INFO8006 Introduction to Artificial Intelligence

Exercises 7: Learning

Learning outcomes

At the end of this exercise session, you should be able to:

- Apply Maximum Likelihood Estimation (MLE) to estimate the parameters of a model from data.

Exercise 1: PacBaby [1]

Pacman and Mrs. Pacman have been searching for each other in the Maze. Mrs. Pacman has been pregnant with a baby, and just this morning she has given birth to Pacbaby (Congratulations, Pacmans!). Because Pacbaby was born before Pacman and Mrs. Pacman reunited in the maze, he has never met his father. Naturally, Mrs. Pacman wants to teach Pacbaby to recognize his father, using a set of Polaroids of Pacman. She also has several pictures of ghosts to use as negative examples. Because the polaroids are black and white, and were taken from strange angles, Mrs. Pacman has decided to teach Pacbaby to identify Pacman based on more salient features: the presence of a bowtie (b), hat (h), or mustache (m).
The following table summarizes the content of the Polaroids. Each binary feature is represented as 1 (meaning the feature is present) or 0 (meaning it is absent). The subject y of the photo is encoded as $+1$ for Pacman or -1 for ghost.

(m)	(b)	(h)	Subject (y)
0	0	0	$+1$
1	0	0	$+1$
1	1	0	$+1$
0	1	1	$+1$
1	0	1	-1
1	1	1	-1

Table 1: Matrix of Pacman polaroids data.

Suppose Pacbaby has a Naive Bayes based brain.

1. Write the Naive Bayes classification rule for this problem (i.e. write a formula which given a data point $x = (m, b, h)$ returns the most likely subject y). Write the formula in terms of conditional and prior probabilities. Be explicit about which parameters are involved, but you do not need to estimate them yet.

   The naive Bayes model assumes that the inputs are independent conditionally to Y, which can be summarised by the Bayesian network of Figure 1. The classification rule for the triplet of observations (m, b, h) is given by:

   $$y^\star = \arg\max_y p(y|m, b, h) = \arg\max_y p(y, m, b, h) = \arg\max_y p(y)p(m|y)p(b|y)p(h|y)$$

2. What are the parameters of this model? Give estimates of these parameters for the classification rule based on the Polaroids.

   The parameters of the model are the conditional probability tables (CPT). By marginalizartion and conditioning we directly obtain:

	$y=+1$	$y=-1$
P(y)	$\frac{4}{6}$	$\frac{2}{6}$

	$y=+1$	$y=-1$
P(m=1|y)	$\frac{2}{4}$	$\frac{2}{2}$
P(b=1|y)	$\frac{2}{4}$	$\frac{1}{2}$
P(b=1|y)	$\frac{1}{4}$	$\frac{2}{2}$

3. Suppose a character comes by wearing a hat but without a mustache or bowtie. What would happen if Pacbaby had to guess the identity of the character?

   To answer this question we have to apply the decision rule given in question 1. If $y = 1$, the probability of seeing this data point $x = (0, 0, 1)$ is

   $$P(y = 1)P(m = 0|y = 1)P(b = 0|y = 1)P(h = 1|y = 1) = \frac{4 \times 2 \times 2 \times 1}{6 \times 4 \times 4 \times 4} = \frac{1}{24}$$

   If y=-1, the probability $P(m = 0|y = -1) = 0$, so the probability of seeing this is 0. Pacbaby would classify this as Pacman.

