TCS H1: On Statistical Inference and Modeling Data with the Ising Model

Clélia de Mulatier

Menu

1. **Short exercise: Modeling the activity of a single neuron.**
2. **Modeling binary data with the Ising model.**
   In this exercise, we will see how to use the tools seen in class in the context of the Ising model to model binary data, and learn more about statistical inference.
3. **Application to the analysis of the US supreme Court.**
   In this exercise, you will perform the statistical inference analysis introduced in exercise 2 to voting data from the US Supreme court. The aim will be to reproduce some of the results of the paper *Statistical mechanics of the US Supreme Court* by E. D. Lee, C. P. Broedersz, and W. Bialek [1].

General information

The homework is long, and can be consider as composed of two combined homework assignments (exercise 2 and 3) that you can submit at the same time. I advise you to start working on it as soon as possible.

Hand in the answers of your homework as a single pdf-file, including the figures. Ideally, it would be great if you can type your answers. Remember to **add titles and axis labels to all your graphs**. Please also upload the program that you used to create the graphs.

Regarding the programming bonus question:
write down the main line of the algorithm you have implemented (not the program itself) in the answer pdf-file. Additionally, please upload your program as a separate file. You can give back: a Python file, a Jupyter notebook, a C or C++ file, or a mathematica notebook, fortran is also ok. It is possible to send a matlab notebook, but only if you cannot work with another programming language.

1 Modeling the activity of a single neuron

Neuronal activity of a single neuron is often modeled by a homogeneous Poisson process with a *refractory period* of the order of a millisecond. This means that each time the neuron spikes, there is a waiting time τ_0 (of the order of a millisecond) during which the neuron cannot spike again.

In the attached dataset "`Data_neuron.txt`", you will find the neuronal activity of one neuron, i.e. the value of the successive times at which the neuron spikes. The time unit is the millisecond.

Q1. Can you plot the distribution $P(\tau)$ of the time intervals τ between successive spikes? Check that there is indeed a refractory period, i.e. a time interval τ_0 after each spike, during which the neuron doesn't spike again. What is the duration τ_0 of this time interval?

Q2. Can you check that the decay of the distribution $P(\tau)$ of inter-spike intervals is indeed exponential? Measure the corresponding decay rate λ.

Q3. Can you deduce an analytical expression for the distribution of inter-spike time interval $P(\tau)$ of the delayed Poisson process as a function of λ and τ_0? Compare your model distribution to the one obtained from the data.

Q4. Using your model, can you generate another 1000 (spike times) datapoints?

Q5. What is the average spiking rate f of the neuron in the data? How is f analytically related to τ_0 and λ that you have previously measured?

2 Modeling binary data with the Ising model

In general, datasets can contain more than a single variable, and variables can also be correlated with each other. In this exercise, we will see how to use the tools seen in class in the context of the Ising model to model binary data, and learn more about statistical inference.

Introduction

In class, we have discussed the use of the Ising model as a model for the ferromagnet-paramagnet phase transition. As simple as it is, the Ising model and its variants are widely used to model and study the behavior of complex systems in many contexts. The Ising model is also used in a completely different way to model binary data, and extract information and patterns from data.

Binary dataset as a spin system. A binary dataset is composed of many observations of binary variables. To model the data with an Ising model, we consider that the binary variables are spins that can take the values $+1$ or -1. We denote the variables by $s_i \in \{-1, +1\}$ and call them *spin variables.* The set of binary variables thus forms a spin system. A datapoints is a value of all the variables $\boldsymbol{s} = (s_1, \cdots, s_n)$, which corresponds to a state of the system. With n binary variables, there are 2^n possible different datapoints. A dataset is composed of many observations of this spin system. In many cases, the number N of datapoints is small compared to 2^n.

