (possibly hidden) pattern in the given data without any explicit label. As we already studied the previous two lectures, the k-means and Gaussian Mixture Model (GMM) algorithms are two common ways to cluster data in an unsupervised way by introducing latent variables.

In this lecture we focus on the essence of the previous two algorithms, namely to iterate on expectation and maximization steps. This alternation of expectation and optimization (a.k.a. expectation maximization) will lead us to a more general framework for a set of clustering algorithms. Specifically we generalizes the k-means and GMM methods into the problem of finding empirical lower bound. We prove that Expectation Maximization (EM) procedure correctly compute the empirical lower bound, which is also identical to what the Concave-convex procedure (CCCP) would produce.

Objectives:

- Generalization of expectation maximization (EM)
- Derive and compute the empirical lower bound
- Similarity between the Concave-convex procedure (CCCP)

2 Background

2.1 Recap on Gaussian Mixture Model (GMM)

Recall in the previous lecture, the cost function for GMM is to minimize the log-likelihood over the entire dataset:

$$\min_{\pi,\mu,\sigma} -\log \prod_{i\in\mathcal{D}} p\left(x^{(i)} \mid \pi,\mu,\sigma\right) \tag{1}$$

$$= \min_{\pi,\mu,\sigma} -\sum_{i\in\mathcal{D}} \log \sum_{k=1}^{K} \pi_k \mathcal{N}\left(x^{(i)} \mid \mu_k,\sigma_k\right) \quad \text{s.t.} \quad \sum_{k=1}^{K}\pi_k = 1, \quad \pi_k \geq 0 \tag{2}$$

where π_k, μ_k and σ_k denotes the weight, mean and standard deviation of k^{th} Gaussian distribution, respectively. Given an initial cluster assignment, the GMM algorithm then calculate/estimate the posterior $r_{ik} = p\left(z_{ik} \mid x^{(i)}\right) = \frac{\pi_k \mathcal{N}\left(x^{(i)}|\mu_k,\sigma_k\right)}{\sum_{\hat{k}} \pi_{\hat{k}} \mathcal{N}\left(x^{(i)}|\mu_{\hat{k}},\sigma_{\hat{k}}\right)}$ in each E-step (expectation), and optimize the parameters with respect to the calculated posterior r_{ik} in M-step (maximization). Update in M-step is in closed form, which makes GMM easy to implement and compute.

Here we do not make a separate case for k-means algorithm because it is a special case of GMM when the all the variances σ_k^2 above reach their limits to zero. It is sufficient to use equation 2 for generalization.

During the formulation of posterior r_{ik}, we introduced the latent variable z_{ik}, where $\forall i \in \mathcal{D}, z_{ik} \in \{0, 1\}$ with $\forall i, \sum_{k=1}^{K} z_{ik} = 1$. With the help of latent variables, we can compute (sometimes hard-to-compute) marginal probabilities of observed variables using (relatively easy-to-compute) joint probability distributions of both observed and latent variables. We can rewrite equation 2 as the following:

$$\min_{\pi,\mu,\sigma} -\sum_{i\in\mathcal{D}} \log \sum_{\mathbf{z}_i} p\left(x^{(i)} \mid \mathbf{z}_i\right) p\left(\mathbf{z}_i\right) \tag{3}$$

$$= \min_{\pi,\mu,\sigma} -\sum_{i\in\mathcal{D}} \log \sum_{\mathbf{z}_i} p\left(x^{(i)}, \mathbf{z}_i\right) \tag{4}$$

where $\mathbf{z}_i = \begin{bmatrix} z_{i1} & \cdots & z_{iK} \end{bmatrix}^\intercal$.

2.2 The Problem in General

GMM assumes the distribution in the model to be normal distributions. But the general problem need not to be composition of normal distributions. More generally, the summation of the joint distribution over latent variable generate the marginal probability of observed variable (i.e., $x^{(i)}$). For simplicity, we will drop the summation over the dataset for now and only consider the minimization problem for a single data point.

$$p_\theta\left(x^{(i)}\right) = \sum_{\mathbf{z}_i} p_\theta\left(x^{(i)} \mid \mathbf{z}_i\right) p_\theta\left(\mathbf{z}_i\right) = \sum_{\mathbf{z}_i} p_\theta\left(x^{(i)}, \mathbf{z}_i\right) \tag{5}$$

