P. Koumoutsakos
ETH Zentrum, CLT E 13
CH-8092 Zürich

HW 2 - Probability & Bayesian Inference

Issued: March 4, 2019
Due Date: March 18, 2019 10:00am

1-Week Milestone: Solve tasks 1 to 3; install python (version > 3.0) and make yourself comfortable with python (task 4.1) (March 11).

Task 1: Probability Theory Reminders

In this exercise we fix the notation we will use during this course and refresh our memory on basic properties of random variables. Present your answers *in detail*.

a) [10pts] A random variable with normal (or Gaussian) distribution $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$ has probability density function (pdf) given by

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} . \tag{1}$$

Show that the mean and the variance of X are given by $\mathbb{E}\left[X\right] = \mu$ and $\mathbb{E}\left[(X-\mu)^2\right] = \sigma^2$, respectively.

b) [10pts] The probability that a random variable X with pdf f_X is less or equal than any $x \in \mathbb{R}$ is given by,

$$P(X \le x) \,=\, F_X(x) \,=\, \int_{-\infty}^{x} f_X(z)\,\mathrm{d}z\,. \tag{2}$$

The function F_X is called the cumulative distribution function (cdf).
The Laplace distribution with parameters μ and β has pdf,

$$f(x) = \frac{1}{2\beta}\exp\left(-\frac{|x-\mu|}{\beta}\right). \tag{3}$$

i) Find the cdf of the Laplace distribution.

ii) Use the cdf to find the median of the Laplace distribution.

c) [10pts] The pdf of the quotient $Q = X/Y$ of two random variables X, Y is given by,

$$f_Q(q) \,=\, \int_{-\infty}^{\infty} |x|\, f_{X,Y}(qx, x)\,\mathrm{d}x\,, \tag{4}$$

where $f_{X,Y}$ is the joint pdf of X and Y.
Assume that X and Y are independent random variables with pdfs $f_X(x) = \mathcal{N}\left(x|0, \sigma_X^2\right)$ and $f_Y(y) = \mathcal{N}\left(y|0, \sigma_Y^2\right)$.

i) Find the joint pdf of X and Y.

ii) Show that $Q = X/Y$ follows a Cauchy distribution with zero location parameter and scale $\gamma = \sigma_X/\sigma_Y$. The pdf of a Cauchy distribution with location parameter x_0 and scale γ is given by,

$$f(x) = \frac{1}{\pi}\frac{\gamma}{(x - x_0)^2 + \gamma^2}\,. \tag{5}$$

Task 2: Bayesian Inference

You are given a set of points $\boldsymbol{d} = \{d_i\}_{i=1}^N$ with $d_i \in \mathbb{R}$. You make the *modelling assumption* that the points come from N realisations of N *independent* random variables X_i, $i = 1, \dots, N$, that follow normal distribution with unknown parameter μ and known parameter $\sigma = 1$.

a) [10pts] Formulate the *likelihood function* of μ,

$$\mathcal{L}(\mu) := p(\boldsymbol{d}|\mu)\,, \tag{6}$$

where p is the conditional pdf of $\boldsymbol{d}$ conditioned on μ.

b) [15pts] Find the *maximum likelihood estimate* (MLE) of μ, i.e.,

$$\hat{\mu} = \arg\max_{\mu} \; \mathcal{L}(\mu). \tag{7}$$

You may find useful that $\arg\max_\mu \mathcal{L}(\mu) = \arg\max_\mu \log \mathcal{L}(\mu)$.

c) [30pts] Before observing any data $\boldsymbol{d}$ you had the belief that μ follows a normal distribution with mean μ_0 and variance σ_0^2. After observing the dataset $\boldsymbol{d}$ you *update your belief* by using Bayes' theorem. Identify the *posterior distribution* $p(\mu|\boldsymbol{d})$ of μ conditioned on $\boldsymbol{d}$. Calculate the mean and the variance of $p(\mu|\boldsymbol{d})$.

d) [5pts] Find the *maximum a posteriori* (MAP) estimate of μ, i.e.,

$$\hat{\mu} = \arg\max_{\mu} \; p(\mu|\boldsymbol{d})\,. \tag{8}$$

e) [10pts] Perform (c) and (d) using as prior an uninformative distribution, i.e. a uniform distribution in $\mathbb{R}$, and compare the MAP with the MLE. Although this not a distribution, since it is not integrable over $\mathbb{R}$, we are allowed to use it in Bayes' theorem al long as the posterior is a distribution. These priors are called *improper priors* and a common choice in practical applications when there is no prior information on the parameters.