Example, the US Supreme court. The attached dataset "`US_SupremeCourt_n9_N895.txt`" contains the values of the votes of the 9 judges of the US Supreme court over 895 cases, which were used in this paper [1]. The original votes were just the answers, *yes* or *no*, for all the cases that were judged by the court. The authors of the paper then mapped each *yes/no* vote onto a *right/left* decision depending on the political orientation of the corresponding case. In the dataset available in canvas, the 1's correspond to conservative-oriented votes (right), while the 0's correspond to liberal-oriented votes (left). These values can easily be converted to ±1. For instance, for this homework we will **map all the** 0**'s to** -1**'s**, and therefore we will have the **conservative oriented votes labeled as** $+1$ and the **liberal oriented votes labeled as** -1.

2.1 Pairwise spin model

It is common to model the collective behavior of systems of binary variables with Ising-like models. To do so, we assume that the system is in a stationary state, and therefore that the datapoints are independently sampled from the same stationary probability distribution. We take this probability distribution to have the general form of an Ising model:

$$p_{\boldsymbol{g}}(\boldsymbol{s}) = \frac{1}{Z(\boldsymbol{g})} \exp\left(\sum_{i=1}^{n} h_i\, s_i + \sum_{pair(i,j)} J_{ij}\, s_i\, s_j \right) , \tag{1}$$

where n is the number of spin variables, where $pair(i, j)$ denotes a summation over all possible pairs of distinct spin variables, where $\boldsymbol{g} = (h_1, \cdots, h_n, J_{1,2}, \cdots, J_{n-1,n})$ is a vector of (real) parameters, and where $Z(\boldsymbol{g})$ is a normalization factor. There are several differences compared to the Ising model we have seen in class:

- there is a different external field h_i for each spin s_i, which can take any real value. In particular, the h_i's are not necessarily all positive or all negative.
- there is a different coupling parameter J_{ij} for each pair of spins s_i and s_j. The parameter J_{ij} parametrises the strength of the coupling between s_i and s_j, and can take any real value. In particular, the J_{ij}'s are not necessarily all positive or all negative.
- the summation is over all possible pairs (i, j) of spins, and not just over the "nearest neighbors". The reason is that, in a general dataset, we have a priori no idea if there exists an underlying structure between the variables and if so, what that structure is, and therefore we don't know which variables are "nearest neighbors".

The general goal of the problem is to infer the set of parameters $\boldsymbol{g} = (h_1, \cdots, h_n, J_{1,2}, \cdots, J_{n-1,n})$ that is the most appropriate to model the data, i.e. to find the parameters $\boldsymbol{g}$ for which the probability distribution in Eq. (1) best fits the data. This way, we would infer from the data the underlying structure between the variables, and find which variables tend to be strongly influenced by an external parameter, and which pairs of variables tend to be more strongly coupled.

Q1.1. How many terms are in the sum over the $pair(i, j)$? Can you deduce what is the number of parameters in the vector $\boldsymbol{g} = (h_1, \cdots, h_n, J_{1,2}, \cdots, J_{n-1,n})$? Can you re-write the sum over the $pair(i, j)$ as a double sum over i and j (without counting twice each pair)?

Q1.2. Can you write down explicitly the terms in the exponential of Eq. (1) for a system with $n = 3$ spins?

Q1.3. In Eq. (1), we can recognize the Boltzmann distribution, in which the parameter $\beta = 1/(k_b T)$ was taken equal to 1 (more precisely, the constant k_B was taken equal to 1, and the temperature parameter T was absorbed in the parameters h_i and J_{ij}). What is the energy function associated with the Boltzmann distribution in that case? What is the partition function and what is its general expression?

Q1.4. Take a spin s_i: if h_i is positive, which direction will s_i tend to turn to, i.e. which direction of s_i will minimize the associated energy $-h_i s_i$? Take a pair of spins s_i and s_j: if J_{ij} is positive, which configurations of (s_i, s_j) minimize the coupling energy $-J_{ij} s_i s_j$?
Assume that we have inferred the best parameters h_i and J_{ij} for the US supreme court dataset discussed in section 2. How would you interpret the sign of the inferred parameters h_i and J_{ij} in this context?

2.2 Observables

The important observables of the system are the local average magnetisations $\langle s_i \rangle$ and the local correlations $c_{ij} = \langle s_i s_j \rangle - \langle s_i \rangle \langle s_j \rangle$, where the angle brackets $\langle A(\boldsymbol{s}) \rangle$ denotes the ensemble average (or thermal average) of the microscopic quantity $A(\boldsymbol{s})$.