Then the negative log-likelihood of a single data point is:

$$-\log p_\theta\left(x^{(i)}\right) = -\log \sum_{\mathbf{z}_i} p_\theta\left(x^{(i)}, \mathbf{z}_i\right) \tag{6}$$

where θ denotes the parameters of the probability distribution. This is the almost the same as the setup in GMM (equation 4) except the summation over the entire dataset. Minimizing the negative log-likelihood is equivalent to maximizing the log-likelihood. Thus, we can consider maximizing the following problem in general:

$$\log p_\theta\left(x^{(i)}\right) = \log \sum_{\mathbf{z}_i} p_\theta\left(x^{(i)}, \mathbf{z}_i\right) \tag{7}$$

3 Expectation Maximization (EM)

To maximize the general problem in equation 7 we have two options:

- Empirical Lower Bound
- Concave-Convex Procedure/Majorize-Minimize

They may appear to be quite different on the surface, but it turns out that both are identical. We will start with empirical lower bound first.

3.1 Empirical Lower Bound

We can introduce another new distribution $q(\mathbf{z})$ over latent variable $\mathbf{z}$ (we will have different $\mathbf{z}_i$ for different $x^{(i)}$ in the final setup, but for now we drop the subscript). Then equation 7 can be rewritten as:

$$\log p_\theta\left(x^{(i)}\right) = \mathcal{L}\left(p_\theta\left(x^{(i)}, \mathbf{z}\right), q(\mathbf{z})\right) + D_{KL}\left(q(\mathbf{z}), p_\theta\left(\mathbf{z} \mid x^{(i)}\right)\right) \tag{8}$$

where the lower bound $\mathcal{L}$ and Kullback-Leibler divergence (KL divergence) D_{KL} are defined as:

$$\mathcal{L}\left(p_\theta\left(x^{(i)}, \mathbf{z}\right), q(\mathbf{z})\right) = \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p_\theta\left(x^{(i)}, \mathbf{z}\right)}{q(\mathbf{z})} \tag{9}$$

$$D_{KL}\left(q(\mathbf{z}), p_\theta\left(\mathbf{z} \mid x^{(i)}\right)\right) = \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{q(\mathbf{z})}{p_\theta\left(\mathbf{z} \mid x^{(i)}\right)} \tag{10}$$

After we prove the correctness of equation 8 in section 3.1.1, we will introduce Kullback-Leibler divergence and some of its properties, which will become quite useful in section 3.1.5 when we derive empirical lower bound.

As a side note, we could not express equation 9 into the form of KL divergence. This is because variable space of two distributions x and z in equation 9 are different whereas the two distributions in KL divergence must have the same variable space. Otherwise, it cannot be defined using KL divergence.

3.1.1 Proof of Equation 8

The right hand side of equation 8 can be written from the definitions:

$$\mathcal{L}\left(p_\theta\left(x^{(i)}, \mathbf{z}\right), q(\mathbf{z})\right) + D_{KL}\left(q(\mathbf{z}), p_\theta\left(\mathbf{z} \mid x^{(i)}\right)\right) \tag{11}$$

$$= \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p_\theta\left(x^{(i)}, \mathbf{z}\right)}{q(\mathbf{z})} + \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{q(\mathbf{z})}{p_\theta\left(\mathbf{z} \mid x^{(i)}\right)} \tag{12}$$

$$= \sum_{\mathbf{z}} q(\mathbf{z}) \log \left(\frac{p_\theta\left(x^{(i)}, \mathbf{z}\right)}{q(\mathbf{z})} \times \frac{q(\mathbf{z})}{p_\theta\left(\mathbf{z} \mid x^{(i)}\right)}\right) \tag{13}$$

$$= \sum_{\mathbf{z}} q(\mathbf{z}) \log \left(\frac{p_\theta\left(x^{(i)}, \mathbf{z}\right)}{p_\theta\left(\mathbf{z} \mid x^{(i)}\right)}\right) \tag{14}$$

$$= \sum_{\mathbf{z}} q(\mathbf{z}) \log \left(\frac{p_\theta\left(\mathbf{z} \mid x^{(i)}\right) p_\theta\left(x^{(i)}\right)}{p_\theta\left(\mathbf{z} \mid x^{(i)}\right)}\right) \tag{15}$$

$$= \sum_{\mathbf{z}} q(\mathbf{z}) \log \left(p_\theta\left(x^{(i)}\right)\right) = \log \left(p_\theta\left(x^{(i)}\right)\right) \sum_{\mathbf{z}} q(\mathbf{z}) \tag{16}$$

$$= \log \left(p_\theta\left(x^{(i)}\right)\right) \tag{17}$$

Equation 17 is the same as left hand side of equation 8. Consequently, equation 8 holds true. The original problem is to minimize negative log-likelihood, which is the same as maximizing log-likelihood.

