Complete

Principles of Machine Learning

Exercises

Victor Verreet *victor.verreet@cs.kuleuven.be*
Laurens Devos *laurens.devos@cs.kuleuven.be*

Fall, 2021

Exercise Session 7: Learning Models

7.1 Missing at Random

The systolic blood pressure of 50 people is measured. Each person is measured twice, one month apart. The measurements are positively correlated. We simulate three possible causes for missing values.

1. There was a problem with the hard drive and no backup is available. Data were corrupted and some values are missing.
2. Everybody is tested twice, but the second value is recorded only if it is higher than some threshold τ.
3. Only people with a pressure higher than τ are invited to the second test.

Draw the graphical model for each case and explain whether the data is missing at random, completely at random or not at random. Use T_1 and T_2 as test variables and M_2 as missingness variable.

7.2 Counting Outcomes

Consider an experiment that can have three outcomes, O_1, O_2 and O_3, with respective unknown probabilities θ_1, θ_2 and $\theta_3 = 1 - \theta_1 - \theta_2$. We want to learn these probabilities from data. The experiment is run N times and outcome O_i is observed N_i times, with $N_3 = N - N_1 - N_2$. Show that maximizing the likelihood as a function of θ_i amounts to counting, so that $\theta_i = N_i/N$. Do this as follows.

1. Write down the likelihood $\mathcal{L}$ for the observed data in terms of θ_1 and θ_2.
2. Optimize the log-likelihood by differentiating it and deriving a set of equations for θ_1 and θ_2.
3. Solve this set of linear equations for θ_1 and θ_2.
4. Verify that $\theta_3 = 1 - \theta_1 - \theta_2$ is also obtained by counting.

7.3 Naive Bayes

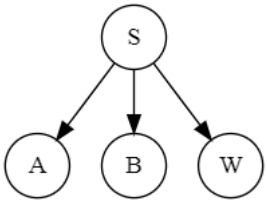


Figure 1: The naive Bayes assumption for butterfly species and attributes.

A biologist is studying three different species of butterflies, the monarch butterfly, speckled wood and holly blue. Within the species $S \in \{\text{monarch}, \text{wood}, \text{blue}\}$ there is still some variation in three attributes, namely antenna length $A \in \{\text{long}, \text{short}\}$, body colour $B \in \{\text{light}, \text{dark}\}$ and wing opacity $W \in \{\text{opaque}, \text{translucent}\}$. The biologist wants to get an estimate of the relative frequency for each species and of the values of the attributes within each species. Therefore he has collected this data for a couple of butterflies, seen in table 1. Using the naive Bayes assumption 1, help the biologist estimate $P[S]$, $P[A|S]$, $P[B|S]$ and $P[W|S]$.

Table 1: Butterfly species and attributes, observed for 9 instances I.

I	1	2	3	4	5	6	7	8	9
S	wood	wood	monarch	blue	wood	monarch	monarch	blue	monarch
A	long	short	long	short	short	short	long	long	short
B	dark	dark	dark	dark	light	dark	light	light	dark
W	opaque	translucent	opaque	opaque	translucent	opaque	opaque	opaque	opaque

7.4 Expectation Maximization

A Bayesian network imposes the factorization $P[a, b] = P[a]P[b|a]$ on the boolean variables a and b, but the probabilities $P[a = 1] = \theta_1$, $P[b = 1|a = 0] = \theta_2$ and $P[b = 1|a = 1] = \theta_3$ are unknown. They can be found by applying the expectation maximization algorithm to some data. The observed instances I are tabulated in 2. Expectation maximization works iteratively. The unknown probabilities $\theta^{(0)}$ are first initialised to some value. Then an expectation and maximization is step is performed in iteration n to obtain new values $\theta^{(n+1)}$ from $\theta^{(n)}$.

I	1	2	3	4
a	1	1	0	?
b	1	?	0	0

Table 2: Partial observations for (a, b) instances.