Q2.1. Given a stationary probability distribution of the state $p_{\boldsymbol{g}}(\boldsymbol{s})$, what are the definitions of $\langle s_i \rangle$ and of $\langle s_i s_j \rangle$?

Q2.2. Consider a dataset $\hat{\boldsymbol{s}}$ composed of N independent observations of the spins: $\hat{\boldsymbol{s}} = (\boldsymbol{s}^{(1)}, \cdots, \boldsymbol{s}^{(N)})$. Let us denote by $\langle s_i \rangle_D$ and $\langle s_i s_j \rangle_D$ the empirical averages of s_i and of $s_i s_j$ respectively (i.e., their average value in the dataset). How would you compute $\langle s_i \rangle_D$ and $\langle s_i s_j \rangle_D$ from the data?

Q2.3. Assume that the data is stationary and that each datapoint has been randomly sampled from $p(\boldsymbol{s})$. Can you show that the empirical averages, $\langle s_i \rangle_D$ and $\langle s_i s_j \rangle_D$, converge to the model averages, respectively $\langle s_i \rangle$ and $\langle s_i s_j \rangle$, as the number N of datapoints goes to infinity? (very large dataset)

2.3 Maximum Entropy models

In many papers, the authors refer to the generalised Ising model defined in Eq. (1) as a *maximum entropy model.* In this section, we will show that the probability distribution defined in Eq. (1) is indeed the (most general) probability distribution that maximizes the Shannon entropy $S[p_{\boldsymbol{g}}(\boldsymbol{s})]$ given a set of constraints (which we will specify).

Q3.1. Consider a spin system with stationary probability distribution $p(\boldsymbol{s})$. Can you recall the definition of the Shannon entropy $S[p(\boldsymbol{s})]$? As mentioned above for the Boltzmann distribution, we will take $k_B = 1$.

The Ising model in Eq. (1) can be seen as a *Maximum Entropy Model*, constrained to reproduce the data local magnetisation and local correlation, i.e. constrained to reproduce all the data averages $\langle s_i \rangle_D$ and $\langle s_i s_j \rangle_D$ (for all spin s_i and s_j). We also want $p(\boldsymbol{s})$ to be normalised, which introduces the additional constraint $\sum_{\boldsymbol{s}} p(\boldsymbol{s}) = 1$. To summarize, we are looking for the set of 2^n probabilities $p(\boldsymbol{s})$ such that $S[p(\boldsymbol{s})]$ is maximal, and such that:

$$\sum_{\boldsymbol{s}} p(\boldsymbol{s}) = 1 \qquad \text{and} \qquad \sum_{\boldsymbol{s}} p(\boldsymbol{s})\, s_i(\boldsymbol{s}) = \langle s_i \rangle_D \qquad \text{and} \qquad \sum_{\boldsymbol{s}} p(\boldsymbol{s})\, s_i(\boldsymbol{s}) s_j(\boldsymbol{s}) = \langle s_i s_j \rangle_D \tag{2}$$

where $\langle s_i \rangle_D$ and $\langle s_i s_j \rangle_D$ are constants that are computed from the data for all distinct s_i and s_j. Note that to be more precise, we wrote $s_i(\boldsymbol{s})$ (instead of just s_i) to specify that this is the value of s_i in the state $\boldsymbol{s}$ (this will help with the next questions).

Q3.2. How many constraints are there in total?

To find the shape of the distributions $p(\boldsymbol{s})$ that maximizes the entropy while satisfying these constraints, we introduce an auxiliary function:

$$\begin{aligned} U[p(\boldsymbol{s})] = S[p(\boldsymbol{s})] + \lambda_0 \left(\sum_{\boldsymbol{s}} p(\boldsymbol{s}) - 1 \right) + \sum_{i=1}^{n} \alpha_i \left(\sum_{\boldsymbol{s}} p(\boldsymbol{s})\, s_i(\boldsymbol{s}) - \langle s_i \rangle_D \right) \\ + \sum_{pair(i,j)}^{n} \eta_{ij} \left(\sum_{\boldsymbol{s}} p(\boldsymbol{s})\, s_i(\boldsymbol{s}) s_j(\boldsymbol{s}) - \langle s_i s_j \rangle_D \right) \end{aligned} \tag{3}$$

where we have introduced a parameter in front of each constraint we want to impose. These parameters (λ_0, α_i and η_{ij}) are called Lagrange multipliers. To find $p(\boldsymbol{s})$ one must maximize this auxiliary function with respect to the 2^n probabilities $p(\boldsymbol{s})$.

