Data Science: Revision Exercises

Maximum Likelihood Estimator (MLE)

Extensions to the Supervision Work

Exercise 1 : In E1Q11, you computed the MLE using a single sample.

1. A friend of yours wants to hash with multiple hash functions and report the mean of the MLEs. Perform an experiment to see how well this performs in practice for $m = 1000$ and 10 repetitions.
2. Now, derive the proper MLE for having multiple samples. Compare the accuracy of this estimator with your friend's estimator.

Exercise 2 : We have two biased coins with probability p_1 and p_2. We are goining to throw N coin tosses, starting with the first coin and after K coin tosess we switch to the second coin (where K is chosen at the beginning).

1. Write code which given p_1, p_2, N and K generates a sequence of coin tosses.
2. Compute the likelihood function for p_1, p_2 given N and K.
3. Write code to find the MLE estimate for p_1, p_2 and K given the observations. *Hint:* You will need to loop over the possible values of K. *Hint 2:* Compute the log-likelihood to avoid underflow.
4. Run the MLE code for the example dataset you generated in (1). How close is your estimate?
5. (optional) What is the time complexity of your algorithm (in terms of N)?
6. (optional) How can you efficiently generalise this for L coins? What is the running time in this case? *Hint:* You need to use dynamic programming.

Exercise 3 : In the lecture notes, you derived the MLE for the Binomial distribution when p is unknown. Now, we will investigate the case where n is unknown.

Suppose that k is a realisation of a random sample from a binomial distribution with parameters (n, p), where p is known and n is unknown. (For example, this could correspond to flipping a coin with a known probability, but not knowing how many times the coin was flipped).

1. Write an expression for the likelihood $\text{lik}(n)$.
2. Show that $\text{lik}(n) = 0$ for $n < k$.
3. Show that

$$\frac{\text{lik}(n+1)}{\text{lik}(n)} \geq 1 \Rightarrow n \leq \frac{k}{p} - 1$$

4. Deduce that $\lfloor k/p \rfloor$ is an MLE. Show that if k/p is an integer, then there are two MLEs.

Exercise 4 : In Exercise 1.1.2 of the lecture notes, you used the plug-in principle also known as the invariance principle.

1. State and prove the invariance property for the MLE. *Hint:* Look here if you are having trouble.
2. Find the MLE estimate for the variance of the Binomial distribution with known n and unknown p.
3. Find the MLE estimate for $\mathrm{Pr}(X = 0)$ given observations $x_1, \ldots, x_n$ from $\mathrm{Po}(\lambda)$.
4. Find the MLE estimate for $\mu^2 + \sigma^2$ given observations $x_1, \ldots, x_n$ from $\mathcal{N}(\mu, \sigma)$.
5. Why is the invariance property important?

Exercise 5 Are the MLE estimators biased?:

1. Recall from Part IA Probability, what it means for an estimator to be *biased*.
2. Is the estimator of E1Q2 biased?
3. Is the estimator of E1Q4 biased? See E1Q11(a) for the distribution of the maximum. What happens as $n \to \infty$?
4. Is the estimator of E1Q6 biased?
5. (optional) If you are interesting in learning more about the MLE being unbiased, see the Fisher Information Project.

Exercise 6 : Download the dataset containing the total expences per MP. Using the analysis of E1Q5, compare the average expenses for MPs in London and MPs out of London. Do the assumptions apply here?

(optional) Perform a similar analysis for the average expenses between parties (or among genders). (You will need to find another dataset that maps MPs to their parties).

Exercise 7 : (+) The unigram model in natural language processing models the probability of a sentence as s as $\mathbb{P}(s) = p_{s_1} \cdot p_{s_2} \cdot \ldots \cdot p_{s_n}$ where $s_1, \ldots, s_n$ are the n words of the sentence. Given M sentences $s^1, \ldots, s^M$, show that the MLE for the parameters p_w are $\frac{c_w}{W}$, where c_w is the number of times word w occurs in any sentence and W is the total number of words in all sentences.

Hint: This question is essentially asking you for the MLE of the multinomial distribution.

Additional exercises

Exercise 8 : A very inexperienced archer shoots an arrow n times at a disc of (unknown) radius θ. The disc is hit every time, but at completely random places. Let $r_1, \ldots, r_n$ be the distances of the various hits to the center of the disck.