The first step in an iteration is to complete the missing data based on the current values for $\theta^{(n)}$. The weight of each completion is the probability of that completion conditioned on the observed values for that instance. For example, the fourth instance $(?, 0)$ has two completions, $(0, 0)$ and $(1, 0)$, with respective weights $P[a = 0|b = 0]$ and $P[a = 1|b = 0]$. Write down the completed table of instances together with their weights as a function of the $\theta^{(n)}$ values. This is called the expectation step.

The second step, the maximization, is to update the θ values to maximize the likelihood of the completed data. For completed data, likelihood maximization amounts to counting instances with their weights. Give the formulas that express $\theta^{(n+1)}$ as a function of $\theta^{(n)}$ by performing a weighted count on the completed data.

Lastly, carry out one iteration of the algorithm with initial values $\theta_1^{(0)} = 0.4$, $\theta_2^{(0)} = 0.3$ and $\theta_3^{(0)} = 0.2$.

7.5 Kullback-Leibler Divergence

Given two continuous probability distributions $p(x) > 0$ and $q(x) > 0$ for a variable $x \in \mathbb{R}$, then by definition

$$\mathrm{KL}(p, q) = \int_{\mathbb{R}} p(x) \log\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x$$

is the Kullback-Leibler divergence of these distributions. It is a sort of measure for how different the distributions are. Smaller values of $\mathrm{KL}(p, q)$ indicate that p and q are more similar. Using that $\log(s) \geq 1 - s^{-1}$ for all $s \in \mathbb{R}_0^+$, show that $\mathrm{KL}(p, q) \geq 0$ and that $\mathrm{KL}(p, p) = 0$.

Now consider the normal distributions

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x - \mu_p)^2}{2}\right) \quad \text{and} \quad q(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x - \mu_q)^2}{2}\right)$$

with respective means μ_p and μ_q. Calculate the Kullback-Leibler divergence $\mathrm{KL}(p, q)$ in terms of μ_p and μ_q. When is $\mathrm{KL}(p, q) = 0$? Does this agree with the Kullback-Leibler divergence giving some notion of dissimilarity of distributions?

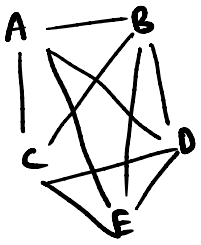
7.6 Most Likely Model

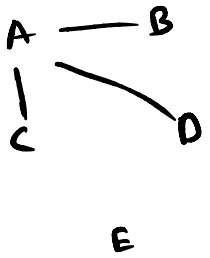
There are two models M_1 and M_2 that can generate boolean data points $x \in \{0, 1\}$. Under model M_1 the probability of generating $x = 1$ is $P[x = 1|M_1] = 0.6$ whereas for the second model $P[x = 1|M_2] = 0.8$. We do not know which model is the true model that generates the data, but five data points x have been observed from it. The outcome of these measurements is the sequence 1, 1, 0, 1, 0. Before seeing the data we had a prior belief that model M_1 is more likely by $P[M_1] = 0.7$. What is the probability that M_1 is the actual underlying model for the data? What is the probability for M_2? Which model is most likely the underlying model? Given these previous observations, what is the probability that the next observation will be $x = 1$, taking into account the uncertainty of the model?

7.7 Skeleton Learning

In a dataset we have observed the independences $\mathcal{I} = \{(A, B, C, D \perp E), (B, C \perp D|A), (B \perp C|A)\}$ for the variables A, B, C, D and E. Apply the PC algorithm to learn a skeleton that conforms to these independences.