3.1.2 Jensen's Inequality

Jensen's inequality states that for a **convex** function f and random variable x, the follow inequality always hold:

$$f\left(E\left[x\right]\right) \leq E\left[f\left(x\right)\right] \tag{18}$$

Conversely, for a **concave** function f:

$$f\left(E\left[x\right]\right) \geq E\left[f\left(x\right)\right] \tag{19}$$

3.1.3 Kullback-Leibler Divergence

Kullback-Leibler divergence (KL divergence) defines the difference between two distributions over the same variable or variable space. Assume distribution p and q are two distributions over discrete variable x, then KL divergence is defined as:

$$D_{KL}\left(p,q\right) = E_{p(x)}\left[\log p\left(x\right) - \log q\left(x\right)\right] = \sum_{x} p\left(x\right) \frac{p\left(x\right)}{q\left(x\right)} \tag{20}$$

If the two distribution are the original distribution and approximation distribution, then KL divergence tells us how far away is the two using the expectation of log difference (see chapter 2.6 in [2] for more detailed information).

3.1.4 KL Divergence Property

One of the important properties of KL divergence in equation 8 is that D_{KL} is always non-negative, namely,

$$D_{KL}\left(p,q\right) \geq 0 \tag{21}$$

Proof: Since log function is concave, we can apply Jensen's inequality to equation 10.

$$D_{KL}\left(q\left(\mathbf{z}\right), p_\theta\left(\mathbf{z} \mid x^{(i)}\right)\right) = \sum_{\mathbf{z}} q\left(\mathbf{z}\right) \log \frac{q\left(\mathbf{z}\right)}{p_\theta\left(\mathbf{z} \mid x^{(i)}\right)} \tag{22}$$

$$\geq \log \sum_{\mathbf{z}} q\left(\mathbf{z}\right) \frac{q\left(\mathbf{z}\right)}{p_\theta\left(\mathbf{z} \mid x^{(i)}\right)} \tag{23}$$

Equation 23 can be rewritten as:

$$\log \sum_{\mathbf{z}} q\left(\mathbf{z}\right) \frac{q\left(\mathbf{z}\right)}{p_\theta\left(\mathbf{z} \mid x^{(i)}\right)} = -\log \sum_{\mathbf{z}} q\left(\mathbf{z}\right) \frac{p_\theta\left(\mathbf{z} \mid x^{(i)}\right)}{q\left(\mathbf{z}\right)} \tag{24}$$

$$= -\log \sum_{\mathbf{z}} p_\theta\left(\mathbf{z} \mid x^{(i)}\right) \tag{25}$$

$$= 0 \tag{26}$$

Thus, we see that $D_{KL} \geq 0$ always holds true.

This non-negativity is also known as Gibbs' inequality. Zero KL divergence only holds when $p = q$ (i.e., when the two distributions are identical).

3.1.5 Maximize Empirical Lower Bound

Section 3.1.1 proves that equation 8 holds true for any distribution $q(z)$.

$$\max_{\theta} \log p_\theta \left(x^{(i)}\right) = \max_{\theta,q} \mathcal{L}\left(p_\theta\left(x^{(i)}, \mathbf{z}\right), q\left(\mathbf{z}\right)\right) + D_{KL}\left(q\left(\mathbf{z}\right), p_\theta\left(\mathbf{z} \mid x^{(i)}\right)\right) \tag{27}$$

Equation 27 becomes our new optimization problem. From section 3.1.4 we have $D_{KL} \geq 0$, thus:

$$\log p_\theta \left(x^{(i)}\right) \geq \mathcal{L}\left(p_\theta\left(x^{(i)}, \mathbf{z}\right), q\left(\mathbf{z}\right)\right) \tag{28}$$

Instead of maximizing the log-likelihood, we can maximize $\mathcal{L}$ which is known as the empirical lower bound. The problem then becomes:

$$\max_{\theta,q} \mathcal{L}\left(p_\theta\left(x^{(i)}, \mathbf{z}\right), q\left(\mathbf{z}\right)\right) \tag{29}$$

We have successfully transformed the original problem into a new maximization problem. But the following question still remains.

