Probability & Statistics 2

Robert Šámal

November 10, 2022

1 Markov Chains

This is work in progress, likely full of typos and small (or larger) inprecisions. If you find something that look wrong, feel free to reach out to me. Thanks to Kryštof Višňák for helping me with spotting typos in a previous version.

1.1 Introduction, basic properties

Two examples to start with A machine can be in two states: working or broken. For simplicity, we assume that the state stays the same for the whole day. Then, during the night, the state changes at random according to the figure below: for instance, if the machine is working one day, it will work the next day with probability 0.99, with probability 0.01 it breaks over night. Crucially, we assume that this probability does not depend on the age of the machine, nor on the previous states.

A fly is moving in a corridor, that we consider as a collection of four spaces, labeled 0, 1, 2, 3. If the fly is in spaces 1 or 2, it stays at the same space with probability 0.4. Otherwise, it moves equally likely one step left or right. At positions 0 and 3 is a spider and the fly can never leave. Again, we assume that "the fly has no memory", so the probabilities do not depend on the past trajectory of the fly.

TODO: add figures

What are the common features of these examples? We consider a sequence of random variables, so called *random process*. We do not care about the numerical value of these variables, as we consider them as mere labels – so we will not ask about expected value of a position of the fly, for instance. We may assume that all the random variables have range contained in set S of labels. For simplicity we assume S to be finite or countable (and frequently we will assume that $S = \{1, \dots, s\}$ or $S = \mathbb{N}$). We also want to prescribe *transition probabilities* $p_{i,j}$ such that $P(X_{t+1} = j \mid X_t = i) = p_{i,j}$. However, there is more subtlety to this: we want to explicitly forbid the history (values of $X_0, \dots, X_{t-1}$ to have an influence on X_{t+1}.

Definition 1 (Markov chain)**.** *Let S be any finite or countably infinite set. A sequence $(X_t)_{t=0}^{\infty}$ of random variables with range S is a (discrete time, discrete space, time-homogeneous)* Markov chain *if for every $t \ge 0$ and every $a_0, \dots, a_{t+1} \in S$ we have*

$$P(X_{t+1} = a_{t+1} \mid X_t = a_t \,\&\, \dots \,\&\, X_0 = a_0) = P(X_{t+1} = a_{t+1} \mid X_t = a_t) = p_{a_t, a_{t+1}},$$

for some collection of transition probabilities $p_{i,j}$. The condition is only required when the conditional probabilities are defined, that is when $P(X_t = a_t \,\&\, \ldots \,\&\, X_0 = a_0) > 0$.

TODO: explain alternatives

Transition matrix is a matrix P such that $P_{i,j} = p_{i,j}$, that is the entry at i-th row and j-th column is the probability of transition from state i to state j. As a consequence of the definition, all entries in the transition matrix are nonnegative, and each row sums to 1. We can describe this succintly by writing $Pj = j$ with j denoting the column vector all 1's.

Let P denote the transition matrix for the machine example and Q for the fly example. We have

$$P = \begin{pmatrix} 0.99 & 0.01 \\ 0.9 & 0.1 \end{pmatrix} \qquad Q = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.3 & 0.4 & 0.3 & 0 \\ 0 & 0.3 & 0.4 & 0.3 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Transition graph/diagram is a directed graph with vertex set S and arcs (directed edges) (i, j) *for every* $i, j \in S$ *such that* $p_{i,j} > 0$. We label arc (i, j) by $p_{i,j}$. In other words, the figures above (TODO) are transition graphs.

Describing the distribution. We will again use the basic tool to describe a random variable, namely a PMF (probability mass function), that is giving a probability of each state (element of S). A common notation is

$$\pi_i^{(t)} = P(X_t = i).$$

For any $t \geq 0$ we also consider $\pi^{(t)}$ as a row vector with coordinates $\pi_i^{(t)}$ for $i \in S$.

Transition of the distribution Suppose we know $\pi^{(0)}$, what can we say about $\pi^{(1)}$, and $\pi^{(t)}$ in general? By law of total probability we have

$$P(X_1 = j) = \sum_{i=1}^{s} P(X_0 = i) \cdot P(X_1 = j \mid X_0 = i) \qquad \text{So, in other notation}$$

$$\pi_j^{(1)} = \sum_{i=1}^{s} \pi_i^{(0)} \cdot P_{i,j} \qquad \text{and using matrix multiplication:}$$

$$\pi^{(1)} = \pi^{(0)} P$$

From this we easily get the following theorem:

Theorem 2. *For any Markov chain and any $k \geq 0$ we have*

$$\pi^{(k)} = \pi^{(0)} P^k$$

and, more generally, $\pi^{(t+k)} = \pi^{(t)} P^k$.

