Assignment 3
From *Estimation* to *Determination*

Please note:

1. What you must hand in includes the assignment report (.pdf) and – if necessary – source codes (.m). Please zip them all together into an archive file named according to the following template: HW3_XXXXXXXX.zip
Where XXXXXXXX must be replaced with your student ID.

2. Some problems are required to be solved *by hand* (shown by the ✎ icon), and some need to be implemented (shown by the ◢ icon).

3. As for the first type of the problems, you are free to solve them on a paper and include the picture of it in your report. Here, cleanness and readability are of high importance.

4. Your work will be evaluated mostly by the quality of your report. Don't forget to explain what you have done, and provide enough discussions when it's needed.

5. 5 points of each homework belongs to compactness, expressiveness and neatness of your report and codes.

6. By default, we assume you implement your codes in MATLAB. If you're using Python, you have to use equivalent functions when it is asked to use specific MATLAB functions.

7. Your codes must be separated for each question, and for each part. For example, you have to create a separate .m file for part b. of question 3. Please name it like p3b.m.

8. Problems with bonus points are marked by the ☆ icon.

9. **Please upload your work in Moodle, before the end of May 3ʳᵈ.**

10. If there is *any* question, please don't hesitate to contact us through the following email addresses:

    - **b.roshanfekr@yahoo.com**

11. Unfortunately, it is quite easy to detect copy-pasted or even structurally similar works, no matter being copied from another student or internet sources. Try to send us your own work, without being worried about the grade! ;)

1. Maximum Likelihood Approach for Parameter Estimation	(15+3 Pts.)

Keywords: *Parameter Estimation, Maximum Likelihood Estimation, Probability Mass Function, Sufficient Statistics*

Maximum Likelihood Estimation (**MLE**) is a **Parameter Estimation** method which tries to find the parameter values of a statistical model that maximise the **Likelihood Function**, given the observations. The resultant is called **Maximum Likelihood Estimate**, abbreviated as **MLE**.

Here, we are going to practice MLE in three different sub-problems.

Let X be a discrete random variable with the following probability mass function, with parameter θ where $0 \le \theta \le 1$.

X	1	2	3	4
$P(X)$	$2(1-\theta)/3$	$2\theta/3$	$(1-\theta)/3$	$\theta/3$

The following 10 independent observations were generated from this distribution: $(1, 3, 2, 1, 4, 2, 3, 4, 1, 4)$.

    a. What is the likelihood function $L(\theta)$?
    b. Find the log likelihood function.
    c. Using one of the above functions, determine the maximum likelihood estimate of θ.

Now, assume $x_1, x_2, ..., x_n$ be i.i.d. samples from *Erlang* distribution, with unknown parameter θ:

$$f(x \mid \theta) = \frac{1}{(m-1)!} \left(\frac{1}{\theta} \right)^m x^{m-1} e^{-\frac{x}{\theta}}, \quad x \ge 0, \quad \theta > 0$$

    d. Find the maximum likelihood estimate of the parameter θ.

Finally, suppose that n samples $y_1, y_2, ..., y_n$ are drawn independently from a continuous distribution given by:

$$f(y, \theta) = \frac{y^2}{2\theta^3} e^{-\frac{y}{\theta}}, \quad y \ge 0, \quad \theta > 0$$

    e. Find $\hat{\theta}_{ML}$.

    f. Prove that $\hat{\theta}_{ML}$ is unbiased for θ and find its variance.

    g. Show that $T(y) = \sum_{n=1}^{N} y_n$ is a sufficient statistic for θ.

    ⭐ h. Find the **Cramer-Rao Lower Bound** on the variance of the unbiased estimators of θ. Is $\hat{\theta}_{ML}$ efficient? Explain.

Useful Solved Examples: [1], [2], [3], [4]

2. Acquiring MLE in More Challenging Problems   (20 Pts.)

Keywords: *Parameter Estimation, Maximum Likelihood Estimation*

Following the previous problem, here we are going to deal with more complicated **Parameter Estimation** problems using **Maximum Likelihood Estimation**.