Task 3: Bayesian Inference: Linear Model

[20pts] You are given the linear regression model that describes the relation between variables x and y,

$$y = \beta x + \epsilon,$$

where β is the regression parameter, y is the output quantity of interest (QoI) of the system, x is the input variable and ϵ is the random variable accounting for model and measurement errors. The model error is quantified by a Gaussian distribution $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

You are given one measurement data point, $D = \{x_0, y_0\}$. Consider an uninformative prior for β and identify the posterior distribution of β after observing D. Calculate the MAP and the standard deviation of $\beta|D$.

Task 4: Data Analysis

This task consists of python implementation and work on paper. Python is one of the most widespread[1] programming languages across research and industry, and hence we would like you to get your hands on it.
Your task will be to analyse the data set provided in the file *task4.npy* and to implement some simple statistical functions. Finally you will visualise your findings in order to support your analysis.
The prepared python script *task4.py* contains a few lines of code to start with and some recommendations in comment sections on how to complete this task. You must not follow our structure and you are free to implement your own solution.

1. **Read Data** [0pts]

   - Make yourself comfortable with python and know how to call methods from python packages. Use the following command to make sure that your python version is newer than 3.0:
     ```
     $ python −−version
     ```
     On Euler you have to load a newer version:
     ```
     $ module load python/3.6.0
     ```

   - Study and run the python script *task4.py*.

   (Info: no documentation needed)

2. **Histogram** [0pts]

   - Visualise the data $\boldsymbol{d} = \{d_i\}_{i=1}^{20193}$ with following three lines provided in the script:
     ```
     counts, bins = hist(data,numbins,continuous=True)
     plt.bar(bins, counts, width=0.5, align='edge')
     plt.show()
     ```

   - Play around with the parameter `numbins`.
     Understand what happens inside the function `hist(xarr, nbins, continuous =True)`.

   (Info: no documentation needed)
   (Hint: as you advance with your implementation you might want to comment the lines above in the script in order to avoid interruptions during testing)

3. **Normalized Histogram** [5pts]
   Adjust the histogram function (where indicated in the script) such that it returns an estimated pdf for the data $\boldsymbol{d}$ (for continuous and discrete distributions), i.e. the input vector `xarr` .
   After that, visualize the normalized histogram and self-check if your solution is reasonable.

[1] https://stackoverflow.blog/2017/09/06/incredible-growth-python/?_ga=2.199625454.1908037254.1532442133-221121599.1532442133

(Hint: for a continuous pdf it holds that $\int_{-\infty}^{\infty} f_X(u)du = 1$; for a discrete pdf it holds that $\sum_{i=1}^{N} p(x_i) = 1$)

4. **Distribution Function** [5pts]
   By now you should be familiar with the data, make an educated guess what the underlying distribution function could be. Document your hypothesis and rationalize your choice (2 - 5 sentences).

   (Hint 1: we generated the data with a function provided in the *numpy.random* package:
   https://docs.scipy.org/doc/numpy-1.14.0/reference/routines.random.html)
   (Hint 2: it's not the Gaussian distribution)

5. **Likelihood and Log-likelihood** [5+15pts]
   - Write down the likelihood $\mathcal{L}(\boldsymbol{\theta})$ and log-likelihood $\ell(\boldsymbol{\theta}) = \log \mathcal{L}(\boldsymbol{\theta})$ of your density function.
   - Implement both functions in python where indicated in the script. The likelihood, respectively log-likelihood, should take an input for each of the parameters $\boldsymbol{\theta}$ of the pdf, as well as an input for the data data $\boldsymbol{d}$.

   (Info: For one liners you might want to use lambda functions
   https://www.programiz.com/python-programming/anonymous-function)

6. **MLE** [5+10pts]
   - Derive the MLE of the parameters $\hat{\boldsymbol{\theta}} = \arg\max \mathcal{L}(\boldsymbol{\theta})\}$ and document your result.
   - Add a few lines to the python script to calculate the MLE of the parameters $\boldsymbol{\theta}$ from the full data set $\boldsymbol{d}$.

   (Hint: you might want to start from the definition of the log-likelihood function in order to derive the MLE).

7. **Comparison with Gaussian Distribution** [5+15+10pts]
   - In python, compute the maximum likelihood and maximum log-likelihood of your pdf given the full data set $\boldsymbol{d}$ (reuse the functions implemented in subtask 5).
   - Implement a likelihood and log-likelihood function for the Gaussian distribution too (according to subtasks 5 & 6).
   - Compare the results from bullet 1 with the maximum likelihood, respectively maximum log-likelihood, of the Gaussian distribution (use the Gaussian MLE $\boldsymbol{\theta} = (\hat{\mu}, \hat{\sigma}^2)$). Answer the following questions and document: Which distribution function provides a better likelihood, respectively log-likelihood? Do you encounter any problems during the calculation of the likelihood - or why is it advisable to work in the log space? (in total 2-5 sentences).