Proof. By induction. TODO □

k-step transition To look at the above theorem in different way, we define the following:

$$\begin{aligned} r_{i,j}(k) &= P(\text{we get from } i \text{ to } j \text{ in } k \text{ steps}) \\ &= P(X_k = j \mid X_0 = i) \end{aligned}$$

As we will se, we also have $r_{i,j}(k) = P(X_{t+k} = j \mid X_t = i)$ for any $t > 0$, but this remains to be seen, there may be a dependency on t.

Theorem 3 (Chapman–Kolmogorov)**.** *For any Markov chain and any $k, \ell \geq 0$ we have*

- $r_{i,j}(k) = (P^k)_{i,j}$
- $r_{i,j}(k + \ell) = \sum_{u=1}^{s} = r_{i,u}(k) r_{u,j}(\ell)$
- $r_{i,j}(k + 1) = \sum_{u=1}^{s} = r_{i,u}(k) p_{u,j}$

1.2 Classification of states

Definition 4 (Accessible states)**.** *For states i, j of a Markov chain we say that j* is accessible from i*, if starting at i we have nonzero probability of reaching j in the future. For short we write $j \in A(i)$ or $i \to j$. In formula:*

$$j \in A(i) \Leftrightarrow P((\exists t \geq 0) X_t = j | X_0 = i) > 0.$$

It is easy to observe (TODO) that $j \in A(i)$ is equivalent with existence of a directed path from i to j in the transition graph.

Definition 5 (Commuting states)**.** *We say that states i, j of a Markov chain* commute *if $i \in A(j)$ and $j \in A(i)$. For short we write $i \leftrightarrow j$.*

Theorem 6. *For any Markov chain the relation $\leftrightarrow$ is an equivalence on the set of states.*

1.3 Convergence to stationary distribution

Chapman-Kolmogorov theorem gives us a way how to describe the behaviour of a Markov chain in a short time: If we start with known $\pi^{(0)}$ (distribution if X_0, the state at time 0), we can compute $\pi^{(k)}$. Next, we turn to describing the long-term behaviour.

transient vs. recurrent states

convergence to stationary distribution

1.4 Probability of absorption, time to absorption

Yet another way to look at long-term behaviour of a Markov chain is to study *absorbing states*, states that can never be left. Formally, $a \in S$ is absorbing state if $p_{a,a} = 1$. Not every Markov chain has such state, but for those that do, two natural questions arise: how long (on average) will it take till we reach an absorbing state? And if there is more than one such state, what is the probability of reaching each of them? Both questions are easily answers, if one approaches it right: it is significantly easier to compute these times and probabilities for all states at the same time, than do to it just for one state.

In the following, assume $A \subseteq S$ is a nonempty set of absorbing states; also assume $0 \in A$. For every $i \in S$ we define μ_i to be the expected time to absorption starting from i, formally

$$\mu_i = \mathbb{E}(T \mid X_0 = i), \text{where } T = \min\{t : X_t \in A\}.$$

Further, we let a_i be the probability we end at state 0, starting from i.

$$a_i = P(\exists t : X_t = 0 \mid X_0 = I).$$

Here we tacitly assume that A contains more absorbing states than just 0, otherwise $a_i = 1$.

Theorem 7. *The probabilities a_i are the unique solution to the following system of equations:*

$$\begin{aligned} a_0 &= 1 \\ a_i &= 0 && \textit{for } 0 \neq i \in A \\ a_i &= \sum_{j \in S} p_{i,j} a_j && \textit{otherwise}. \end{aligned}$$

TODO: proof simple by law of total probability.

Theorem 8. *The expected times μ_i are the unique solution to the following system of equations:*

$$\begin{aligned} \mu_i &= 0 && \textit{for } i \in A \\ \mu_i &= 1 + \sum_{j \in S} p_{i,j} \mu_j && \textit{otherwise}. \end{aligned}$$

TODO: proof simple by law of total expectation.

Example: random walk on a path

1.5 Application: algorithm for 2-SAT, 3-SAT

2 Bayesian statistics

2.1 Two approaches to statistics

In the first semester we looked at the *classical (frequentists')* approach to statistics. In this approach:

- Probability is a long-term frequency (out of 6000 rolls of the dice, a six was rolled 1026 times, the ratio converges to the true probability). It is an objective property of the real world.
- Parameters are fixed, unknown constants. We can't make meaningful probabilistic statements about them.
- We design statistical procedures to have desirable long-run properties. E.g. 95 % of our interval estimates will cover the unknown parameter.

