Probabilistic graphical models

Siddharth Bhat

Spring 2020

Contents

Chapter 1

Background and aims

Consider a distribution of binary random variables $x_1, x_2, \ldots x_n, y$. Note that to define the value of $P(x_1)$, we need just one value: $P(x_1 = 0)$. We can derive $P(x_1 = 1) \equiv 1 - P(x_1 = 0)$.

However, the full joint distribution $P(x_1, x_2, \ldots, x_n, y)$ needs $2^{n+1} - 1$ values to fully define.

However, let us assume that $P(x_i|y)$ are all independent. Hence, we can rewrite the above distribution as $P(y) \prod_{i=1}^{n} P(x_i|y)$. Now, we need to know $P(x_i|y = 0), P(x_i|y = 1)$. Both of these are binary random variables which need one value to define. So in toto, we need $2n + 1$ values for the above (factored) joint distribution.

So, we will study how to represent, perform inference, and perfom bayesian updates (learning). Also, connections to boltzmann distributions and whatnot will be explored. Connections to graph theory as well. We are also going to study MCMC (Markov chain monte carlo) methods. I hope we study more than just metropolis hastings: I want to understand Hamiltonian and Lavengin Monte Carlo more deeply (NUTS sampling, slice sampling, their interactions with HMC, etc). Later, we will see some connections to Learning theory (PAC learning - defined by Valiant).

The textbook is "Kohler and Friedman".

1.0.1 A teaser problem

We start with an ordered deck. We propose a shuffling mechanism: take the top card and move it to somewhere in the deck. Eg. If we start form $(1, 2, 3)$, we can move this to $(2, 1, 3)$, or $(2, 3, 1)$. Now, when the card 3 comes to the top, note that we had placed all other numbers in the deck with uniform probability. So, when the card 3 comes to the top, all the other cards are uniformly distributed. We now need to place 3 uniformly in the deck.

Let T_1 be the random variable of the first round at which a single card is placed *underneath* n.

There are $n - 1$ slots where can place any top card, so the likelihood of hitting the bottom slot is $1/(n - 1)$.

$$
\begin{aligned}
P(T_1 = 1) &\equiv \frac{1}{n-1} \\
P(T_1 = 2) &\equiv \left(1 - \frac{1}{n-1}\right) \frac{1}{n-1} = \frac{n-2}{n-1} \\
P(T_1 = i) &\equiv (1 - P(T_1 = i-1)) \frac{1}{n-1} = \left(1 - \frac{1}{n-1}\right)^{i-1} \frac{1}{n-1} = \frac{(n-2)^{i-1}}{(n-1)^i}
\end{aligned}
$$

This is a geometric distribution with parameter $\frac{1}{n-1}$. The expectation is going to be $\mathbb{E}[T_1] \equiv n-1$.

We now define T_2 to be the random variable which is the time from when the first card went underneath the nth card, to when the second card went underneath the nth card. We have two locations at the bottom. Eg. if we had $(1,2,3,4)$ to start with, and after T_1, we are now at $(2,3,4,1)$. We now have two positions $(2,3,4,\circ,1,\circ)$ to be underneath the card 4.

$$
\begin{aligned}
P(T_2 = 1) &\equiv \frac{2}{n-1} \\
P(T_2 = i) &\equiv \left(1 - \frac{2}{n-1}\right)^{i-1} \frac{2}{n-1}
\end{aligned}
$$

This is a geometric distribution with parameter $\frac{2}{n-1}$. The expectation is going to be $\mathbb{E}[T_1] \equiv n-2$.

The total time for the nth card to reach the top is going to be $T \equiv T_1 + T_2 + \cdots + T_n$. So the expectation is going to be $\mathbb{E}[T] = \sum_i \mathbb{E}[T_i] = \sum_i \frac{1}{n-i}$

1.0.2 Another problem

There are three balls, numbered 1, 2, 3, and there are three numbered bins. We throw the first ball into each of the three bins with equal probability. Independently, throw the second ball and the third ball to each of the three bins Independently.