(Hint 1: the log-likelihood of the Gaussian pdf for the full data set equals -40287.57)
(Hint 2: you can self-check your implementations by comparing the output of the likelihood function with the exponential of the output of the log-likelihood function)

8. **Visualisation** [5+5+5pts]
   - Plot your distribution function on top of the normalized histogram from subtask 3.
   - Plot the distribution function of the Gaussian on top of the other two graphs.
   - Which graph provides a qualitatively better fit to the histogram from subtask 3? Do your findings agree with the numbers calculated in subtask 7? State your answer in 2-5 sentences and include the final plot into your report.

(Hint: you can uncomment `plt.show()` from subtask 2 & 3, add two more plot statements `plt.plot(x,y)` and then insert `plt.show()` at the end of the script)

Task 5: 1-D Laplacian Approximation

The Laplace method approximates a continuous distribution function $p(x)$ with a Gaussian distribution. For the approximation we perform a Taylor expansion of the logarithm of $p(x)$ around the maximum $\hat{x}$.
In this task we derive the Laplace approximation of the Cauchy distribution and then plot both distributions.

1. **Maximum** [5pts]
   Let $x \in \mathbb{R}$ be a parameter with a continuous probability distribution function $p(x)$. Which conditions do **not** hold for the global maximum $\hat{x}$ of the pdf $p(x)$:
   (i) $\frac{\partial p}{\partial x}\Big|_{\hat{x}} < 0,$
   (ii) $\frac{\partial p}{\partial x}\Big|_{\hat{x}} = 0,$
   (iii) $\frac{\partial p}{\partial x}\Big|_{\hat{x}} > 0,$
   (iv) $\frac{\partial^2 p}{\partial x^2}\Big|_{\hat{x}} < 0,$
   (v) $\frac{\partial^2 p}{\partial x^2}\Big|_{\hat{x}} = 0,$
   (vi) $\frac{\partial^2 p}{\partial x^2}\Big|_{\hat{x}} > 0.$

2. **Taylor expansion** [5pts]
   The logarithm of the probability density is given by $\ell(x) = \log p(x)$. The Taylor expansion of the logarithm of $p(x)$ around the maximum $\hat{x}$ is given by

$$\ell(x) = \ell(\hat{x}) + \frac{1}{2}\left.\frac{\partial^2 \ell}{\partial x^2}\right|_{\hat{x}} (x-\hat{x})^2 + \mathcal{O}(x-\hat{x})^3, \tag{9}$$

   The 1-D Laplace approximation of the probability distribution $p(x)$ can be approximated by

$$p(x) \approx A \exp\left(\frac{1}{2}\left.\frac{\partial^2 l}{\partial x^2}\right|_{\hat{x}} (x-\hat{x})^2\right). \tag{10}$$

   Which steps and assumptions did we do to go from 9 to 10? Also find an expression for the constant A.

3. **Laplace approximation** [10pts]
   One can see that the 1-D Laplace approximation (equation 10) has the form of a Gaussian distribution with mean $\mu = \hat{x}$. First, derive an expression for the variance σ^2. Second, rewrite the Laplace approximation (equation 10) as a product of the pdf of the Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ and constant terms left to define.

4. **Cauchy distribution** [5+5+5pts]
   The Cauchy distribution is given by

$$p(x|x_0, \gamma) = \frac{1}{\pi}\frac{\gamma}{(x-x_0)^2 + \gamma^2} \tag{11}$$

- Derive analytically the location of the maximum $\hat{x}$ of the Cauchy distribution and document it.
- Write down the Taylor expansion up to order 2 of the logarithm of the Cauchy distribution around the maximum. Start with the derivation of the second derivative of $\ell(x|x_0, \gamma)$.
- Write down the Laplace approximation of the Cauchy distribution. Also define the normalization constant if you choose to use one.

5. **Visualisation** [20pts]

   In this subtask you are asked to create a plot with two graphs. You can solve it with any programming language you like. You might want to reuse your routines from task 4.

   - Plot the Cauchy $p(x|x_0, \gamma)$ distribution with parameters $x_0 = -2.0$ and $\gamma = 1.0$.
   - Add a graph of the Laplace approximation of $p(x|x_0, \gamma)$.

   Include the final plot with both graphs into your report (code not needed). Do not forget to add labels for the axis and a legend for the graphs. Make sure that the modes are inferable from the plot.

Guidelines for reports submissions:

- Submit two files via Moodle: a **pdf** of your report (including plots from task 4 & 5) plus the solution to the coding exercise *task4.py*.