1. Show that given θ the pdf is $f_\theta(r) = \frac{2r}{\theta^2}$ for $0 \leq r \leq \theta$.
2. Determine the Maximum Likelihood estimate for θ.

Exercise 9 : Suppose that $x_1, \ldots, x_n$ is a dataset, which is a realisation of a random sample from a Rayleigh distribution, which is the continuous distribution with pdf $f_\theta(x) = \frac{x}{\theta^2} \exp\left\{-\frac{1}{2}\frac{x^2}{\theta^2}\right\}$ for $x \geq 0$. Determine the maximum likelihood estimator for θ.

Exercise 10 : Suppose that $x_1, \ldots, x_n$ is a dataset, which is a realisation of a random sample from a distribution with pdf

$$f_\theta(x) = \frac{\theta}{(x+1)^{\theta+1}} \text{ for } x > 0$$

Determine the MLE for θ.

Exercise 11 : Suppose that $x_1, \ldots, x_n$ is a dataset, which is a realisation of a random sample from a distribution with pdf

$$f_\theta(x) = \begin{cases} e^{\theta - x} & \text{for } \ x > \theta \\ 0 & \text{for } \ x \leq \theta \end{cases}$$

1. Determine the MLE for θ.
2. Is there anything weird with this distribution?

Exercise 12 : (+) Suppose that $x_1, \ldots, x_n$ is a dataset, which is a realisation of a random sample from a distribution with pdf

$$f_\theta(x) = \frac{1}{2} e^{-|x-\theta|} \text{ for } -\infty < x < \infty$$

Determine the maximum likelihood estimator for θ.

Exercise 13 : Suppose that $x_1, \ldots, x_n$ is a dataset, which is a realisation of a random sample from a distribution with pdf

$$f_{\mu,\lambda}(x) = \left(\frac{\lambda}{2\pi x^3}\right)^{1/2} \exp\left\{-\lambda(x-\mu)^2/(2\mu^2 x)\right\} \text{ for } x > 0$$

Determine the maximum likelihood estimator for μ and λ.

Exercise 14 : (+) Create your own MLE exercise.

Random variable transformations

Revision

1. You are given a random variable X, its cdf F_X and a function f. Let $Y = f(X)$.
   1. Find the cdf for Y for the case where f is non-decreasing.
   2. Find the cdf for Y for the case where f is non-increasing.
   3. Describe how you would find the cdf for Y for the case where f is not monotonic.
2. What does it mean to generate a random variable?
3. How is generating a sample from a distribution related to random variable transformation (for increasing functions)?
4. Describe the general approach for generating a sample from a distribution with cdf F_X. What is the graphical interpretation for this?

Transformation

Exercise [Linear transformation]: Consider r.v. X and $Y = aX + b$ for constants a ($a \neq 0$) and b. Find the cdf of Y in terms of the cdf of X.

Exercise [Constant transformation]: Consider r.v. X with cdf F_X and the transformation $Y = f(X) = a$ where a is a constant. Find the pdf Y?

Exercise [Cubic transformation]: Consider r.v. $X \sim U[0,1]$ and $Y = X^3$. Find the cdf and pdf for X.

Exercise: Redo exercise 1 from Example Sheet 1.

Exercise [Sine transformation]: Suppose $X \sim U[0, 2\pi]$. Consider $Y = \sin^2(X)$. Find the cdf for Y.

Exercise [Square transformation]: Suppose X is a continuous random variable. Find the cdf and pdf for $Y = X^2$.

Exercise [Normal-gamma squared relationship]: Let $X \sim \mathcal{N}(0,1)$, show that $Y = X^2$ follows a gamma distribution.