Question: How can we optimize the new maximization problem as defined in equation 29?

Answer: By alternating optimization w.r.t q and θ (the same idea in GMM and k-means).

Figure 1 (taken from chapter 9 of [1]) shows one successive update in EM algorithm. The red curve represents the log-likelihood of the given data x. The blue curve depicts an intermediate empirical lower bound $\mathcal{L}_i$, which is produced when performing the E-step, using the old parameters θ_{old}. Then we would perform the M-step using the maximum value of $\mathcal{L}_i$ to get a newly generated θ_{new}. The next step is to perform E-step again to get the next (and better) lower bound $\mathcal{L}_{i+1}$, the green curve in the graph. This alternating process will continue until convergence. Note that every curve is tangential to the original input data curve.

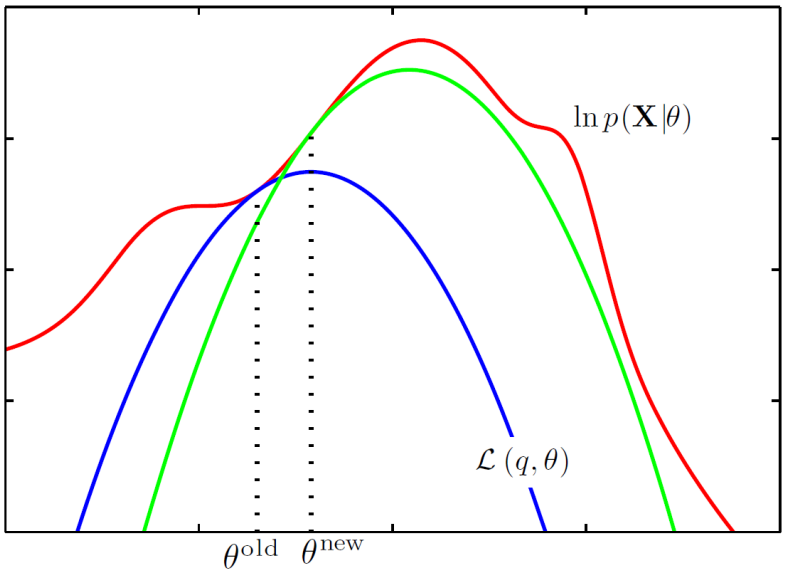

Figure 1: Illustration of EM optimization. The red curve is the input data curve (in the form of maximum log-likelihood). Blue curve is one of the empirical lower bound and the green curve is the empirical lower bound after the blue one.

3.1.6 Optimization with respect to q

How can we optimize empirical lower bound $\mathcal{L}$ with respect to a distribution $q(\mathbf{z})$?

Our first guess is to set $q(\mathbf{z}) = p_\theta\left(\mathbf{z} \mid x^{(i)}\right)$. The following reasoning is intuitive:

- Maximize with respect to q Since distribution q is can be arbitrary, we can make set it to any function that we need. From section 3.1.4 we know that D_{KL} is zero only when the two distribution are exactly identical to each other. Thus, When $q(\mathbf{z})$ is the same as $p_\theta\left(\mathbf{z} \mid x^{(i)}\right)$, $D_{KL} = 0$ makes the empirical lower bound in equation 29 the same as the original problem in equation 8.

But the reasoning of $q(\mathbf{z}) = p_\theta\left(\mathbf{z} \mid x^{(i)}\right)$ is based on the idea that we should pick distribution that yields the tightest empirical lower bound (chapter 11 in [3]).

An alternative way to derive is to use the Lagrangian formulation:

$$\mathcal{L}\left(p_\theta\left(x^{(i)}, \mathbf{z}\right), q(\mathbf{z})\right) = \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p_\theta\left(x^{(i)}, \mathbf{z}\right)}{q(\mathbf{z})} \tag{30}$$

$$= \sum_{\mathbf{z}} q(\mathbf{z}) \log p_\theta\left(x^{(i)}, \mathbf{z}\right) - \sum_{\mathbf{z}} q(\mathbf{z}) \log q(\mathbf{z}) \tag{31}$$

$$= \sum_{\mathbf{z}} q(\mathbf{z}) \log p_\theta\left(x^{(i)}, \mathbf{z}\right) + H(q(\mathbf{z})) \tag{32}$$

where entropy $H(q(x)) = -\sum q(x) \log q(x)$.