Let X be the number of balls in the first balls in the first bin. Let N be the number of non-empty bins.

Write down $P(X), P(N), P(X,N)$ where $P(X,N)$ is the joint distribution. Also find $P(X|N)$, $P(N|X)$.

Chapter 2

Review of Probability

The first thing we should know is the sample space Ω. Next, we care about events $E \subseteq \mathcal{P}(\Omega)$. The probability is a function $P : \Omega \to \mathbb{R}_{\geqslant 0}$ such that:

- $P(\Omega) = 1$
- if $A \cap B = \emptyset$, then $P(A \cap B) = P(A) + P(B)$

A random variable is a function $X : \Omega \to \mathbb{R}$.
The expectation of a random variable $\mathbb{E}[X] \equiv \sum_i i \cdot P(X = i)$. That is, it is the average.
Markov's inequality says that $\mathbb{P}[X > a] \leqslant \mathbb{E}[X]/a$.
Chebyshev's inequality says that $\mathbb{P}[|X - \mu| > a] \leqslant Var(X)/a^2$.

2.1 Conditional Probability

To control to sample space and smooth interpolate probabilities a we restrict to a smaller sample space, we use *conditioning*. Hence we define the conditional probability: $P(B|A) \equiv P(B \cap A)/P(A)$

We can have a chain of conditioning:

$$P(A \cap B \cap C) = P(A)P(B \cap C|A) = P(A)P(B|A)P(C|A \cap B)$$

Example 1 *We have 3 bins and 4 balls. We throw the 4 balls into the 3 bins independently. Each ball will land into some bin.*

Let X *be the number of balls in bin 1.* N *be the number of non-empty bins.*

The total number of possibilities is 3^4*, since each ball has* 3 *possible bins, and there are* 4 *independent balls.*

$P(X = 2 \wedge N = 1) = 0$*. We have have only two balls in bin* 1 *which occupies one bucket. Hence,the other two balls need another bin.* N *can be at minimum* 2*.*

$$P(X = 2 \wedge N = 2) = \textit{ways to send 2 balls to 1st bin} \cdot \textit{leftover bin to send the leftover 2 balls} = \frac{\binom{4}{2} \cdot 2}{3^4}$$

Example 2 *We have a drunken man who takes a step forward with* $p = 1/2$*, and 2 steps backward with probability* $p = 1/2$*. at time* $t = 0$*,* $x = 0$*. Let* T *be the time of passing out. Let* pos_T *be the set of positions man could have been in when he passes out at time* T*.*

For example, at $T = 0, \mathrm{pos}_0 = \{0\}$*. At* $T = 1, \mathrm{pos}_1 = \{-2, 1\}$*. At* $T = 2, \mathrm{pos}_2 = \{-4, -1, 2\}$*.* $\mathrm{pos}_3 = \{0, \dots\}$*.*

Let Y *be a random variable such that* Y *is* 1 *with probability* 0.5*, and* -2 *with probability* 0.5*. Let* X *be the random variable that is the position after* T *steps. For example,* $\mathbb{E}[X|T = 1] = \mathbb{E}[Y]$*,* $\mathbb{E}[X|T = 2] = \mathbb{E}[Y + Y] = \mathbb{E}[Y] + \mathbb{E}[Y]$*. Linearity of distribution saves us here.*

Example 3 *We have two distributions on the same sample sample,* $p, q : \Omega \to [0, 1]$*. We need to distinguish between* p *and* q*. We have an oracle* O *that provides numbers distributed according to either* p *or* q*. That is, we are given access to* O_r *which provides numbers distributed according to distribution* r*. Return whether* $r = p$ *or* $r = q$*.*

We should probably look at the event $a \equiv \max_{A \subseteq \Omega} p(A) - q(A)$*, the element that exhibits maximum discrepancy. Then we should draw events from the set which maximised* a *and see what happens.*

Now, this number a *happens to be equal to* $\frac{1}{2} \sum_{w \in \Omega} |p(w) - q(w)|$

Chapter 3

Belief networks

There are two people, Alice and Bob. They are neighbours, and there's a wall between their houses. Alice's house has a sprinkler. One day, she wakes up and notices that her lawn is wet. So now, we have the random variables:

- A: Alice's lawn is wet/not wet.
- B: Bob's lawn is wet/not wet.
- S: Alice's sprinkler was switched on/not switched on.
- R: Rained or not.