For each of the scenarios, determine the following items as well as those required specifically:

1. The likelihood function
2. The log-likelihood function
3. The maximum likelihood estimator for the unknown parameter
4. The numerical value of the MLE (if data are given)

a. Suppose $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are errors of some measurement, and are normally distributed with zero mean and variance σ^2. What would be the estimate for the parameter if $n = 5$, and $\varepsilon_1 = 1.3$, $\varepsilon_2 = -2.4$, $\varepsilon_3 = 2.1$, $\varepsilon_4 = -0.8$ and $\varepsilon_5 = -0.3$?

b. A radioactive sample is emitting Neutron particles according to a Poisson process at an average rate of λ per minute. The number of particles emitted during the *i*-th minute of an experiment is denoted by X_i, where $i = 1, \ldots, n$. If $n = 100$ and $\sum_{i=1}^{100} X_i = 217$, what would be the estimate for the parameter?

c. Assume a biased coin which when flipped, lands heads-up with probability p. Suppose the coin is flipped until it lands heads-up. Note that X – number of flips required – has a geometric distribution with parameter p. Find the MLE for this parameter. If 4 flips are needed before the coin lands heads-up, estimate p.

d. A student sends n statistically independent application letters for PhD position, where the probability of success on a single application request is p. Let X be the number of successful application letters sent. Find the MLE for parameter p in terms of X and n. If the student sends 28 application letters of which 6 are successful, estimate p.

3. Generalizing MLE: Maximum A Posteriori Estimation   (20+4 Pts.)

Keywords: *Parameter Estimation, Maximum Likelihood Estimation, Maximum A Posteriori (MAP) Estimation, Posterior Distribution, Prior Distribution*

Another well-known method for **Parameter Estimation** is **Maximum A Posteriori Estimation**, where the estimate of the unknown quantity equals the **Mode** of the **Posterior Distribution**. **MAP** estimation is nearly identical to **MLE**, where the only difference is the inclusion of **Prior Distribution** in MAP. In other words, the likelihood is now weighted with some weight coming from the prior. Therefore, MAP estimation can be interpreted as a **Regularization** of ML estimation.

In this problem, you are to devise ML and MAP estimators for a simple model of an uncalibrated sensor. X Is a random variable which ranges over the real numbers and determines the sensor output. Assume that when tested over a range of environments, the sensor outputs are uniformly distributed on some unknown interval $[0, \theta]$, such that

$$p(x \mid \theta) = \begin{cases} \dfrac{1}{\theta} & 0 \le x \le \theta \\ 0 & otherwise \end{cases} = \frac{1}{\theta} \mathbf{I}_{0,\theta}(x)$$

where $\mathbf{I}_{0,\theta}(x)$ denotes an indicator function which equals 1 when $0 \le x \le \theta$, and 0 otherwise. This distribution is denoted by $X \sim U(0,\theta)$. It is desirable to infer θ in order to characterise the sensor's sensitivity.

Consider n i.i.d. observations $x_1, x_2, ..., x_n$, where $X_i \sim U(0,\theta)$.

   a. Find the likelihood function $p(x \mid \theta)$.
   b. What is the maximum likelihood estimator for θ?
   c. Express an informal proof that the obtained estimator is actually the ML estimator.

Now suppose that the following prior distribution has been put on the parameter θ:

$$p(\theta) = \alpha \beta^\alpha \theta^{-\alpha-1} \mathbf{I}_{\beta,\infty}(\theta)$$

This distribution – known as a *Pareto* distribution – is denoted by $\theta \sim Pareto(\alpha, \beta)$.

   d. Considering the following hyperparameter choices, plot the three prior probability densities corresponding to them:

   $(\alpha, \beta) = (0.1, 0.1)$ $\qquad\qquad$ $(\alpha, \beta) = (2.0, 0.1)$ $\qquad\qquad$ $(\alpha, \beta) = (1.0, 2.0)$

   e. If n uniformly distributed observations $X_i \sim U(0,\theta)$ are observed, such that $\theta \sim Pareto(\alpha, \beta)$, find the posterior distribution $p(\theta \mid x)$.
   f. Is the obtained result in the previous part a member of any standard family?
   g. Find the corresponding MAP estimator of θ for the derived posterior in part e.
   h. How does MAP estimator you derived in the previous part compare to the ML estimator?
⭐ i. The **Quadratic Loss** could be defined as $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$. For the posterior derived in part e., what estimator of θ minimises the posterior expected quadratic loss?
   j. Let $x = (0.7, 1.3, 1.7)$ be our observations. Find the posterior distribution of θ for each of the priors given in part d., and plot the corresponding posterior densities.
   k. In the previous part, what would be the MAP estimator for each of the hyperparameter choices?
⭐ l. In part j., what estimator minimises the quadratic loss for each of the hyperparameter choices?