Next we write the standard form with constraints on $q(\mathbf{z})$,

$$\max_q \sum_{\mathbf{z}} q(\mathbf{z}) \log p_\theta\left(x^{(i)}, \mathbf{z}\right) + H(q(\mathbf{z})) \quad \text{s.t.} \quad -q(\mathbf{z}) \leq 0, \quad 1 - \sum_{\mathbf{z}} q(\mathbf{z}) = 0 \tag{33}$$

By finding the stationary point of Lagrangian, we can find the optimum occurs when

$$q(\mathbf{z}) = \frac{p_\theta\left(x^{(i)}, \mathbf{z}\right)}{\sum_{\mathbf{z}} p_\theta\left(x^{(i)}, \mathbf{z}\right)} = p_\theta\left(\mathbf{z} \mid x^{(i)}\right) = r_i \tag{34}$$

3.1.7 Empirical Lower Bound in Gaussian Case

The empirical lower bound $\mathcal{L}$ in the Gaussian case is (note the change of summation indices in the last few steps):

$$\mathcal{L}\left(p_\theta\left(x^{(i)}, \mathbf{z}\right), q(\mathbf{z})\right) = \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p_\theta\left(x^{(i)}, \mathbf{z}\right)}{q(\mathbf{z})} \tag{35}$$

$$= \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{\prod_{k=1}^{K} \pi_k^{z_{ik}} \mathcal{N}\left(x^{(i)} \mid \mu_k, \sigma_k\right)^{z_{ik}}}{q(\mathbf{z})} \tag{36}$$

$$= \sum_{\mathbf{z},k} q(\mathbf{z}) \log \left(\pi_k^{z_{ik}} \mathcal{N}\left(x^{(i)} \mid \mu_k, \sigma_k\right)^{z_{ik}}\right) + H(q(\mathbf{z})) \tag{37}$$

$$= \sum_{k} r_{ik} \log \left(\pi_k \mathcal{N}\left(x^{(i)} \mid \mu_k, \sigma_k\right)\right) - \sum_{k} r_{ik} \log r_{ik} \tag{38}$$

The first term recovers the cost function in GMM lecture.

3.1.8 Empirical Lower Bound in General Case

In general the joint distribution of observed and latent variables is:

$$p_\theta\left(x^{(i)}, \mathbf{z}\right) = \frac{1}{Z\left(\theta\right)} \exp\left(F\left(x^{(i)}, \mathbf{z}, \theta\right)\right) \tag{39}$$

where $Z\left(\theta\right)$ is the partition function (to ensure normalization of the probabilities).
Now we can calculate the negative of empirical lower bound $\mathcal{L}$:

$$-\mathcal{L}\left(p_\theta\left(x^{(i)}, \mathbf{z}\right), q\left(\mathbf{z}\right)\right) = -\sum_{\mathbf{z}} q\left(\mathbf{z}\right) \log \frac{p_\theta\left(x^{(i)}, \mathbf{z}\right)}{q\left(\mathbf{z}\right)} \tag{40}$$

$$= -\sum_{\mathbf{z}} q\left(\mathbf{z}\right) \log\left(\frac{1}{Z\left(\theta\right)} \frac{\exp\left(F\left(x^{(i)}, \mathbf{z}, \theta\right)\right)}{q\left(\mathbf{z}\right)}\right) \tag{41}$$

$$= \log\left(Z\left(\theta\right)\right) - \sum_{\mathbf{z}} q\left(\mathbf{z}\right) \exp\left(F\left(x^{(i)}, \mathbf{z}, \theta\right)\right) - H\left(q\left(\mathbf{z}\right)\right) \tag{42}$$

This is an important conclusion as we will see that Concave-Convex Procedure (CCCP) can be constructed into form, again, exactly identical to the above.

3.2 Concave-Convex Procedure (CCCP)

The Concave-Convex Procedure is a way to construct discrete time iterative dynamic systems which are guaranteed to monotonically decrease global optimization/energy functions. This procedure can be applied to almost any optimization problem and many existing algorithms can be interpreted in terms of it. In particular, all EM algorithms can be re-expressed in terms of CCCP.
CCCP contains several theories, summary of those are:

- Any function, subject to weak conditions, can be expressed as sum of a convex and concave part(this decomposition is not unique). This will imply that CCCP can be applied to almost any optimization problem.
- CCCP procedure converges to a minimum or a saddle point of the energy function.
- In many cases, it is possible to solve for x^{t+1} analytically. In other cases, we can re-express CCCP in terms of minimizing a time sequence of *convex update energy functions* $E_{t+1}(x^{t+1})$ to obtain the updates x^{t+1}.