Q3.3. Let us fix a choice of a state $\boldsymbol{s}$. The probability $p_{\boldsymbol{s}} = p(\boldsymbol{s})$ is a parameter of $U[\boldsymbol{p}]$, where $\boldsymbol{p}$ is the vector of the 2^n probabilities. Can you show that:

$$\frac{\partial U[\boldsymbol{p}]}{\partial p_{\boldsymbol{s}}} = -\log p_{\boldsymbol{s}} - 1 + \lambda_0 + \sum_{i=1}^{n} \alpha_i\, s_i(\boldsymbol{s}) + \sum_{pair(i,j)} \eta_{ij}\, s_i(\boldsymbol{s}) s_j(\boldsymbol{s}) \ \ ? \tag{4}$$

Q3.4. Can you show that the most general expression of $p_{\boldsymbol{s}}$ with maximal entropy that satisfying the constraints in Eq. (2) is Eq. (1)? Give the relation between λ_0 and the partition function Z. How are the parameters α_i and η_{ij} related to the parameters h_i and J_{ij}?

2.4 Statistical inference: model with no couplings

Consider the model with no couplings (all the $J_{ij} = 0$):

$$p_{\boldsymbol{g}}(\boldsymbol{s}) = \frac{1}{Z(\boldsymbol{g})} \exp\left(\sum_{i=1}^{n} h_i\, s_i \right) . \tag{5}$$

The vector $\boldsymbol{g} = (h_1, \cdots, h_n)$ now only contains n local field parameters.

Q4.1. Can you show that in that case the model is assuming that the variables are independent from each other, i.e. that we can write the joint probability distribution as a product of a probability distribution over each variable: $p_{\boldsymbol{g}}(\boldsymbol{s}) = \prod_{i=1}^{n} p_{\boldsymbol{h_i}}(s_i)$. What is the probability distribution $p_{\boldsymbol{h_i}}(s_i)$ for the spin variable s_i?

Q4.2. Take one of the spin variable s_i. We recall that $\langle s_i \rangle_D$ is the average value of s_i in the data (given a dataset, this quantity is a constant), and that $\langle s_i \rangle = \sum_{\boldsymbol{s}} p(\boldsymbol{s})\, s_i$ is the model average of s_i. Can you show that the value of the parameter h_i that satisfies the constraint $\langle s_i \rangle = \langle s_i \rangle_D$ is:

$$h_i = \tanh^{-1}(\langle s_i \rangle_D)\,, \tag{6}$$

where $\tanh^{-1}(x)$ denotes the inverse of the hyperbolic tangent? In particular, in that case the probability distribution over s_i in the model is exactly equal to the empirical distribution of s_i.

Q4.3. In Eq. (6), we observe that:

- if $\langle s_i \rangle_D > 0$, then the inferred h_i is also positive;
- reciprocally, $\langle s_i \rangle_D < 0$, then the inferred h_i is also negative.

How does this connect with the tendency of the i-th judge to vote on average more liberal or more conservative? Is this result coherent with the general comments that we did in Question Q1.4.?

2.5 Statistical inference: maximizing the log-likelihood function

Introducing the likelihood function. Looking more closely at Eq. (1), one can see that it does not just define a single probability distribution, but many of them: there is one probability distribution for each value of the set of parameters $\boldsymbol{g}$. More precisely, the distribution in Eq. (1) changes continuously as one continuously varies the parameters in $\boldsymbol{g}$. We say that Eq. (1) defines a *parametric family of probability distributions*. The inference procedure consists in finding the value of the parameters $\boldsymbol{g}$ that maximizes the probability that the model $p_{\boldsymbol{g}}(\boldsymbol{s})$ produces the data.