---

4. Dealing with a Basic Parameter Estimation Problem in MATLAB (15 Pts.)

Keywords: *Parameter Estimation, Maximum Likelihood Estimation, Maximum A Posteriori (MAP) Estimation, Posterior Distribution, Prior Distribution*

After practicing some problems related to **Parameter Estimation** by applying both **ML** and **MAP** estimation methods, it's time to deal with a more realistic problem. The following list of 20 numbers were generated by sampling a binomial distribution with unknown number of trials N, and probability of success p equals to 0.2 for each trial.

$$\{1, 3, 2, 2, 3, 0, 1, 2, 1, 2, 1, 1, 2, 2, 1, 3, 3, 2, 1, 3\}$$

a. Plot a bar graph which indicates the values in the list and their number of occurrences.
b. Normalise the plot in part a. and plot the empirical probability mass function.
c. Plot the probability mass function of a binomial distribution corresponding to the following parameters:

$$N = 5, p = 0.2 \qquad\qquad N = 10, p = 0.2 \qquad\qquad N = 20, p = 0.2$$

d. Which one of the probability mass functions in part c. is more likely to have generated the given list of 20 numbers?
e. Plot the likelihood function of the given numbers as a function of the parameter N.
   Note: Your graph must have integer numbers from 1 to 20 as its x-axis, and $p(x \mid N)$ as its y-axis.
f. Plot the log likelihood function.
g. Determine the value of N that maximises both likelihood and log likelihood.

Now, suppose that a prior distribution for N is given, such that it is equal to 10 with probability 0.1, and is equal to other numbers in the range of 1 to 20 ($1, 2, ..., 9, 11, ..., 20$) with probability $0.9/19$.

h. For each of the values of parameter N, plot the likelihood times the prior.
i. Is the function plotted in part h. a probability mass function?
j. Plot the posterior function by normalising the curve obtained in part h.
k. Is the function plotted in part j. a probability mass function?
l. Determine the value of N that maximises both the curves obtained in part h. and j.
m. Compare the results in part l. with the one obtained before normalisation.

Recommended MATLAB functions: `bar()`, `binornd()`, `binopdf()`

---

5. MLE vs. MAP: Which One Dominates in Image Retrieval Problem?   (15+8 Pts.)

Keywords: *Parameter Estimation, Image Retrieval, Maximum Likelihood Estimation (MLE), Maximum A Posteriori (MAP) Estimation, Image Thresholding, Dirichlet Smoothing*

In this problem, you are to apply **Parameter Estimation** methods in a real-world problem, i.e. **Image Retrieval**, where given a query image and based on features extracted from some query image, the goal is to rank a set of images according to how similar their distribution of features are to the query image.

Consider K distinct types of features that can be measured at each pixel of an image. Assume there are n_1 pixels where a feature of type 1 exists, n_2 pixels where a feature of type 2 exists, and so on. Therefore, each image can be represented by a categorical distribution with K parameters $u_i = n_i/N$, where $N = n_1 + ... + n_K$. These values are essentially MLE estimates, obtained by forming a histogram of feature counts n_i over the whole image and then dividing by the sum of all counts to get a probability mass function that sums to one.

Now let's define a scoring function which computes how similar the categorical distribution $[u_1, u_2, ..., u_k]$ acquired for image I is to the categorical distribution $[q_1, q_2, ..., q_k]$ describing the query image Q. In order to define this score, the probability that the features in Q were generated from the categorical distribution representing image I is considered. Assuming that the features

Statistical Pattern Recognition   Amirkabir University
of Technology
(Tehran Polytechnic)   *Spring 2019*

in Q are drawn independently from the categorical distribution of I , one can compute this score such that

$$score(Q, I) = cP(Q|u_1, \dots, u_k) = u_1^{q_1} u_2^{q_2} \dots u_K^{q_k}$$

Here, c is a positive constant and will be ignored. Note that each q_i is a proportional to an integer count n_i , since $q_i = n_i/N$ for some value N . As long as the computed scores are going to be used to rank different images I with respect to the same query image Q , it would be OK to use the q_i values rather than n_i .