NumPy and SciPy

Exercise 15 : Write efficient NumPy expressions for the following:

1. Given a matrix A, compute the sum of all its entries.
2. Given a matrix A, compute the sum of the squares of all entries.
3. Given a vector v, return the items at odd indices.
4. Given a vector v, return the sum of entries at odd indices.
5. Given a matrix A, vectors x and y, compute $(Ax - y)^2$.
6. Given two vectors of observations x and y, keep the observations with $x_i \geq t$.
7. Find the inverse of a square matrix A.
8. Given a list of linear models and a matrix X of feature vectors, find a confidence interval for each point. (*Hint:* See E3Q8)
9. Given a vector v compute the sum of all entries except for the last one.
10. Given a vector p, write code that verifies that this is a probability vector.
11. Given a finite probability vector p, write code that computes the expectation and variance.
12. Given a finite probability vector p, compute the CDF.
13. Greate a matrix A of size n with the checkboard pattern.
14. (+) Given two vectors a and b, find the ordering of the elements of b that minimises $||a - b||$.
15. Given a matrix A, subtract the mean from each row.
16. Given a matrix A and a value x find the entry in A that is closest to x.
17. Given a matrix A, compute its rank.
18. Given a vector v, find the k largest values.
19. (optional) Implement Conway's Game of Life in NumPy.

Exercise 16 Counterexamples for fmin:

1. Construct an example of a function for which `scipy.optimize.fmin` finds a local minimum.
2. Construct an example of a fucntion for which `scipy.optimize.fmin` fails to find a local minimum.

Linear models

Revision

- Define a linear model.
- State and prove the connection between linear models and least squares.
- What does it mean for the parameters of a model to be identifiable?
- When are the parameter values interpretable?
- What is a one-hot encoding?
- How can we incorporate non-linearity in linear models?
- What is a residual plot? How can we use it to improve our model?

Extensions to supervision work

Exercise 17 : Repeat E1Q14, but this time don't assume that the inflection point is at 1980.

Exercise 18 : Extension of E1Q17.

1. Generate a dataset where the st. deviation changes with (sigma x_i)^3 (or 4). Show how would you fit this.
2. (optional) Generalise your code, so that it works for general k.

Bayesian inference

Revision

1. State Bayes' rule.
2. What is the *posterior*, *prior* and *likelihood*?
3. What is the main difference between Bayesian and frequentist inference?
4. Derive the posterior distribution for a Normal distribution where the mean is distributed according to a Normal prior. (E2Q3)
5. Derive the posterior distribution for a Beta distribution where the mean is distributed according to a Beta prior. (E2Q6)

Recommended reading: Chapter 7 from "The Science of Uncertainty" (see here)

Additional exercises

Exercise 19 : Extension of E2Q1.

1. Choose a relatively simple shape (e.g. a cartoonised tree) and create a PDF that generates
2. How does the plotting method affect the shape generated? What would happen if we run for sufficiently long?
3. How could you create a density function that "draws" any shape given as a pixel matrix (i.e. a matrix where each entry is between $[0, 1]$ indicating the intensity of a pixel at that (x, y) point).

Exercise 20 : We are given a random sample $x_1, \ldots, x_n$ from the Poisson distribution $\mathrm{Po}(\theta)$, where $\theta \sim \Gamma(\alpha, \beta)$ for constant α and β. Determine the posterior $\Pr(\theta \mid x_1, \ldots, x_n)$.

Exercise 21 : We are given a random sample $x_1, \ldots, x_n$ from the exponential distribution $\mathcal{E}(\theta)$, where $\theta \sim \Gamma(\alpha, \beta)$ for constant α and β. Determine the posterior $\Pr(\theta \mid x_1, \ldots, x_n)$.

Exercise 22 : Example 7.14 from “The Science of Uncertainty” (see here)

Linear independence

Revision

- What does it mean for n vectors to be linearly independent?
- Define the space spanned by n vectors.

Exercises

Exercise 23 : Are the following vectors independent?

1. $v_1 = (1, 2)$ and $v_2 = (-5, 3)$
2. $v_1 = (1, 2)$ and $v_2 = (-4, -8)$
3. $v_1 = (2, -1, 1)$, $v_2 = (3, -4, 2)$ and $v_3 = (5, -10, -8)$

Exercise 24 : Show that the if a vector v belongs to the span of vectors v_1, v_2, v_3 then $\{v, v_1, v_2, v_3\}$ are not linearly independent.

Exercise 25 : Using NumPy (`np.linalg.matrix_rank`) determine whether the following vectors are linearly dependent:

1. $v_1 = [0.1, 0.3, 0.4, 0.5], v_2 = [0.2, 0.1, 0.4, 0.1], v_3 = [0.2, 0.2, 0.1, 0.4]$.
2. $v_1 = [0.1, 0.3, 0.4, 0.5], v_2 = [0.3, 0.4, 0.8, 0.6], v_3 = [0.1, -0.2, 0., -0.4]$.
3. $v_1 = [0.3, 0.1, -0.2, 0.3], v_2 = [0.4, 0.8, 0.9, 1.4], v_3 = [0.1, -0.1, -0.3, -0.1]$.
4. $v_1 = [0.3, 0.1, -0.2, 0.3], v_2 = [0.4, 0.8, 0.9, 1.4], v_3 = [0., 0.1, -0.3, 0.3]$.

