Introduction to Artificial Intelligence (INFO8006)

Exercise session 4

Maximum likelihood estimation

Given a set of i.i.d. observations $\mathcal{D} = \{x_1, ..., x_N\}$, a set of unknown parameters $\theta = [\theta_1, ..., \theta_K]$ and the likelihood function $P(x_i \mid \theta)$ of one observation given the parameters, we derive the likelihood of the parameters for this set of observations as

$$P(\mathcal{D} \mid \theta) = \prod_{i=1}^{N} P(x_i \mid \theta).$$

From which we can recover the maximum likelihood estimate θ^* of the parameters as

$$\theta^* = \underset{\theta}{argmax} P(\mathcal{D} \mid \theta).$$

This can typically be found by cancelling the derivative of the associated log-likelihood w.r.t. each parameter

$$\frac{\partial LL(\mathcal{D}; \theta)}{\partial \theta_k} = 0, \quad \forall k.$$

Bayesian learning and maximum a posteriori

We can treat parameters as random variables to incorporate uncertainty about their values. To do so, we have to specify a prior distribution $P(\theta)$ over the parameters. When new observations $\mathcal{D}$ are collected, the distribution over parameters can be updated, leading to the posterior

$$P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta) P(\theta).$$

When the latter is not analytically tractable, we can still compute the maximum a posteriori

$$\theta^* = \underset{\theta}{argmax} P(\theta \mid \mathcal{D}).$$

Cheat sheet for Gaussian models

From the joint

$$p\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} = \mathcal{N}\left(\begin{pmatrix}\mathbf{a}\\ \mathbf{b}\end{pmatrix}, \begin{pmatrix}\mathbf{A} & \mathbf{C}\\ \mathbf{C}^{\mathbf{T}} & \mathbf{B}\end{pmatrix}\right),$$

the marginal and conditional distributions are given by

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}(\mathbf{a}, \mathbf{A})\\ p(\mathbf{y}) &= \mathcal{N}(\mathbf{b}, \mathbf{B})\\ p(\mathbf{x} \mid \mathbf{y}) &= \mathcal{N}(\mathbf{a} + \mathbf{C}\mathbf{B}^{-\mathbf{1}}(\mathbf{y} - \mathbf{b}), \mathbf{A} - \mathbf{C}\mathbf{B}^{-\mathbf{1}}\mathbf{C}^{\mathbf{T}})\\ p(\mathbf{y} \mid \mathbf{x}) &= \mathcal{N}(\mathbf{b} + \mathbf{C}^{\mathbf{T}}\mathbf{A}^{-\mathbf{1}}(\mathbf{x} - \mathbf{a}), \mathbf{B} - \mathbf{C}^{\mathbf{T}}\mathbf{A}^{-\mathbf{1}}\mathbf{C}). \end{aligned}$$

On the other hand, from

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}(\mathbf{m}, \mathbf{P})\\ p(\mathbf{y} \mid \mathbf{x}) &= \mathcal{N}(\mathbf{H}\mathbf{x} + \mathbf{u}, \mathbf{R}), \end{aligned}$$

we can recover the joint distribution as

$$p\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} = \mathcal{N}\left(\begin{pmatrix}\mathbf{m}\\ \mathbf{H}\mathbf{m} + \mathbf{u}\end{pmatrix}, \begin{pmatrix}\mathbf{P} & \mathbf{P}\mathbf{H}^{\mathbf{T}}\\ \mathbf{H}\mathbf{P} & \mathbf{H}\mathbf{P}\mathbf{H}^{\mathbf{T}} + \mathbf{R}\end{pmatrix}\right),$$

In session exercises: Ex. 1, Ex. 2

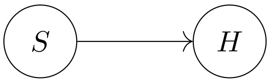
Exercise 1 Nuclear power plant (AIMA, Ex 14.11)

In your local nuclear power station, there is an alarm that senses when a temperature gauge exceeds a given threshold. The gauge measures the temperature of the core. Consider the Boolean variables A (alarm sounds), F_A (alarm is faulty), and F_G (gauge is faulty) and the multi-valued nodes G (gauge reading) and T (actual core temperature).