Now we are going to look at an alternative, so called *Bayesian approach*:

- Probability describes how much we believe in a phenomenon, how much we are willing to bet:
  (Prob. that T. Bayes had a cup of tea on December 18, 1760 is 90 %.)
  (Prob. that COVID-19 virus did leak from a lab is ?50? %.)
- We can make probabilistic statements about parameters (even though they are fixed constants): the "choice of universe" is the underlying elementary event.
- We compute the distribution of ϑ and form point and interval estimates from it, etc.

2.2 Bayesian method – basic description

- The unknown parameter is treated as a random variable Θ
- We choose *prior distribution*, the pmf $p_\Theta(\vartheta)$ or the pdf $f_\Theta(\vartheta)$ independent of the data.
- We choose a statistical model $p_{X|\Theta}(x|\vartheta)$ or $f_{X|\Theta}(x|\vartheta)$ that describes what we measure (and with what probability), depending on the value of the parameter.
- After we observe $X = x$, we compute the *posterior distribution* $f_{\Theta|X}(\vartheta|x)$
- and then derive what we need e.g. find a, b so that $P(a \le \Theta \le b \mid X = x) = \int_a^b f_{\Theta|X}(\vartheta|x)d\vartheta \ge 1 - \alpha$

2.3 Bayes theorem

Theorem 9 (Bayes theorem for discrete r.v.'s)**.** *X, Θ are discrete r.v.'s*

$$p_{\Theta|X}(\vartheta|x) = \frac{p_{X|\Theta}(x|\vartheta)p_\Theta(\vartheta)}{\sum_{\vartheta' \in Im\Theta} p_{X|\Theta}(x|\vartheta')p_\Theta(\vartheta')}.$$

(terms with $p_\Theta(\vartheta') = 0$ are considered to be 0).

Theorem 10 (Bayes theorem for continuous r.v.'s)**.** *X, Θ are continuous r.v.'s with pdf's f_X, f_Θ and joint pdf $f_{X,\Theta}$*

$$f_{\Theta|X}(\vartheta|x) = \frac{f_{X|\Theta}(x|\vartheta) f_\Theta(\vartheta)}{\int_{\vartheta' \in \mathbb{R}} f_{X|\Theta}(x|\vartheta') f_\Theta(\vartheta') d\vartheta'}.$$

(terms with $f_\Theta(\vartheta') = 0$ with $f_\Theta(\vartheta') = 0$ are considered 0).

Theorem 11 (Bayes theorem for discrete r.v.'s)**.** *X be discrete and Θ continuous r.v. Then*

$$f_{\Theta|X}(\vartheta|x) = \frac{p_{X|\Theta}(x|\vartheta) f_\Theta(\vartheta)}{\int_{\vartheta' \in Im\Theta} p_{X|\Theta}(x|\vartheta') f_\Theta(\vartheta')}.$$

(terms with $p_\Theta(\vartheta') = 0$ are considered to be 0).

2.4 Bayesian point estimates – MAP and LMS

Even when we know a distribution of a random variable it is unclear what is the best numerical value that represents it. is it the mean (expected value)? Or the mode (moste probable value)? Or the median? It turns out all choices have their justification. In the context of Bayesian statistics, we are interested in a random variable Θ conditioned on the event $X = x$. (You may concentrate on the discrete case, where the conditioning is easy to understand.)

MAP – Maximum A-Posteriori We choose $\hat{\vartheta}$ to maximize

- $p_{\Theta|X}(\vartheta|x)$ in the discrete case
- $f_{\Theta|X}(\vartheta|x)$ in the continuous case
- Essentially, we are replacing the random variable by its mode.
- Similar to the ML method in the classical approach if we choose a "flat prior" – Θ is supposed to be uniform/discrete uniform.

LMS – Least Mean Square Also the conditional mean method.

- We choose $\hat{\vartheta} = \mathbb{E}(\Theta \mid X = x)$, so we replace the random variable by its mean.
- What we get is an Unbiased point estimate that has the smallest possible LMS (least mean square) error:

$$\mathbb{E}((\Theta - \hat{\vartheta})^2 | X = x)$$

- (we will show this later.)

Similarly, if we take median (number m such that $P(\Theta \leq m \mid X = x) = 1/2$) then we minimize absolut value of an error $\mathbb{E}((\Theta - \hat{\vartheta})^2 | X = x)$. But we will not pursue this approach further.

2.5 Bayesian inference – examples

2.5.1 Naive Bayes classifier – both Θ and X are discrete

This techniques can be used for any classification of objects into finite number of categories, using finite number of discrete features. For concreteness, we will explain it as a way to test whether some email is a spam or ham (that is, not spam). We let Ω be the set of all emails (together with the probability of receiving each of them). We can't possibly list of elements of Ω, but we consider the emails delivered to our inbox as sampling from this probability space.