[1] Source: `http://ai.berkeley.edu/sections/section_10_sp14.pdf`

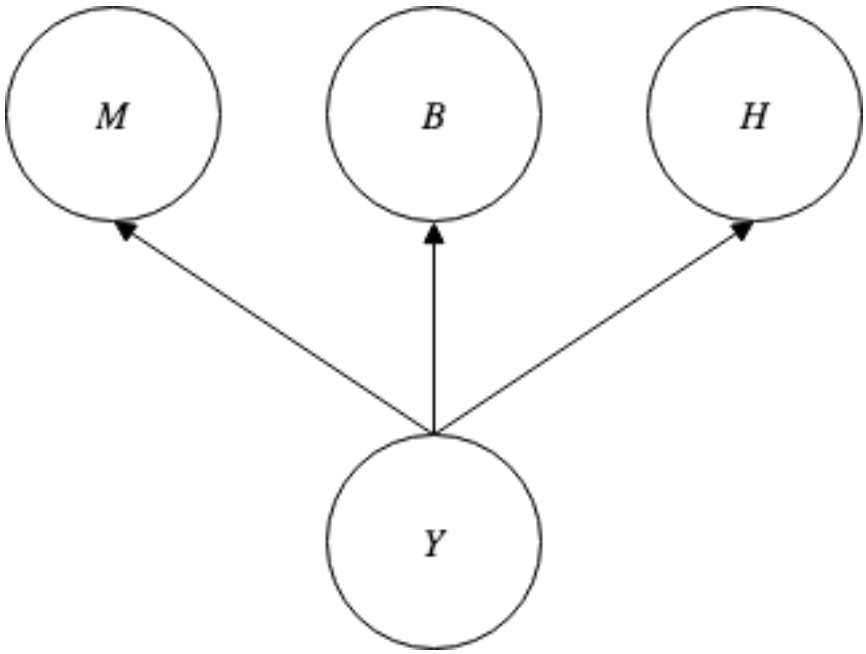

Figure 1: Naive Bayes network

Exercise 2: Predict your grade

The Bayesian network shown in Figure 2 represents how the final mark of this class is computed. In this model, X_1, X_2, X_3 respectively denote the marks obtain by a student at the homework, the project and at the exam. The teaching assistant who marks the exam also marks the homework, which implies a slight bias in the exam correction, such that we can write $X_3 = a_1 X_2 + \mathcal{N}(\mu_3, \sigma_3^2)$. The random variable $C = a_2 X_1 + a_3 X_2 + \mathcal{N}(\mu_C, \sigma_C^2)$ is a linear combination of the homework and project grades plus a computation error which is normally distributed. Finally, $Y = a_4 C + a_5 X_3 + \mathcal{N}(\mu_y, \sigma_y^2)$ stands for the final grade obtained by the student, which is a linear combination of the exam grades and the grades obtained for the work done during the semester plus some Gaussian noise on it due to rounding errors. Answer the following questions about this model.

1. Assume the parameters of the model are known, give the prediction rule of Y given X_1, X_2, X_3.

   Let's first rewrite the random variable Y in term of the input random variables X_1, X_2, X_3:

   $$Y = a_4 C + a_5 X_3 + \mathcal{N}(\mu_y, \sigma_y^2) \tag{1}$$
   $$= a_4(a_2 X_1 + a_3 X_2 + \mathcal{N}(\mu_C, \sigma_C^2)) + a_5 X_3 + \mathcal{N}(\mu_y, \sigma_y^2) \tag{2}$$

   Due to properties of Normal distribution this implies that $P(Y|x_1, x_2, x_3) = \mathcal{N}(a_2 a_4 x_1 + a_3 a_4 x_2 + a_5 x_3 + a_4 \mu_C + \mu_y, a_4^2 \sigma_C^2 + a_5^2 \sigma_3^2 + \sigma_y^2)$. The decision rule for the input triplet (x_1, x_2, x_3) is given by:

   $$y^\star = \arg\max_y P(y|x_1, x_2, x_3) \tag{3}$$
   $$= \arg\max_y \mathcal{N}(a_2 a_4 x_1 + a_3 a_4 x_2 + a_5 x_3 + a_4 \mu_C + \mu_y, a_4^2 \sigma_C^2 + \sigma_y^2) \tag{4}$$
   $$= a_2 a_4 x_1 + a_3 a_4 x_2 + a_5 x_3 + a_4 \mu_C + \mu_y \tag{5}$$

2. Simplify this rule and give the minimal set of parameters of this model that are required to predict Y given X_1, X_2, X_3.

   The distribution $p(Y|x_1, x_2, x_3)$ can be simplified if we are only interested by using it to predict the value of y, it becomes $p(y|x_1, x_2, x_3) = \mathcal{N}(w_1 x_1 + w_2 x_2 + w_3 x_3 + b, \sigma)$.
   The decision rule for the input triplet (x_1, x_2, x_3) is given by:

   $$y^\star = \arg\max_y p(y|x_1, x_2, x_3) \tag{6}$$
   $$= \arg\max_y \mathcal{N}(w_1 x_1 + w_2 x_2 + w_3 x_3 + b, \sigma) \tag{7}$$
   $$= w_1 x_1 + w_2 x_2 + w_3 x_3 + b \tag{8}$$

3. Can you recognise a model seen in class?

   Yes it is a linear regression.