1. Draw a Bayesian network for this domain, given that the gauge is more likely to fail when the core temperature gets too high.
2. From your network topology, can you conclude $T \perp F_A \mid F_G$ and $T \perp F_A \mid F_G, A$?
3. Suppose there are just two possible actual and measured temperatures: low (l) and high (h). The probability that the gauge gives the correct temperature is x when it is working, but y when it is faulty. Give the conditional probability table associated with G.
4. Suppose the alarm is always triggered by high measured temperatures, unless it is faulty, in which case it never sounds. Give the conditional probability table associated with A.
5. Suppose the gauge is not faulty and the alarm is triggered. Calculate an expression for the probability that the temperature of the core is too high, in terms of the various conditional probabilities in the network.

Exercise 2 Student–TA relationship

The teaching assistant of the course is worried about the time he spends to prepare the exercise sessions. He wants to investigate the influence of students can have on his schedule. To do so, he builds the following model



where S is a r.v. denoting the number of students leaving the room between a theoretical lecture and the practical session that follows, and H is a r.v. representing the number of hours spent to prepare the next session. He chooses the following relations for each variable

$$P(S = s) = \texttt{Poisson}(S = s; \lambda) = \frac{\lambda^s e^{-\lambda}}{s!} \tag{1}$$

$$H = \omega S + H_0 + \mathcal{N}(0, \sigma^2) \tag{2}$$

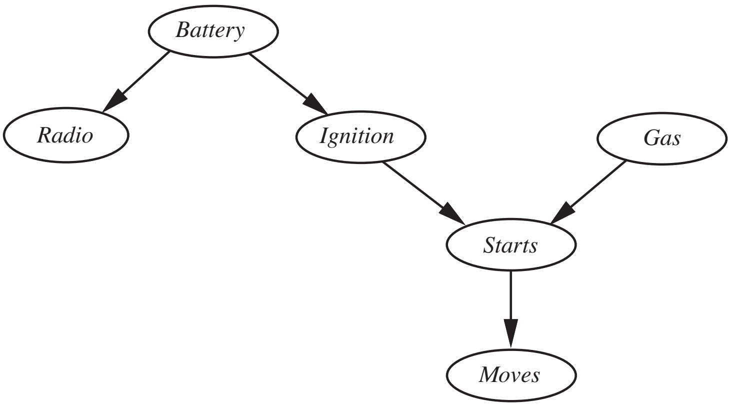
where H_0 corresponds to the number of hours the TA should spend per session regarding its contract and ω is the influence weight of students over the TA.

1. Identify the parameters of the Bayesian network and give the joint log-likelihood of one pair (s_i, h_i) given the model.
2. The TA collects a set $\mathcal{D} = \{(s_i, h_i)\}_{i=1}^N$ of independent observations. Derive the maximum likelihood estimate of the parameters.
3. The TA has just started to collect observations of students leaving the classroom. Being an adept of the Bayes' school, he knows that using MLE can lead to overfitting. Hence, he decides to incorporate uncertainty in his analysis of the parameter λ as he has no strong knowledge about it. Using a prior $\lambda \sim \texttt{Gamma}(\alpha, \beta)$, what would be the posterior distribution $P(\lambda \mid \mathcal{S} = \{s_i\}_{i=1}^N)$?

   Hint: $\texttt{Gamma}(x \mid \alpha, \beta) = \frac{1}{Z} x^{\alpha-1} e^{-\beta x}$, with Z a normalizing constant.
4. Knowing that the mean, the variance and the mode of $\texttt{Gamma}(\alpha, \beta)$ are respectively $\frac{\alpha}{\beta}$, $\frac{\alpha}{\beta^2}$ and $\frac{\alpha - 1}{\beta}$, interpret how they evolve w.r.t. N between the prior and the posterior.

Exercise 3   Car diagnosis (AIMA, Ex 14.8)

Let be the following Bayesian network describing some features of a car's electrical system and engine. Each variable is Boolean, and the true value indicates that the corresponding aspect of the vehicle is in working order.



1. Extend the network with the Boolean variables IcyWeather and StarterMotor.
2. According to your knowledge of cars, give reasonable conditional probability tables for all the nodes.
3. How many independent values are contained in the joint probability distribution for eight Boolean nodes, assuming that no conditional independence relations are known to hold among them?
4. How many independent probability values do your network tables contain?