We have a joint probability distribution $P(A\ B\ S\ R)$. We're now going to factor this using Bayes rule:

$$\mathbb{P}[A|BSR]\,\mathbb{P}[BSR] = \mathbb{P}[A|BSR]\,\mathbb{P}[B|SR]\,\mathbb{P}[R|S]\,\mathbb{P}[S]$$

For $\mathbb{P}[A|BSR]$ we need to provide $\mathbb{P}[A = 0|B = b\ S = s\ R = r]$. We can calculate

$$\mathbb{P}[A = 1|B = b\ S = s\ R = r] = 1 - \mathbb{P}[A = 0|B = b\ S = s\ R = r]$$

So we need to provide 8 values for all possible choices of (b, s, r).

Similarly, for $\mathbb{P}[B|SR]$ we need 4 numbers, and $\mathbb{P}[R|S]$ we need 2 numbers, and $\mathbb{P}[S]$ we need 1 number. The total is $8 + 4 + 2 + 1 = 15$, which is how many we need to describe the *full* joint distribution $P(A, B, S, R)$.

Now, for example, if we know the status of rain, we don't *need to know the status of Bob's lawn to comment on Alice's lawn*. So, we can rewrite $P(A|BSR) = P(A|SR)$. This now needs only 4 variables, from 8. Similarly, if we know the status of the rain, the sprinkler in Alice's yard cannot affect Bob's yard. So, we can rewrite $P(B|SR) = P(B|R)$. This dropped 4 numbers to 2 numbers. Finally, the status of rain does not affect whether the sprinkler was on or not. Hence, $P(S|R) = P(S)$. This lowers 2 numbers to 1 numbers. Hence, we now have a total of $4 + 2 + 1 + 1 = 8$, which is *half* of what we started with.

We can draw our relationships as a graph:

Let's assume we have an instantiation of this model on some concrete numbers:

$\mathbb{P}[R=1]=0.2$ $\mathbb{P}[S=1]=0.1$

$\mathbb{P}[B=1|R=1]=1$ $\mathbb{P}[B=1|R=0]=0.2$

$\mathbb{P}[A=1|R=0\ S=0]=0$ $\mathbb{P}[A=1|R=1\ S=1]=1$

$\mathbb{P}[A=1|R=1\ S=0]=1$ $\mathbb{P}[A=1|R=0\ S=1]=0.9$

We can now begin number-crunching and compute what $\mathbb{P}[S=1|A=1]$ is going to be:

$$\begin{aligned}\mathbb{P}[S=1|A=1] &= \frac{\mathbb{P}[S=1\ A=1]}{\mathbb{P}[A=1]} \\ &= \frac{\sum_{R\ B}\mathbb{P}[S=1\ A=1\ R\ B]}{\sum_{R\ B\ S}\mathbb{P}[A=1\ R\ B\ S]} \\ &= \frac{\sum_{R\ B}\mathbb{P}[A=1|R\ S=1]\,\mathbb{P}[B|R\ S=1]\,\mathbb{P}[R]\,\mathbb{P}[S]}{\sum_{R\ B\ S}\mathbb{P}[A=1|RSB]\,\mathbb{P}[B|RS]\,\mathbb{P}[R]\,\mathbb{P}[S]}\end{aligned}$$

Suppose we find that $\mathbb{P}[S=1|A=1]=0.3382$. Now, what is $\mathbb{P}[S=1|A=1\ B=1]$?

$$\begin{aligned}\mathbb{P}[S=1|A=1\ B=1] &= \frac{\mathbb{P}[S=1\ A=1\ B=1]}{\mathbb{P}[A=1\ B=1]} \\ &= \frac{\sum_{R}\cdots}{\sum_{S,R}\mathbb{P}[A=1\ B=1\ S\ R]}\end{aligned}$$

We find that $P(S=1|A=1\ B=1)=0.1604$. That is, if both alice and bob had wet lawns, then it's unlikely that the water was from the sprinkler.