$S_{B,D} = \{A\}$ $S_{C,D} = \{A\}$ $S_{B,C} = \{A\}$





7.1

1. $t_1 \quad m_{t_1}$ MCAR
   $\downarrow$
   $t_2 \quad m_{t_2}$

2. $t_1 \quad m_{t_1}$ MNAR
   $\downarrow$
   $t_2 \longrightarrow m_{t_2}$

3. $t_1 \longrightarrow m_{t_2}$
   $\downarrow$
   $t_2 \quad m_{t_1}$

7.2

$$L(\theta|D) = P(D|\theta) = \prod_{i=1}^{N} P(d_i|\theta)$$

$$= \theta_1^{N_1}\, \theta_2^{N_2} (1-\theta_1-\theta_2)^{N_3}$$

$$L_L = \log L = N_1 \log\theta_1 + N_2 \log\theta_2 + N_3 \log(1-\theta_1-\theta_2)$$

$$\frac{\partial L_L}{\partial \theta_1} = \frac{N_1}{\theta_1} - \frac{N_3}{1-\theta_1-\theta_2} = 0$$

$$\Rightarrow N_1(1-\theta_2) - N_1\theta_1 = N_3\theta_1$$

$$\Rightarrow \theta_1 = \frac{N_1(1-\theta_2)}{N_1+N_3}$$

$$\frac{\partial L_L}{\partial \theta_2} = \frac{N_2}{\theta_2} - \frac{N_3}{(1-\theta_1-\theta_2)} = 0$$

$$\Rightarrow N_2(1-\theta_1) = (N_2+N_3)\theta_2$$

$$\Rightarrow \theta_2 = \frac{N_2(1-\theta_1)}{N_2+N_3}$$

$$\theta_1 = \frac{N_1\left(1 - \frac{N_2(1-\theta_1)}{N_2+N_3}\right)}{N_1+N_2}$$

$$\Rightarrow \theta_1 = N_2\theta_1 - N_2 = 0$$

$$\Rightarrow (N_2+N_3)\theta_2 + N_2\theta_1 - N_2 = 0$$

$$N_1\theta_2 + (N_1+N_3)\theta_1 - N_1 = 0$$

$$\Rightarrow N_1(N_2+N_3)\theta_2 + N_1N_2\theta_1 - N_1N_2 = 0$$

$$N_1(N_2+N_3)\theta_2 + (N_2+N_3)(N_1+N_3)\theta_1 - N_1(N_2+N_3) = 0$$

$$\theta_1\left\{N_1N_2 - (N_2+N_3)(N_1+N_3)\right\} + N_1N_3 = 0$$

$$\Rightarrow \theta_1 = \frac{N_1N_3}{(N_2+N_3)(N_1+N_3) - N_1N_2} = \frac{N_1}{N}$$

$$(N_1+N_3)(N_2+N_3)\theta_2 + (N_1+N_3)\theta_1 N_2 - N_2(N_1+N_3)$$

$$N_1N_2\theta_2 + N_2(N_1+N_3)\theta_1 - N_1N_2 = 0$$

$$\Rightarrow \left\{(N_1+N_3)(N_2+N_3) - N_1N_2\right\}\theta_2 = N_2N_3$$

$$\theta_2 = \frac{N_2N_3}{\left\{(N_1+N_3)(N_2+N_3) - N_1N_2\right\}} = \frac{N_2}{N}$$

$$\theta_3 = 1 - \frac{N_1 N_3 + N_2 N_3}{\{(N_1+N_3)(N_2+N_3) - N_1 N_2\}}$$

$$= \frac{N_3^2}{\{(N_1+N_3)(N_2+N_3) - N_1 N_2\}}$$

$$= \frac{N_3}{N}$$

7.3

$$P(S, A, B, W) = P(S)\,P(A|S)\,P(B|S)\,P(W|S)$$

$$P(S=m) = \frac{4}{9}$$

$$P(S=w) = \frac{1}{3}$$

$$P(S=b) = \frac{2}{9}$$

$P(A=1\|S)$	m	w	b
	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{2}$

S

$P(B=1\|S)$	m	w	b
	$\frac{3}{4}$	$\frac{2}{3}$	$\frac{1}{2}$

S

$P(W=a\|S)$	m	w	b
	1	1	$\frac{1}{3}$

7.4

$P(a=1) = \theta_1$

$P(b=1|a=0) = \theta_2$

$P(b=1|a=1) = \theta_3$

a	b		weights
1	1		1
1	0	$q^2(b=0)$	
1	1	$q^2(b=1)$	
0	0		1
1	0	$q^3(a=1)$	
0	0	$q^4(a=0)$	