Our interest lies in random variable Θ that is equal to 1 for spams and to 2 for hams. (Recall Θ is a function from Ω to $\mathbb{R}$, so for each email $\omega \in \Omega$ we need to define value of $\Theta(\omega)$.) In order to estimate value of Θ, we measure data: a list of Bernoulli variables $X_1, \dots, X_n$, where $X_i(\omega) = 1$ if ω contains word w_i (and $X_i(\omega) = 0$ otherwise). So we imagine $w_1, \dots, w_n$ is a list of all words that are useful to detect spams.

By the Bayes theorem we have

$$p_{\Theta|X}(\vartheta|x) = \frac{p_{X|\Theta}(x|\vartheta)p_\Theta(\vartheta)}{\sum_{t=1}^2 p_{X|\Theta}(x|t)p_\Theta(t)}.$$

TODO finish it

2.5.2 Estimating bias of a coin – Θ is continuous, X is discrete

Consider a loaded coin with probability of heads being ϑ (which we assume to be an evaluation of a random variable Θ). Btw, everything applies to any procedure generating a Bernoulli random variable, but we stick with a coin for concreteness. Our goal is to find out the value of ϑ. In tune with the Bayesian methodology, we start with a prior distribution, that is a pdf f_Θ. (As we want to allow any real number in $[0, 1]$ as the value of ϑ, we must take Θ to be a continuous random variable.) Then we take measurements: we choose a number n of coin tosses and check how many heads we get. If we know the value of θ, the distribution of this number (call it X), is clearly $Bin(n, \vartheta)$. So we get

$$p_{X|\Theta}(k|\vartheta) = \binom{n}{k}\vartheta^k(1-\vartheta)^{n-k}.$$

It remains to apply Theorem 11. We still haven't decided what prior to choose though. If we don't known anything (say it is not a real coin but a digital generator), we may take *flat prior* $\Theta \sim U(0, 1)$. However, we need something more versatile to allow us to encode some prior knowledge.

Beta distribution It is convenient to use the following type of distribution for Θ:

$$f_\Theta(\vartheta) = \begin{cases} c\vartheta^{\alpha-1}(1-\vartheta)^{\beta-1} & \text{for } 0 < \vartheta < 1 \\ 0 & \text{otherwise} \end{cases}$$

Here c is a normalizing constant that makes the following function a pdf. It is typically written as $1/B(\alpha, \beta)$, the reciprocal of a *beta function*. The r.v. Θ is said to have *beta distribution*. We will collect some useful properties of this distribution. All are easy to verify using basic knowledge of calculus, details are omitted though.

- $f_\Theta(\vartheta)$ is maximal for $\vartheta = \frac{\alpha-1}{\alpha+\beta-2}$ (mode of the distribution). This can be verified by a simple differentiation.
- $\mathbb{E}(\Theta) = \frac{\alpha}{\alpha+\beta}$ (mean of the distribution). This follows from the next part and easy calculation.
- $B(\alpha, \beta) = 1/\binom{\alpha+\beta-2}{\alpha-1}$. This can be shown by per-partes and induction over $\alpha + \beta$.

Now we have all set up to apply Theorem 11. Fortunately, we don't need to compute the integral in the denominator.

$$\begin{aligned} f_{\Theta|X}(\vartheta|k) &= c_1 p_{X|\Theta}(k|\vartheta) f_\Theta(\vartheta) \\ &= c_2 \vartheta^k (1-\vartheta)^{n-k} \vartheta^{\alpha-1} (1-\vartheta)^{\beta-1} \\ &= c_2 \vartheta^{\alpha+k-1} (1-\vartheta)^{\beta+n-k-1} \end{aligned}$$

The calculation is only valid for $\vartheta \in [0, 1]$, otherwise $f_\Theta(\vartheta) = 0$, so the updated (posterior) pdf is also 0. How to find out c_2, if we need to? We use the fact that after conditioning on the event $\{X = k\}$ the random variable Θ still only attains values in $[0, 1]$. Thus, c_2 takes such value that makes $f_{\Theta|X}(\vartheta|k)$ a pdf, a function with integral 1. Based on what we learned about Beta distribution, $c_2 = 1/B(\alpha', \beta')$ and $\Theta|X = k$ follows the Beta distribution with parameters $\alpha' = \alpha+k$ and $\beta' = \beta+n-k$.

TODO: wrap up

2.5.3 Estimating normal random variables – both Θ and X are continuous