3.2.1 Model

In the lecture, our model is :

$$p_\theta(x^{(i)}, \mathbf{z}) = \frac{1}{Z(\theta)} \exp F((x^{(i)}, \mathbf{z}), \theta) \tag{43}$$

To maximize the likelihood, rewrite the above function as,

$$\min_\theta - \ln \sum_{\mathbf{z}} \frac{\exp F((x^{(i)}, \mathbf{z}), \theta)}{Z(\theta)} \tag{44}$$

$$= \min_\theta \quad \ln Z(\theta) - \ln \sum_{\mathbf{z}} \exp F((x^{(i)}, \mathbf{z}), \theta) \tag{45}$$

The former part $\ln Z(\theta)$ is convex and the latter part $\ln \sum_{\mathbf{z}} \exp F((x^{(i)}, \mathbf{z}), \theta)$ is also convex. Note that the negative of the second term (convex) result in a concave function.

3.2.2 Procedure

- Initialize θ
- Repeat:
  - Decompose the concave part into convex + concave part for current θ
    One possible way (by introducing a new distribution q):

$$\ln \sum_{\mathbf{z}} \exp F((x^{(i)}, \mathbf{z}), \theta) \tag{46}$$

$$= \ln \sum_{\mathbf{z}} q(\mathbf{z}) \frac{\exp F((x^{(i)}, \mathbf{z}), \theta)}{q(\mathbf{z})} \quad \text{(Using Jensens's Inequality)} \tag{47}$$

$$\leq \sum_{\mathbf{z}} q(\mathbf{z}) \ln \frac{\exp F((x^{(i)}, \mathbf{z}), \theta)}{q(\mathbf{z})} \tag{48}$$

$$\leq \sum_{\mathbf{z}} q(\mathbf{z}) \{\ln \exp F((x^{(i)}, \mathbf{z}), \theta) - \ln q(\mathbf{z})\} \tag{49}$$

$$= \max_{q(\mathbf{z})} \left(\sum_{\mathbf{z}} q(\mathbf{z}) F((x^{(i)}, \mathbf{z}), \theta) + H\left(q(\mathbf{z})\right) \right) \tag{50}$$

    By approximation of $q(z)$ and entropy H, we can approximate the original cost function. Also note that equation 50 is very similar to an inference problem. Section 3.2.3 gives a simple yet concrete example of this similarity.
    The final equation should be:

$$\min_{\theta, q} \ln Z(\theta) - \sum_{\mathbf{z}} q(\mathbf{z}) F((x^{(i)}, \mathbf{z}), \theta) - H(q(\mathbf{z})) \tag{51}$$

    This is the **same equation** as we derived in Empirical Lower Bound in equation 42.
  - Solve convex program

$$\min_{\theta} \quad \ln Z(\theta) - \ln \sum_{\mathbf{z}} \exp F((x^{(i)}, \mathbf{z}), \theta) \tag{52}$$

3.2.3 Similarity with Inference

A simple and intuitive example of maximizing expectation:

$$E(y) = \sum_{y} p(y) \times F(y)$$

where $p(y)$ is the distribution of variable y and $F(y)$ is the score function.

The score function as $F(y)$ depends on the input data y. Function F could be any function as we defined, for example, the number of input points at a position along the axes. Consider the following function F:

$$F(1) = 5$$
$$F(2) = 7$$
$$F(3) = 2$$

Ideally, the distribution of y should be a delta function so that $p(1) = 0, p(2) = 1, p(3) = 0$, which will produce the maximum expectation value $E(y) = 7$.

Note that this concept of maximizing expectation is actually doing what we've discussed in the inference problem.

4 Summary

In this lecture we studied general EM algorithm, which includes previously discussed k-means and GMM. By selecting the optimal distribution of latent variable, we are able to find the optimum by alternation of optimization. We also show that CCCP algorithm is no different from EM in general. To sum up, we finished discussing the following goals:

- Generalizing EM algorithm
- Getting to know EM's relationship with CCCP
- Seeing the variational form of the partition function
- Observing EM's similarity to inference

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[2] D. J. MacKay and D. J. Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[3] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.