$q^1(b=0) = P(b=0|a=1)$

$= 1 - \theta_3$

$q^2(b=1) = \theta_3$

$q^3(a=1) = P(a=1|b=0)$

$= \frac{P(a=1, b=0)}{P(b=0)}$

$= \frac{P(a=1)\cdot P(b=0|a=1)}{\sum P(a, b=0)}$

$= \frac{\theta_1(1-\theta_3)}{\theta_1(1-\theta_3) + (1-\theta_1)(1-\theta_2)}$

$q^4(a=0) = 1 - \frac{\theta_1(1-\theta_3)}{\theta_1(1-\theta_3) + (1-\theta_1)(1-\theta_2)}$

$= \frac{(1-\theta_1)(1-\theta_2)}{\theta_1(1-\theta_3) + (1-\theta_1)(1-\theta_2)}$

$q^1(b=0) = 0.8$

$q^2(b=1) = 0.2$

$$q^3(a=1) = \frac{0.4 \times 0.8}{0.4 \times 0.8 + 0.6 \times 0.7}$$

$$= \frac{32}{32+42} = \frac{32}{74}$$

$$= \frac{16}{37}$$

$$q^4(a=1) = \frac{21}{37}$$

$$\theta_1^{(1)} = P(a=1)$$

$$= \frac{q^1(b=0) + q^2(b=1) + 1 + q^3(a=1)}{4}$$

$$= \frac{0.8 + 0.2 + 1 + 0.43}{4} = 0.608$$

$\theta_2^1 = 0$

$$\theta_3^1 = \frac{1+0.2}{2.43} = 0.49$$

$$KL(P,Q) = \int_R P(x) \log\left(\frac{Q(x)}{P(x)}\right) dx$$

$A(x) \geq B(x)$

$\int A(x)\,dx \geq \int B(x)$

$$\geq \int_R P(x)\left(1 - \frac{Q(x)}{P(x)}\right) dx$$

$$\geq \int_R P(x)\left(\frac{P(x) - Q(x)}{P(x)}\right) dx$$

$$\geq \int \left(P(x) - Q(x)\right) dx = \int_R P(x)\,dx - \int_R Q(x)\,dx$$

$$= 0$$

$$KL(P,Q) = \int_R P(x) \log\left(\frac{P(x)}{Q(x)}\right) dx$$

$$= \int_R P(x) \left\{\frac{(x-\mu_Q)^2}{2} - \frac{(x-\mu_P)^2}{2}\right\} dx$$

$$= \int_R P(x) \left\{\frac{(\mu_P - \mu_Q)(2x - \mu_P - \mu_Q)}{2}\right\} dx = \mu_P(\mu_P - \mu_Q) - \frac{(\mu_P^2 - \mu_Q^2)}{2}$$

$$= \frac{(\mu_P - \mu_Q)^2}{2}$$

7.6

$$P(M_1|D) = \frac{P(D|M_1)\,P(M_1)}{P(D)}$$

$$= \frac{(0.6)^3 \times (0.4)^2 \times 0.7}{(0.6)^3 \times (0.4)^2 \times 0.7 + (0.8)^3 \times (0.2)^2 \times 0.3}$$

$$= 0.8$$

$P(M_2|D) = 0.2$

$$P(x=1) = P(x=1|M_1)\,P(M_1) + P(x=1|M_2)\,P(M_2)$$

$$= 0.6 \times 0.8 + 0.2 \times 0.8$$

$$= 0.64$$