To do so, we introduce the *log-likelihood function*:

$$\mathcal{L}(\boldsymbol{g}) = \log P_{\boldsymbol{g}}(\hat{\boldsymbol{s}})\,, \tag{7}$$

where $P_{\boldsymbol{g}}(\hat{\boldsymbol{s}})$ is the probability that the model $p_{\boldsymbol{g}}(\boldsymbol{s})$ produces the dataset $\hat{\boldsymbol{s}} = (\boldsymbol{s}^{(1)}, \cdots, \boldsymbol{s}^{(N)})$. Note that $\mathcal{L}(\boldsymbol{g})$ is a function of the parameters $\boldsymbol{g}$. The inference procedure therefore consists in finding the value $\boldsymbol{g}^\star$ of the parameters that maximizes $\mathcal{L}(\boldsymbol{g})$. For the moment we will assume that there exists only a unique such value of $\boldsymbol{g}$.

Q5.1. We assuming that, in the dataset $\hat{\boldsymbol{s}}$, all the datapoints are independently sampled from an underlying distribution $p_{\boldsymbol{g}}(\boldsymbol{s})$. Can you show that the log-likelihood function can be re-written as:

$$\mathcal{L}(\boldsymbol{g}) = N \sum_{\boldsymbol{s}} p_D(\boldsymbol{s}) \log p_{\boldsymbol{g}}(\boldsymbol{s})\,, \tag{8}$$

where $p_D(\boldsymbol{s})$ is the empirical distribution over the states? The empirical distribution is given by $p_D(\boldsymbol{s}) = K(\boldsymbol{s})/N$ where $K(\boldsymbol{s})$ is the number of times that the datapoint $\boldsymbol{s}$ occurs in the dataset.

Ising model. We now take the model distribution $p_{\boldsymbol{g}}(\boldsymbol{s})$ to be given by the Ising model in Eq. (1).

Q5.2. Taking the first derivative of $\mathcal{L}(\boldsymbol{g})$ with respect to a parameter h_i, can you show that at the maximum of $\mathcal{L}(\boldsymbol{g})$ we have that $\langle s_i \rangle = \langle s_i \rangle_D$? Similarly, taking the first derivative of $\mathcal{L}(\boldsymbol{g})$ with respect to a parameter J_{ij}, can you show that at the maximum of $\mathcal{L}(\boldsymbol{g})$ we have that $\langle s_i s_j \rangle = \langle s_i s_j \rangle_D$?

3 Application to the analysis of the US supreme Court

A system with n spin variables can be in 2^n different states. However, most of the times, the number of different states observed in a dataset is very small compared to 2^n.

Q6.1. For the US Supreme court dataset: What is the number n of spin variables, and the total number 2^n of states that can be observed for that system? What is the total number N of datapoints in the datasets? What is the number N_{max} of different states that are observed?

Q6.2. (Bonus question) Numerical solution: For the dataset provided, find numerically the value of the parameters $\boldsymbol{g}$ of the fully connected pairwise model Eq.(1) that maximizes the log-likelihood function $\log L(\boldsymbol{g})$. What are the main computational limitations of your algorithm? How can you improve it?

In Canvas, the files `hi_ussc_unsorted.txt` and `Jij_ussc_unsorted.txt` contain the values of the fitted parameters h_i and J_{ij} for the US supreme court dataset. These are formatted in sequential order, i.e. `hi_ussc_unsorted.txt` contains the local field parameters in the following order:

```
h1
h2
h3
..
h9
```

where `hi` is the value of the fitted local field on the variable s_i, and `Jij_ussc_unsorted.txt` contains the pairwise couplings:

```
J12
J13
..
J19
J23
..
J29
J34
..
J78
J89
```

Please note that only unique couplings are included, i.e. J_{12} is included, but J_{21} is not, since these are redundant (therefore, there are $9 \times 8/2 = 36$ values in `Jij_ussc_unsorted.txt`). The order of these couplings corresponds to the order of the spins (judges) in the given dataset.