Exercise 26 : Given $n + 1$ vectors in $\mathbb{R}^n$, can they be independent? (Do not give a proof for this).

Exercise 27 : What is the minimum number k of vectors in $\mathbb{R}^n$ that can be linearly dependent?

Exercise 28 : Give n vectors in $\mathbb{R}^n$ which are linearly independent.

Confidence intervals

Revision

- Define a confidence interval in a frequentist setting.
- Define a confidence interval in a Bayesian setting.
- Explain the difference between the two.
- Describe how to estimate a confidence interval using resampling in a frequentist setting.
- Explain how this is applied to E3Q3.
- Describe how to estimate a confidence interval by sampling models in a Bayesian setting. Write down the assumptions and procedure.
- Explain how this is applied to E2Q5.

Exercises

Exercise 29 :

1. Repeat E3Q2, and compute a confidence interval for each of the parameters.
2. Try to write generic code that given a linear model, find an α-confidence interval for a specified parameter.

Exercise 30 :

1. Repeat E2Q5, and compute a confidence interval for each of the parameters.
2. Try to write generic code that given a linear model, find an α-confidence interval for a specified parameter.

Exercise 31 : Extension of E3Q8. Create a similar confidence band some other dataset of the lecture notes where you did prediction, e.g. the Iris dataset.

Exercise 32 : Perform a hypothesis test for whether the dataset has variance that changes with the square of variance.

Exercise 33 [Normal confidence interval]: Derive a confidence interval for the mean of the normal distribution (with known variance) given n samples. Why do we usually choose confidence intervals of equal tails?

Exercise 34 [Uniform distribution]: In this exercise, you will construct (meaningful) confidence intervals for the parameter of the uniform distribution. You are given i.i.d. r.vs. $X_1, \ldots, X_n \sim U[0, \theta]$.

1. Consider the r.v. $Q = \frac{1}{\theta} \max_{i=1,\ldots,n} X_i$, show that $\mathbb{P}(Q \leq t) = t^n$ (see order statistics exercises).
2. Show that $\mathbb{P}(Q \leq 1) = 1$.
3. By considering $\mathbb{P}(t \leq Q \leq 1)$, construct a $1 - \alpha$ confidence interval.

Datasets from the course

Exercise 35 : For each of the datasets that you presented in the course, collect the inference tasks and visualisations that you applied on the datasets. Are there some tasks that you can apply only on certain types of datasets?

Now, it is your time to to look in more depth at the data sets you have

Exercise 36 : For the police stop-and-search dataset, find an interesting hypothesis or some sort of modelling task and apply the techniques your learnt to solve it. Use visualisations that support your insights.

Exercise 37 : For the iris dataset, find an interesting hypothesis or some sort of modelling task and apply the techniques your learnt to solve it. Use visualisations that support your insights.

Exercise 38 : For the police stop-and-search dataset, find an interesting hypothesis or some sort of modelling task and apply the techniques your learnt to solve it. Use visualisations that support your insights.

Further datasets

Exercise 39 :

1. Find any reasonable dataset online (or in the extended online notes). (You may want to look at Our World in Data, Kaggle, the UCI repository, the SK-learn datasets, wikipedia)
2. Identity some interesting Data Science task on the dataset.
3. Use the modelling and analysis techniques you learnt in the course to solve the task.
4. Visualise the task. (You may need to iterate between this and the previous two steps multiple times).
5. Evaluate how well you solved the task.

Exercise 40 : Repeat the previous exercise with several datasets, but there is no need to write code.

Exercise 41 : (optional)

1. Read about the Oxford/AtraZeneca vaccine trials in this journal article or about the Pfizer trials in this journal article.
2. Describe the sample selection process.
3. What data science techniques are used in the process? How is efficacy defined?
4. How is expert knowledge incorporated in the analysis?
5. (Optional) Using resampling compute a confidence interval on the efficacy.
6. Is there anything that you would have done differently?