After defining a score for comparing two images, now it would be possible to compare the query image Q with each of M images $I_1, I_2, \dots, I_M$ in the gallery. Clearly, images with the highest scores will be the most similar to the query image Q .
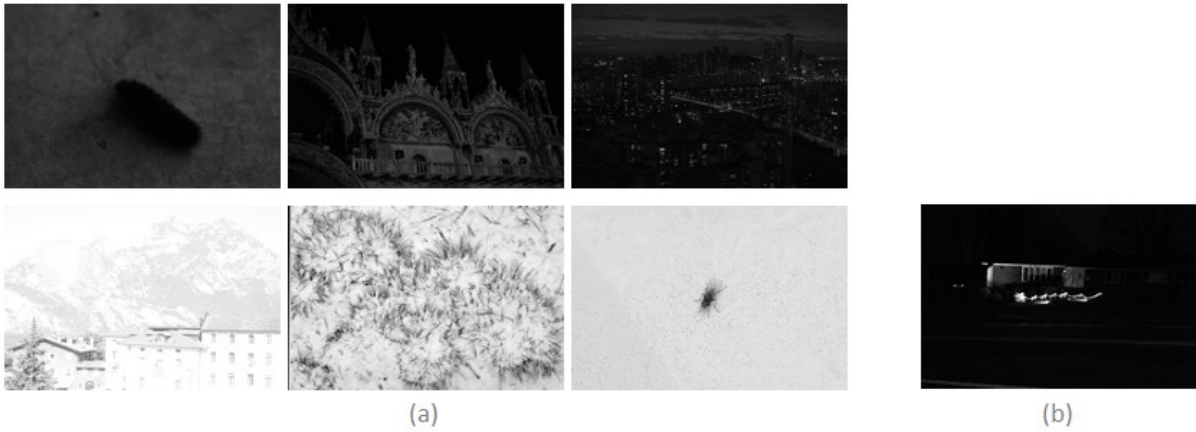

(a)   (b)

Figure 1 A collection of images given for this problem (a) Gallery images (b) Query image

As illustrated in Figure 1, here you will work with a small dataset containing six images $I_1, I_2, \dots, I_6$ as the gallery, and a single test image Q to use as the query. Consider four different feature types ($K = 4$) computed by thresholding a pixel's value g . More precisely, each type of feature for a specific pixel can be defined as

$$\text{Feature type} = \begin{cases} 1 & 0 \le g < 64 \\ 2 & 64 \le g < 128 \\ 3 & 128 \le g < 192 \\ 4 & 192 \le g < 256 \end{cases}$$

a. Compute the maximum likelihood estimate of categorical distribution parameters for all images, by finding the comparison scores between Q and each image I . Sort the scores from highest to lowest and determine which images are deemed to be most similar.

b. Based on the image histograms and the categorical parameters computed from them, try to explain why the above scoring/ranking fails so badly.

c. Now consider a Dirichlet prior to impose knowledge about the expected distribution of feature values across the population of all images in the gallery. To apply this method, known as **Dirichlet Smoothing**, start by using MLE to compute the parameters $\{\rho_1, \rho_2, \dots, \rho_K\}$ of a categorical distribution of features across all the images in the gallery,

in order to get a population distribution. This distribution indicates the overall frequency with which to expect any given feature to appear; a priori. To from a Dirichlet prior to use for MAP estimation, a parameter a_i is applied, such that $a_i = 1 + \eta \rho_i$. Repeat part a. using MAP estimation to determine the categorical distributions representing each image. Play around with different values of η to see how it affects the results. Find the sorted scores, discuss the results and compare them with part a.

Note: Images can be found in "input/P5" folder attached to this homework. Gallery images are "img1.png" to "img6.png", where the query image is "test.png".

Hint 1: Digital images are in fact matrices of pixels, where each pixel is represented by a numerical value. Therefore, when dealing with grayscale images, you are actually working with a matrix of numbers.

Hint 2: As can be seen in Figure 1, the gallery contains three very bright and three very dark images, and the query image is also very dark. Therefore, it is expected to obtain a rank ordering to rank the dark images more highly than the bright ones.

Hint 3: The parameter η acts as a variable smoothing parameter. Setting this parameter nearly equal to the number of pixels in the image will cause the prior information be nearly equal importance as the data. If η sets to be very large, the prior will swamp the data, and if η is very low, the data will dominate. $\eta = 0$ would reduce the process to ML estimation.

Recommended MATLAB functions: `imread(), double(), im2double(), imhist()`

---

6. Some Explanatory Questions   (10 Pts.)

Please answer the following questions as clear as possible:

a. How can you relate ML and MAP estimation?
b. What is penalized MLE? When is it better to apply penalized MLE instead of normal MLE?
c. How does ML and MAP estimation change when the data isn't i.i.d.?
d. How would you relate bias/variance to underfitting/overfitting? Explain.
e. How can you interpret Bayes decision rule in terms of bias/variance?
   Hint: Pay attention to prior distribution shape.

Good Luck!
Ali Abbasi, Behnam Roshanfekr