Lehrstuhl für STEUERUNGS- UND REGELUNGSTECHNIK Technische Universität München Prof. Dongheui Lee	MACHINE LEARNING IN ROBOTICS Exercises 3: Maximum Likelihood & EM

Exercise 1

Show that the Maximum a posteriori (MAP) estimate becomes Maximum likelihood (ML) estimate if we assume uniform prior distribution for the parameters $\boldsymbol{\theta}$.

Solution:

The MAP estimate is given by

$\arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(n)}) = \arg\max_{\boldsymbol{\theta}} \frac{p(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(n)}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(n)})}$

Since denominator is independent of $\boldsymbol{\theta}$, it can be ignored.

$\arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(n)}) \propto \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(n)}|\boldsymbol{\theta})p(\boldsymbol{\theta})$

Now if the prior $p(\boldsymbol{\theta})$ is uniformly distributed then it can be replaced with a constant and the maximization is reduced to

$\arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(n)}) \propto \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(n)}|\boldsymbol{\theta})$

Now if the samples are independent and identically distributed then

$\arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(n)}) \propto \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{x}^{(1)}|\boldsymbol{\theta})p(\boldsymbol{x}^{(2)}|\boldsymbol{\theta}) \ldots p(\boldsymbol{x}^{(n)}|\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{n} p(\boldsymbol{x}^{(i)}|\boldsymbol{\theta})$

which is the well known maximum likelihood approach for parameters estimation. Thus the Bayesian estimator coincides with the maximum-likelihood estimator for a uniform prior distribution on the parameters.

Exercise 2

The Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time. The probability mass function is given by :

$$p(k|\mu) = \frac{\mu^k e^{-\mu}}{k!}$$

where $\mu > 0$. Let $\boldsymbol{X} = x^{(1)}, x^{(2)}, \ldots, x^{(n)}$ be i.i.d. poisson random variables. Use the samples to get a maximum likelihood estimate of μ.

Solution:

$L(\mu) = \prod_{i=1}^{n} \frac{e^{-\mu}\mu^{x^{(i)}}}{x^{(i)}!} = \frac{e^{-n\mu}\mu^{\sum_{i=1}^{n} x^{(i)}}}{\prod_{i=1}^{n} x^{(i)}!}$

$l(\mu) = -n\mu + \sum_{i=1}^{n} x^{(i)} \log\mu - \log\prod_{i=1}^{n} x^{(i)}!$

Evaluating the gradient of $l(\mu)$ to zero we get

$-n + \frac{\sum_{i=1}^{n} x^{(i)}}{\mu} = 0$
$\mu = \frac{\sum_{i=1}^{n} x^{(i)}}{n}$
$\mu = \bar{x}$

Exercise 3

Consider a Gaussian mixture model in which the marginal distribution $p(\mathbf{z})$ for the latent variable is given by $p(\mathbf{z}) = \prod\limits_{k=1}^{K} \pi_k^{z_k}$ (where $z_k \in \{0,1\}$ and $\sum\limits_k z_k = 1$), and the conditional distribution $p(\mathbf{x}|\mathbf{z})$ for the observed variable is given by $p(\mathbf{x}|\mathbf{z}) = \prod\limits_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}$. Show that the marginal distribution $p(\mathbf{x})$, obtained by summing $p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$ over all possible values of $\mathbf{z}$, is a Gaussian mixture of the form $p(\mathbf{x}) = \sum\limits_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

Solution:

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}) = \sum_{\mathbf{z}} p(\mathbf{x}|\mathbf{z})p(\mathbf{z}) = \sum_{\mathbf{z}} \prod_{k=1}^{K} (\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))^{z_k}$$

Since z can only take 1 of K possible state, we can introduce an indicator variable I such that $I_{kj} = 1$ if $k = j$ and 0 otherwise. Now we can rewrite $p(\mathbf{x})$ as

$$p(\mathbf{x}) = \sum_{j=1}^{K} \prod_{k=1}^{K} (\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))^{I_{kj}} = \sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

Exercise 4

Robotic arms are widely used for conducting robotics research. We designed two algorithms $(A1, A2)$ for catching objects thrown towards a robotic arm. Algorithm 1 catches the objects with an unknown success rate of θ $(p_1 = P(Success|A1) = P(1|A1) = \theta)$ while Algorithm 2 has a 50 percent success rate $(p_2 = p(Success|A2) = p(1|A2) = 0.5)$. We ran the two algorithms several number of times and recorded their results (success $x = 1$ or failure $x = 0$). The algorithms were choosen randomly. Unfortunately after n experiments, we realize that we recorded only the results (success or failure) $\boldsymbol{X} = \{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$ without recording the identity ($A1$ or $A2$) of the algorithms $\boldsymbol{Z} = \{\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)}, \ldots, \boldsymbol{z}^{(n)}\}$ when performing experiments.

Since repeating experiments on real robot can be a costly and time consuming process, we are interested in estimating the success rate for Algorithm 1 with EM, by only using the incomplete data.