Conditional probability and independence

Exercise 42 [Extended multiplication rule]:

1. Show that $\mathbb{P}(A, B, C) = \mathbb{P}(A|B, C)\mathbb{P}(B|C)\mathbb{P}(C)$.
2. Show that $\mathbb{P}(A_1, \ldots, A_n) = \mathbb{P}(A_1|A_2, \ldots, A_n) \cdot \ldots \mathbb{P}(A_{n-1}|A_n)\mathbb{P}(A_n)$.

Exercise 43 [Decomposition]: Show that if X is independent of (A, B) then it is independent of A and independent of B.

Exercise 44 [Weak Union]: Show that if X is independent of (A, B), then X is independent of A given B.

Exercise 45 [Contraction]: Show that if X is independent of A given B, and X is independent of B, then X is independent of (A, B).

Exercise 46 [Machine translation]: (optional) In this exercise, you will prove the famous probability formulations of the IBM models. In the context of statistical machine translation, we have a source sequence $\mathbf{s}$ (where s_i is the i-ith element and s_i^j denotes the elements $s_i, \ldots, s_j$), a target sequence $\mathbf{t}$ and an alignment $\mathbf{a}$, indicating connections between the source and the target sequences.

1. (easy) Show that if m is the length of the target sequence then $\mathbb{P}(t, a|s) = \mathbb{P}(t, a, m|s)$.

2. Show that assuming independence between a and t given s,

$$\mathbb{P}(t, a, m|s) = \mathbb{P}(m|s) \prod_{i=1}^{|t|} \mathbb{P}(a_i|a_1^{i-1}, s, m)\mathbb{P}(t_i|f_1^{i-1}, s, m)$$

3. Show that

$$\mathbb{P}(t, a, m|s) = \mathbb{P}(m|s) \prod_{i=1}^{|t|} \mathbb{P}(t_i, a_i|a_1^{i-1}, s, m)$$

4. Show that $\mathbb{P}(t_j|t_1^{j-1}, a_1^j, s, m) \cdot \mathbb{P}(a_j|t_1^{j-1}, a_1^{j-1}, s, m) = \mathbb{P}(a_j, f_j|t_1^{j-1}, a_1^{j-1}, s, m)$. Deduce that

$$\mathbb{P}(t, a, m|s) = \mathbb{P}(m|s) \prod_{i=1}^{|t|} \mathbb{P}(t_i|t_1^{i-1}, a_1^i, s, m) \cdot \mathbb{P}(a_i|t_1^{i-1}, a_1^{i-1}, s, m)$$

Exercise 47 : (Optional) At the station there are three payphones which accept 20p pieces. One never works, another always works, while the third works with probability 1/2. On my way to the metropolis for the day, I wish to identify the reliable phone, so that I can use it on my return. The station is empty and I have just three 20p pieces. I try one phone and it does not work. I try another twice in succession and it works both times. What is the probability that this second phone is the reliable one?

Exercise 48 : Parliament contains a proportion p of Party A members, who are incapable of changing their minds about anything, and a proportion $1 - p$ of Party B members who change their minds completely at random (with probability r) between successive votes on the same issue. A randomly chosen member is noticed to have voted twice in succession in the same way. What is the probability that this member will vote in the same way next time?

Exercise 49 [Polya Urn]: The Polya Urn model is as follows. We start with an urn which contains one white ball and one black ball. At each second we choose a ball at random from the urn and replace it together with one more ball of the same colour. Calculate the probability that when n balls are in the urn, i of them are white.

Markov Chains

Revision questions

1. Define a Markov Chain.
2. What does it mean for a Markov Chain to be periodic?
3. What does it mean for a Markov Chain to be irreducible?
4. What is a stationary distribution of a Markov Chains? Does it always exist?
5. How can you numerically compute the stationary distribution?
6. What are the detailed-balanced equations and how can you use them to compute the stationary distribution? Does it always work?
7. What are hitting times of a Markov Chain?
8. How can you numerically compute these hitting times (and variants)?

Exercises

This section is under construction

The following handouts have a lot of problems on Makov Chains (some out of the scope of this class):

- Handout 1
- Handout 2

When you solve one of these problems, try to think about computational aspects of the Markov Chain. Run simulations and see if these match the theoretical result.