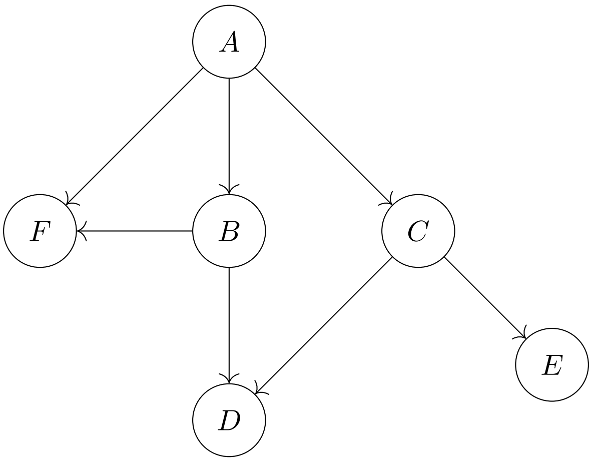
Exercise 4 Gaussian parameter learning with known σ^2

You have in your possession a high-tech laser that emits light of intensity μ_l with great precision. You want to use it to test a new light-intensity measurement device. You assume that your measures y_i follow a Gaussian distribution $\mathcal{N}(\mu_y, \sigma^2)$ with $\mu_y \sim \mathcal{N}(\mu_l, \sigma_l^2)$ corresponding to the model of your laser given in the manual. You assume that σ^2 is fixed and that you can determine it.

You have access to $\mathcal{Y}_c$ a calibration set of C observations and you want to make N new measures with your device. You want to study how you can reduce your uncertainty about the measures using $\mathcal{Y}_c$.

1. Compute the prior and posterior predictive distributions $P(y)$ and $P(y \mid \mathcal{Y}_c)$ if you have access to one observation, i.e. $\mathcal{Y}_c = \{y_o\}$.
2. Looking at you predictive uncertainty, what is the benefit of incorporating $\mathcal{Y}_c$?

Exercise 5 Independence



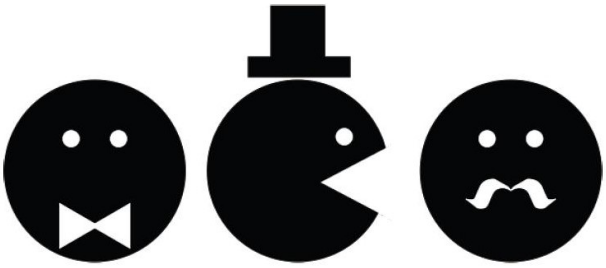
Considering the hereabove Bayesian network, which of the following statements are enforced by the network structure?

1. $P(B, C) = P(B)P(C)$
2. $P(B, C|A) = P(B|A)P(C|A)$
3. $P(F, E) = P(F)P(E)$
4. $P(F|A, D, E) = P(F|A, D)$
5. $P(B, E) = \sum_{a,c,d,f} P(a)P(B|a)P(c|a)P(d|B, c)P(E|c)P(f|a, B)$

For the same network, use inference by variable elimination to compute $P(E|A = 1, B = 1)$.

Exercise 6 Pacbaby (UC Berkeley CS188, Spring 2014)

Pacman and Pacwoman have been searching for each other in the maze. Pacwoman has been pregnant with a baby, and just this morning she has given birth to Pacbaby[1]. Because Pacbaby was born before Pacman and Pacwoman were reunited in the maze, he has never met his father. Naturally, Pacwoman wants to teach Pacbaby to recognize his father, using a set of pictures of Pacman. She also has several pictures of ghosts to use as negative examples.



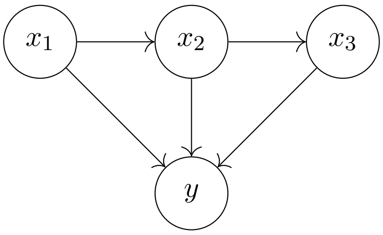
Because the pictures are black and white, and were taken from various angles, Pacwoman has decided to teach Pacbaby to identify Pacman based on salient features: the presence of a bowtie B, hat H or mustache M. The following table summarizes the content of the pictures. Each feature takes realization in $\{0, 1\}$, where 0 and 1 mean the feature is respectively absent and present. The subject of the picture is described by a random variable $S \in \{0, 1\}$, where 0 is a ghost and 1 is Pacman.

