**2nd Assignment**

**General Instructions:**

1. This is the second open book assignment.
2. You are free to consult any resource, including your friends. However, the latter may be in a situation similar to yours.
3. YOU HAVE TO WRITE THE SOLUTION IN YOUR OWN WORDS. Any form of copying will result in zero marks.
4. Write your solution as a word document. You can use the equation editor to write the equations.
5. Handwritten and scanned solutions will not be entertained.
6. **Submission deadline – 5.00 p.m. on 04/11/2020.**
7. **Upload on Moodle only. Please do not send your solution by mail.**
8. Read the assignment carefully, at least twice, right up to the last full stop. Make sure that you understand the problem completely. Then start thinking about the solution.

**The assignment:**

You have studied the Expectation Maximization algorithm in the class. You now have to apply it for the following problem.

The background: Humans can have certain specialized cells. These cells have similar functionality but have can belong to four different types (called **P\_Type)** namely – **X, Y, XY and Z**. One person can have only one type of cell. The **P\_Type** of a person can be observed directly. For example, if a scientist takes a sample from me and finds that I have **W** type then all other samples taken from me will show that I have **W** type. It is also known that the **P\_Type** actually depends on another property called **G\_Type**. However, **G\_Type** of a person can not be observed directly. The basic problem that you have to solve is that given samples from humans, we can observe their **P\_Type**. Thus, we can find a distribution for **P\_Type.** Using these observations, we want to infer something about the distribution of **G\_Type** using the Expectation Maximization method.

The following facts are known:

1. The **G\_Type** comes in three variants, **a, b** and **c**.
2. Each human will have two **G\_Type** cell. For example, I can have **a** and **c** type cells.
3. We can collect samples from various humans, measure their **P\_type** and find the frequencies of occurrence of **X, Y, XY and Z** to get **n(X), n(Y), n(XY)** and **n(Z)**.
4. Let the probability of occurrence of **G\_Type** be p(**a**), p(**b**) and p(**c**).
5. The connection between **G\_Type** and **P\_Type** is given by the following table (remember that each human will have two **G\_Type**  cell

|  |  |  |
| --- | --- | --- |
| **Underlying G\_Type** | Probability | **Observed P\_Type** |
| **aa** | p(a).p(a) | **X** |
| **ac** | 2.p(a).p(c) | **X** |
| **bb** | p(b).p(b) | **Y** |
| **bc** | 2.p(b).p(c) | **Y** |
| **cc** | p(c).p(c) | **Z** |
| **ab** | 2.p(a).p(b) | **XY** |

1. In the above, p(a), p(b) and p(c) are the probabilities of observing **G\_Type** of types **a, b** and **c** respectively.
2. As mentioned earlier, we can find the frequencies of occurrence of **X, Y, XY and Z** to get **n(X), n(Y), n(XY)** and **n(Z)**. These can be converted to probabilities **p(X), p(Y), p(XY)** and **p(Z).**
3. Using the table above we can write

p(**X**) = p(a)2 + 2 p(a) p(c)

p(**Y**) = p(b)2 + 2 p(b) p(c)

p(**Z**) = p(c)2

p(**XY**) = 2 p(a) p(b)

**The Assignment:**

Find expressions for p(a), p(b) and p(c) given **n(X), n(Y), n(XY)** and **n(Z).**

The expressions obtained in step 8 above express the probabilities **p(X), p(Y), p(XY)** and **p(Z)** in terms of the probabilities p(a), p(b) and p(c). In general, you will not be ably to simply invert these to get the desired expressions for p(a), p(b) and p(c) given the probabilities **p(X), p(Y), p(XY)** and **p(Z)**. You basically have three unknowns (i.e. p(a), p(b) and p(c)) and four equations. You will have to apply the Expectation Maximization method to solve this assignment. Thus, the expressions that you derive should be in terms of the EM method. You should write unambiguous expressions for the E step and the M step that pertains to this specific problem.

**Hint:**

Start with writing the likelihood function noting that **p(X), p(Y), p(XY)** and **p(Z)** are multinomial distributions (much like the probabilities used for the loaded dice). Treat the **G\_Type** as some hidden (latent / underlying) process. Express **p(X), p(Y), p(XY)** and **p(Z)** in terms of these latent processes. You may use the formulation of the two dice problem. Now apply the EM method for writing the final expressions for the E step and the M step.

Further note that p(c) = 1 – p(a) – p(b)

Thus, if you can find expressions for p(a) and p(b) then finding p(c) is easy.

*Students who submit only the general EM method (or some algorithm) will not be awarded any marks. Only those submissions that address the specific problem given above will be considered for evaluation.*