1. Write down the complete data log-likelihood if the identity of the algorithm at each trial was also recorded in the form of a discrete vector $\boldsymbol{z}$ where k-th element of $\boldsymbol{z}$ can be either 0 or 1 ($z_k \in \{0,1\}$) and $\sum_k z_k = 1$. (If the algorithm at i^{th} trial is $A1$ then $z_1^{(i)} = 1$ and $z_2^{(i)} = 0$ or simply $\boldsymbol{z}^{(i)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$)

2. In EM we have an old estimate for parameters $\boldsymbol{\theta}^{old}$ and the goal is to derive a better estimate of $\boldsymbol{\theta}$. In E-step $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \mathbb{E}_z \left[\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})|\boldsymbol{X}, \boldsymbol{\theta}^{old}\right]$ is calculated, where $\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ is the complete data log-likelihood (which you have calculated in the previous step). Show that the Q-function for the given problem can be written as:

$$Q(\theta, \theta^{old}) = \sum_{i=1}^{n} \sum_{k=1}^{2} \gamma(z_k^{(i)}) \left\{\log \pi_k + x^{(i)} \log p_k + (1 - x^{(i)}) \log(1 - p_k)\right\}$$

3. In M-step a revised parameter estimate is calculated as $\boldsymbol{\theta}^{new} = \arg\max\limits_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$. Calculate the update equation for the parameter θ (probability of success for Algorithm 1).

Solution:

Part 1

$$L(\theta) = p(\boldsymbol{X}, \boldsymbol{Z}|\theta) = p(\boldsymbol{X}|\boldsymbol{Z}, \theta)p(\boldsymbol{Z}|\theta) = \prod_{i=1}^{n}\prod_{k=1}^{2} \pi_k^{z_k^{(i)}} (p_k^{x^{(i)}}(1-p_k)^{1-x^{(i)}})^{z_k^{(i)}}$$

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n}\sum_{k=1}^{2} z_k^{(i)}(\log \pi_k + x^{(i)} \log p_k + (1-x^{(i)}) \log(1-p_k))$$

where $\pi_1 = \pi_2 = 0.5$, $p_1 = p(Success|A1) = \theta$ and $p_2 = p(Success|A2) = 0.5$.

Part 2

As we don't have the value of the latent variables, we consider the expectation, with respect to the posterior distribution of the latent variables, of the complete data log-likelihood. By using the property $\mathbb{E}[a\boldsymbol{X} + b\boldsymbol{Y} + c] = a\mathbb{E}[\boldsymbol{X}] + b\mathbb{E}[\boldsymbol{Y}] + c$ where a, b and c are constants.

$\mathbb{E}_z\left[\ln p(\boldsymbol{X}, \boldsymbol{Z}|\theta)|\boldsymbol{X}, \theta^{old}\right] = \sum_{i=1}^{n}\sum_{k=1}^{2} \mathbb{E}_z\{z_k^{(i)}\}\left\{\log \pi_k + x^{(i)} \log p_k + (1-x^{(i)}) \log(1-p_k)\right\}$

This expectation is calculated as a posterior distribution of latent variable.

$\mathbb{E}_z\{z_k^{(i)}\} = p(z_k^{(i)}|x^{(i)}, \theta^{old})$

Now using Bayes rule

$p(z_k^{(i)}|x^{(i)}, \theta^{old}) = \frac{p(z_k^{(i)}|\theta^{old})p(x^{(i)}|z_k^{(i)},\theta^{old})}{p(x^{(i)}|\theta^{old})} = \frac{p(z_k^{(i)}|\theta^{old})p(x^{(i)}|z_k^{(i)},\theta^{old})}{\sum\limits_{j=1}^{2} p(z_j^{(i)}|\theta^{old})p(x^{(i)}|z_j^{(i)},\theta^{old})} = \frac{\pi_k(p_k^{x^{(i)}}(1-p_k)^{1-x^{(i)}})}{\sum\limits_{j=1}^{2}\pi_j(p_j^{x^{(i)}}(1-p_j)^{1-x^{(i)}})} = \gamma(z_k^{(i)})$

$\gamma(z_k^{(i)})$ is also called the responsibility. With this the the final form of Q function is

$Q(\theta, \theta^{old}) = \sum_{i=1}^{n}\sum_{k=1}^{2} \gamma(z_k^{(i)})\left\{\log \pi_k + x^{(i)} \log p_k + (1-x^{(i)}) \log(1-p_k)\right\}$

Part 3

The new value of parameters are the one which maximizes the Q function. For this purpose we evaluate the partial derivative of Q w.r.t. to θ and evaluate it to zero while keeping the responsibilities fixed (as they are a function of θ^{old} and will act as constants).

$\frac{\partial Q}{\partial \theta} = 0$

$\frac{\partial Q}{\partial \theta} = \sum\limits_{i=1}^{n}\sum\limits_{k=1}^{2} \gamma(z_k^{(i)})\frac{\partial}{\partial \theta}\left\{\log \pi_k + x^{(i)} \log p_k + (1-x^{(i)}) \log(1-p_k)\right\}$

$\frac{\partial Q}{\partial \theta} = \sum\limits_{i=1}^{n} \gamma(z_1^{(i)})\{\frac{x^{(i)}}{\theta} + \frac{x^{(i)}-1}{1-\theta}\}$

$\frac{\partial Q}{\partial \theta} = \sum\limits_{i=1}^{n} \gamma(z_1^{(i)})\{(x^{(i)})(1-\theta) + (x^{(i)}-1)(\theta)\}$

$\frac{\partial Q}{\partial \theta} = \sum\limits_{i=1}^{n} \gamma(z_1^{(i)})\{x^{(i)} - \theta\} = 0$

$\implies \theta^{new} = \frac{\sum\limits_{i=1}^{n}\gamma(z_1^{(i)})x^{(i)}}{\sum\limits_{i=1}^{n}\gamma(z_1^{(i)})}$