Now, how do we design automated algorithms such that the above calculations we performed can be done *automatically* and *quickly*?

3.1 Computing joint distributions from the graph of the belief net

We first perform a topological sort of the DAG(directed acyclic graph), which is always guaranteed to exist.

We now visit the topo sort and multiply the values corresponding to each of the factors.

Notice that R has no incoming edges, so we need a term of $\mathbb{P}[R]$, Similarly, S has no incoming edges, so we need a term of $\mathbb{P}[S]$. B has an incoming edges from R, so we get a term $\mathbb{P}[B|R]$. Similarly, A has incoming edges from R, S so we get a term $\mathbb{P}[A|S\ R]$. This gives us the final solution:

$$\mathbb{P}[A\ B\ S\ R]=\mathbb{P}[R]\,\mathbb{P}[S]\,\mathbb{P}[B|R]\,\mathbb{P}[A|S\ R]$$

3.2 Conditional independence of random variables

We wish to study situations of the form: "Are random variables X and Y are independent conditioned on C"? If they are, this is notated as:

$$X \perp Y|C \iff \mathbb{P}[X\ Y|C]=\mathbb{P}[X|C]\,\mathbb{P}[Y|C]$$

Example 4 *Let us assume we have x_1, x_2, x_3, all of which are dependent on each other with a network of the form:*

There are 6 graphs that can represent the network ($3! = 6$)

Example 5

- *1.* $(x_1 \to x_3 \leftarrow x_2)$. *We get* $\mathbb{P}[X_1\ X_2\ X_3] = \mathbb{P}[X_1]\ \mathbb{P}[X_2]\ \mathbb{P}[X_3|X_1\ X_2]$
- *2.* $(x_1 \leftarrow x_3 \to x_2)$. *We get* $\mathbb{P}[X_1\ X_2\ X_3] = \mathbb{P}[X_3]\ \mathbb{P}[X_1|X_3]\ \mathbb{P}[X_2|X_3]$
- *3.* $x_1 \to x_3 \to x_2$. *We get* $\mathbb{P}[X_1\ X_2\ X_3] = \mathbb{P}[X_1]\,\mathbb{P}[X_3|X_1]\,\mathbb{P}[X_2|X_3]$.

Note that the first graph is not the same as the second graph, since in the first graph, (X_1, X_2) *are independent, but in the second one, they are not.*

The second and third are equal, and we can prove it!

$$\begin{aligned} 1 \leftarrow 3 \to 2 &= (\mathbb{P}[X_3]\ \mathbb{P}[X_1|X_3])\ \mathbb{P}[X_2|X_3] \\ &= (\mathbb{P}[X_1]\ \mathbb{P}[X_3|X_1])\ \mathbb{P}[X_2|X_3]\ \textit{(Bayes rule)} \\ &= 1 \to 3 \to 2 \end{aligned}$$

Similarly, the third and fourth are equal:

$$\begin{aligned} 2 \to 3 \to 1 &= (\mathbb{P}[X_2]\ \mathbb{P}[X_3|X_2])\ \mathbb{P}[X_1|X_3] \\ &= (\mathbb{P}[X_3]\ \mathbb{P}[X_3|X_2])\ \mathbb{P}[X_1|X_3]\ \textit{(Bayes rule)} \\ &= 2 \leftarrow 3 \to 1 \end{aligned}$$

subgraphs of the form $(x_1 \to x_3 \leftarrow x_2)$ are called as **collisions**. Here, x_1, x_2 are independent, but conditioned on x_3, they may become dependent!