B	H	M	S
0	0	0	1
1	0	0	0
1	0	1	1
1	1	0	0
0	1	0	0
1	1	1	1

1. Suppose Pacbaby has a Naive Bayes based brain. Draw the Bayesian network that would represent the dependencies between S, B, H and M for Pacbaby.
2. Write the Bayesian classification rule for this problem, *i.e.* the formula that given a data point (b, h, m) returns the most likely subject. Write the formula in terms of conditional and prior probabilities. What does the formula become under the assumptions of Pacbaby ?
3. What are the parameters of this model? Give estimates of these parameters according to the pictures provided by Pacwoman.
4. Pacman eventually shows up wearing a bowtie, but no hat or mustache. Will Pacbaby recognize his father?

[1] Congratulations!

Exercise 7 Predict your grade



The hereabove Bayesian network represents how the final grade of a class is computed. In this model, x_1, x_2 and x_3 respectively denote the grades obtained by a student at the homework, project and exam. The teaching assistant that grades the homework also grades the project and the exam, which introduces a slight bias in the corrections. In particular, $x_2 \sim \mathcal{N}(a_1 x_1 + \mu_2, \sigma_2^2)$ and $x_3 \sim \mathcal{N}(a_2 x_2 + \mu_3, \sigma_3^2)$. Finally, $y \sim \mathcal{N}(a_3 x_1 + a_4 x_2 + a_5 x_3 + \mu_y, \sigma_y^2)$ stands for the final grade, which is a linear combination of the grades obtained by the student during the semester plus some Gaussian noise due to rounding errors. Answer the following questions about this model.

1. Assuming the parameters of the model are known, what is the expected value of y given x_1 and x_2.
2. Suppose now that the model's parameters are unknown. Given a learning set $d = \{(x_{i,1}, x_{i,2}, y_i)\}$ of N independent and identically distributed points, determine the model that best describes d.

Exercise 8 Heteroscedastic linear regression

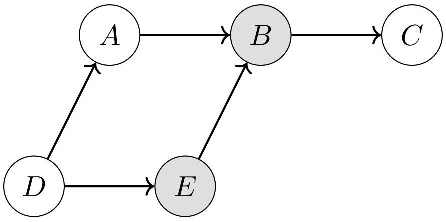
What becomes the expression of the weight vector w in the solution of question 2.2 if the noise is different for each sample? In particular, $y_i \sim N(w^T x, \sigma_i^2)$ and we know the values σ_i.

Quiz

The Normal distribution $\mathcal{N}(\mu, \sigma)$ is described by the density function

- □ $p(x) = \frac{1}{\sigma} \exp\left(-(z + \exp(-z))\right)$, with $z = \frac{x-\mu}{\sigma}$.
- □ $p(x) = \frac{1}{2\sigma} \exp\left(-\frac{|x-\mu|}{\sigma}\right)$.
- □ $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$.
- □ $p(x) = 1 \,/\, \pi\sigma \left[1 + \left(\frac{x-\mu}{\sigma}\right)^2\right]$.

Consider the Bayesian network shown below. We want to infer $\mathbf{P}(A|b, e)$ where A is the query variable, B and E are evidence variables, and C and D are hidden variables. Which of the following statements is true?



- □ $\mathbf{P}(A|b, e) \propto \sum_c \sum_d P(c|b)\mathbf{P}(b|A)P(b|e)\mathbf{P}(A|d)P(e|d)P(d)$
- □ $\mathbf{P}(A|b, e) \propto \sum_c \sum_d P(c|b)\mathbf{P}(b|A, e)\mathbf{P}(A|d)P(e|d)$
- □ $\mathbf{P}(A|b, e) \propto \sum_c \sum_d P(c|b)\mathbf{P}(b|A, e)\mathbf{P}(A|d)P(e|d)P(d)$
- □ $\mathbf{P}(A|b, e) \propto \sum_b \sum_e P(c|b)\mathbf{P}(A|b, e)\mathbf{P}(A|d)P(e|d)P(d)$

In a Bayesian network

- □ Independence between two variables is guaranteed if all path connecting them are inactive.
- □ We can guarantee dependence between two variables using d-separation.
- □ A path between two nodes is active if it contains at least one active triplet.
- □ A cascade triplet is active if the center node is observed.

In a one node Bayesian network with a binary variable following a Bernoulli distribution. If we observe T positive realizations and F negative ones,

- □ the maximum likelihood estimate of the positive probability is $\dfrac{T}{F}$.
- □ the maximum likelihood estimate of the positive probability is $\dfrac{T+F}{T}$.
- □ the maximum likelihood estimate of the positive probability is $\dfrac{T}{T+F}$.
- □ None of the above.