Q6.3. We would like to reproduce **Figure 13** of the paper [1]. Can you plot the $\langle s_i \rangle_D$ as a function of i? Can you re-order the label i so that the values of $\langle s_i \rangle_D$ are ordered from the smallest (negative value) to the largest (positive value), as in Fig. 13.A top? Keeping the new ordering, can you plot a heatmap of the matrix of $\langle s_i s_j \rangle_D$ (see Fig. 13.A bottom)? What can we say about the judges with negative $\langle s_i \rangle_D$? with positive $\langle s_i \rangle_D$? Can you comment on these plots?

Q6.4. Keeping the new ordering of the labels i, can you plot a heatmap of the fitted vector of h_i's and a heatmap of the fitted matrix of J_{ij}'s, as in Fig. 13.B? You can use the fitted values of h_i and J_{ij} provided in Canvas. Can you comment on these plots?
Note that the values of h_i and J_{ij} that are provided in Canvas are following the same order of the variables s_i than in the USSC datafile `US_SupremeCourt_n9_N895.txt` (i.e. the original ordering of the judges).

Q6.5. Scatter plot 1, "Cross-validation": for all the states observed in the data, can you plot the empirical probability of the state, $p_D(\boldsymbol{s})$ against the model probability $p_{\boldsymbol{g}}(\boldsymbol{s})$ with the fitted parameters (h_i and J_{ij})? What can you say from this plot?

Q6.6. Scatter plot 2: checking the fit: for all the spin s_i, can you plot the value of $\langle s_i \rangle_D$ in the data against the value $\langle s_i \rangle$ in the fitted model? for all the pairs of spins s_i and s_j, can you plot the value of $\langle s_i s_j \rangle_D$ in the data against the value $\langle s_i s_j \rangle$ in the fitted model? What can you say from these plots?

One of the question addressed in the paper is: can the pairwise model reproduce higher order patterns of the data better than a model of independent judges? To answer this question, the authors introduce the probability $P(k)$ that there are k-conservative votes as answer to a case (i.e. the probability that a datapoint contains k conservative votes).

Q6.7. Consider a model with no coupling, in which judges vote independently from each others. Each judge s_i has a probability $p_i(+1)$ to vote conservative. In that case, what is the probability $P_I(k)$ that k-judges vote conservative? Note: we added the label "I" to $P(k)$ to specify that it is the probability distribution obtained for an independent model.

In the dataset, how many judges have votes that are more conservative on average? At which value of k do you then expect the maximum of $P_I(k)$ to be? Can you obtain the values of $p_i(+1)$ from the data and plot the values of $P_I(k)$ as a function of k for the independent model?

Q6.8. Let us call $P_D(k)$ the probability distribution $P(k)$ obtained directly from the data. How can you compute the values of $P_D(k)$ from the data for $k = 1$? for $k = 2$? for any k? Can you plot $P_D(k)$ as a function of k and compare the curve to the one obtained previously for the independent model? Where is the maximum of $P_D(k)$? Is the independent model a good model for the US supreme court data?

Q6.9. Let us call $P_P(k)$ the probability distribution $P(k)$ obtained from the fitted Ising model with pairwise couplings (i.e., with the model $p_{\boldsymbol{g}}(\boldsymbol{s})$ in Eq. (1) with the fitted parameters h_i and J_{ij} provided in Canvas). How can you compute $P_P(k)$: can you write the analytical expression of $P_P(k)$ as a function of $p_{\boldsymbol{g}}(\boldsymbol{s})$? Can you plot $P_P(k)$ as a function of k and compare the curve to $P_D(k)$ and $P_I(k)$ previously obtained? (see Figure 16.A of the paper) Which conclusions can you draw?

4 Bonus questions: Fisher Information Matrix and multi-dimensional fluctuation-dissipation theorem

Please ask me if you have finished the rest of the homework and are interested in these bonus questions.

References

[1] Edward D Lee, Chase P Broedersz, and William Bialek. Statistical mechanics of the us supreme court. *Journal of Statistical Physics*, 160(2):275–301, 2015.