4. From a data set of points $d = \{(x_{1,1}, x_{2,1}, x_{3,1}, y_1), \ldots, (x_{1,N}, x_{2,N}, x_{3,N}, y_N)\}$ explain how you would compute the parameters of the simplest model.

We have to solve the following optimisation problem:

$$W = \arg\max_W p(d|W) \tag{9}$$
$$= \arg\max_W \log(p(d|W)) \tag{10}$$
$$= \arg\max_W \sum_{i=1}^{N} \log(p(y_i, X_i|W)) \tag{11}$$
$$= \arg\max_W \sum_{i=1}^{N} \log(p(y_i|X_i, W)) + \log(p(X_i)) \tag{12}$$
$$= \arg\max_W \sum_{i=1}^{N} \log(p(y_i|X_i, W) \tag{13}$$
$$= \arg\max_W \sum_{i=1}^{N} \frac{-1}{2} \log(2\pi\sigma^2) - \frac{(W^T X_i - y_i)^2}{2\sigma^2} \tag{14}$$
$$= \arg\min_W \sum_{i=1}^{N} (W^T X_i - y_i)^2 \tag{15}$$

Where $W = [w_1, w_2, w_3, b]^T$ and $X = [x_1, x_2, x_3, 1]$. The last expression is convex, consequently it can be minimised by finding the value of W that cancels the gradient:

$$\nabla_W \sum_{i=1}^{N} (W^T X_i - y_i)^2 = 0 \tag{16}$$
$$\sum_{i=1}^{N} \nabla_W (W^T X_i - y_i)^2 = 0 \tag{17}$$
$$\sum_{i=1}^{N} 2X_i (W^T X_i - y_i) = 0 \tag{18}$$
$$X^T (XW - y) = 0 \tag{19}$$
$$X^T XW - X^T y = 0 \tag{20}$$
$$W - (X^T X)^{-1} X^T y = 0 \tag{21}$$
$$W = (X^T X)^{-1} X^T y \tag{22}$$

Where $X = [X_1, \ldots, X_N]^T \in \mathbb{R}^{N \times 4}$ and $y = [y_1, \ldots, y_N]^T \in \mathbb{R}^N$.

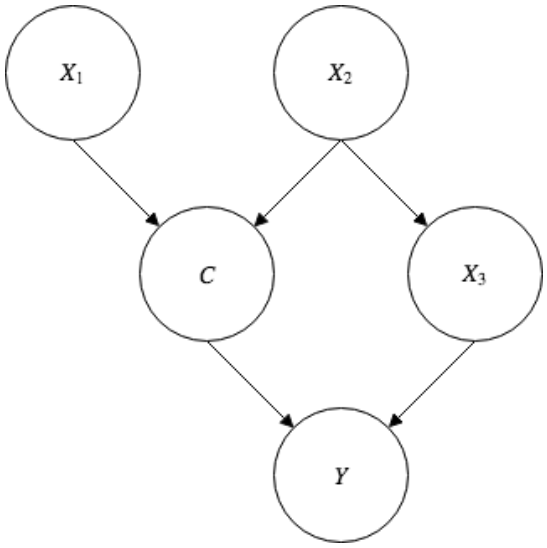

Figure 2: Bayesian network of the final grade calculation.

⋆ Exercise 3: Learning to play Pacman (August 2020)

You observe a Grandmaster agent playing Pacman. How can you use the moves you observe to train your own agent?

(a) Describe formally the data you would collect, the inference problem you would consider, and how you would solve it.

(b) How would you design a neural network to control your agent? Define mathematically the neural network architecture, its inputs, its outputs, its parameters, as well as the loss you would use to train it.

(c) Discuss the expected performance of the resulting agent when (i) the Grandmaster agent is optimal, and (ii) the Grandmaster agent is suboptimal.

⋆ Exercise 4: Heteroscedastic linear regression

What becomes the expression of the weight matrix W in the solution of 2.4 if the noise is different for each sample? In particular, $y_i \sim \mathcal{N}(y|W^T X, \sigma_i^2)$. Suppose you know the values σ_i. Bonus: Can you generalise this expression when the noise is correlated?

⋆ Exercise 5: Ridge regression

Show that linear regression with $L1$ penalisation corresponds to assuming that noise follows a Laplace distribution in contrast to the correspondence between $L2$ penalisation (least squares) and Normal noise distribution.