$$\begin{aligned} &\mathbb{P}[X_1\ X_2|X_3] =_? \mathbb{P}[X_1|X_3]\,\mathbb{P}[X_2|X_3] \\ &\mathbb{P}[X_1|X_3]\,\mathbb{P}[X_2|X_1\ X_3] =_? \mathbb{P}[X_1|X_3]\,\mathbb{P}[X_2|X_3] \\ &\mathbb{P}[X_2|X_1\ X_3] =_? \mathbb{P}[X_2|X_3] \end{aligned}$$

Let X_1, X_2 be binary random variables that take on values $0, 1$ with probability half. Let $X_3 = X_1 \oplus X_2$ ($\oplus$ represents XOR). Now, X_3 is also a random variable. Now, notice that if we know only *one* random variable, the other two random variables are independent. However, as soon as we know *two random variables*, the third one is completely determined! So, $\mathbb{P}[X_1|X_1\ X_3] \in \{0, 1\}$, while $\mathbb{P}[X_2|X_3] = 0.5$ for this example. Hence, the above probabilities are not equal, and therefore X_1, X_2 are not independent when conditioned on X_3.

3.3 d connectivity

We say that X, Y are d-connected wrt Z, where Z is a set of random variables, if there exists an *undirected path* such that:

- For all colliders, either C or a descendant of $C \in Z$.
- No non-colliders should be in Z.

Example 6 $a \to b \to d \leftarrow c$. *Now let's consider whether* a, c *are d-connected with respect to* d. *There exists a path* $a \to b \to d \leftarrow c$. *In this path,* d *is a colliding vertex, which does belong to* Z. b *is a non-colliding vertex that does not belong to* Z. *Hence, this satisfies the conditions for this to be a* d*-connected path.*

Theorem 1 X *is not d-connected to* Y *with respect to* $Z \implies (X \perp Y|C)$ *Note that this is not iff.*

Proof 1

Chapter 4

Belief / Bayesian nets

Missed class

Chapter 5

Markov Networks

Not all distributions which have conditional independences. Not all of them can be modeled by a Bayes net.

If we have an undirected graph G where every person chooses the color of their hair based on their neighbour, and we have a 4-cycle $A \to B \to C \to D \to A$, then we can build a bayes net:

$$x_a \to x_c$$
$$x_a \to x_b$$
$$x_b \to x_d$$
$$x_c \to x_d$$

However, note that our bayes net is directed, while the underlying process was undirected. I can build the same bayes net by starting with x_c:

$$x_c \to x_a$$
$$x_c \to x_d$$
$$x_a \to x_b$$
$$x_d \to x_b$$

According to this, X_C `independent` $X_B | X_A, X_D$, but this is not true in the first bayes net!

To fix this, we just draw the *undirected graph*. For this, we need a tool called as a *potential*: potentials are non-negative functions that are associated with cliques in the graph, denoted by ϕ.

A clique of a graph is a complete sub-graph of a graph.

$$a, b \in \{r, b, g\}$$
$$\phi_{A,B}(a, b) \equiv \begin{cases} 10 & a \neq b \\ 1 & a = b \end{cases}$$

$$P_{X_A X_B X_C X_D} \equiv \prod_{c \in C} \phi_c(x_c) = \frac{1}{Z} \phi_{AB}(a,b)\phi_{AC}(a,c)\phi_{CD}(c,d)\phi_{BD}(b,d)$$

Here, Z is called as the partition function. It is the normalizing constant: $Z \equiv \sum_{a,b,c,d} \phi_{AB}(a,b)\phi_{AC}(a,b)\phi_{CD}($

Chapter 6

Preliminary definitions of Information

Definition 1 *Entropy(H): The entropy of a random variable X with probability distribution $p : X \to \mathbb{R}$ is defined as:*

$$H(X) \equiv -\sum_{x \in X} p(x) \log p(x) = \mathbb{E}\left[[] - \log \circ p\right]$$

Definition 2 *Condtional entropy($H(X|Y)$): The conditional entropy of a random variable X with respect to another variable Y is defined as:*

$$\begin{aligned} H(X|Y) &\equiv -\sum_{y \in Y} p(y) H(X|Y = y) \\ &= \sum_{y \in Y} p(y) \sum_{x \in X} -p(x|y) \log p(x|y) \\ &= \sum_{y \in Y} \sum_{x \in X} -p(y) p(x|y) \log p(x|y) \\ &= \sum_{y \in Y} \sum_{x \in X} -p(y \wedge x) \log p(x|y) \end{aligned}$$

Definition 3 *Kullback-Leibler divergence $D(X||Y)$: The Kullback-Leibler divergence of $X \sim p$ with respect to $X' \sim q$ is:*

$$D(X||X') \equiv \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$

Note that $D(X||X')$ is *not symmetric*.

Intuition: extra cost of encoding X if we thought the distribution were X'.

Useful extremal case to remember: Assume X' has $q(x) = 0$ for some letter $x \in X$. In this case, $D(X||X')$ would involve a term $\frac{p(x)}{0}$, which is ∞. This is intuitively sensible, since X' has no way to represent x, and hence X' is *infinitely far away from encoding* X. However, In this same case, one could have that X is able to encode all of X'.

Definition 4 *Mutual information: I(X;Y): This is the relative entropy between the joint and product distributions.*

$$I(X;Y) \equiv D(p(x,y)||p(x)p(y)) \equiv H(X) - H(X|Y)$$
$$= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Theorem 2 *Proof of equivalence of two definitions of mutual information:*

Proof 2

$$I(X;Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$
$$= \sum_x \sum_y p(x,y) \log \frac{p(x|y)}{p(x))}$$
$$= -\sum_x \sum_y p(x,y) \log p(x) - \big[-\sum_x \sum_y p(x,y) \log p(x|y)\big]$$
$$= -\sum_x p(x) \log p(x) - \big[-\sum_y \sum_x p(x,y) \log p(x|y)\big]$$
$$= -\sum_x p(x) \log p(x) - \big[-\sum_y p(y) \sum_x \log p(x|y)\big]$$
$$= H(X) - H(X|Y)$$

Some notes about mutual information:

- $I(X;Y) = I(Y;X)$. That is, I is symmetric.
- Since $I(X;Y) = H(X) - H(X|Y)$, one can view it as the *reduction in uncertainity* of X, after knowing Y. Alternatively, (to avoid double negatives), it's the *gain in certainty* of X after knowing Y. Another way of saying this is, what is the expected reduction in the number of yes/no questions to be answered to isolate the value of X on knowing the value of Y.
- $I(X;Y) = 0$ iff X, Y are independent. That is, knowing X reduces no uncertainity about Y.
- $I(X;X) = H(X)$. So, knowing X allows us to reduce our uncertainity of X by $H(X)$. ie, we completely know X, since we have *reduced our uncertainity of X* which was initially $H(X)$, by $H(X)$.

Theorem 3 *Chain rule for entropy: Let $X_1, X_2, \ldots X_n \sim p(x_1, x_2, \ldots, x_n)$ Then:*

$$H(X_1, X_2, \ldots X_n) = \sum_i H(X_i|X_{i-1}, X_{i-2}, \ldots X_1)$$

Proof 3

$H(X_1, X_2) = H(X_1) + H(X_2|X_1)$
$H(X_1, X_2, X_3) = H(X_1) + H(X_2, X_3|X_1) = H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1)$
induction for the rest

Definition 5 *Conditional mutual information: Conditional mutual information of random variables* X *and* Y *given* Z *is:*

$$I(X;Y|Z) \equiv H(X|Z) - H(X|Y,Z)$$

Theorem 4 *Chain rule for information:*

$$I(X_1, X_2, \ldots X_n; Z) = \sum_i I(X_i; Z|X_1, X_2, \ldots X_{i-1}$$

Proof 4 *TODO: finish*

Chapter 7

Variational auto encoders