1 Learning (13/12/2018)

1.1 Objectives

At the end of this repetition you should be able to:

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

1.2 Exercises

a PacBaby ($\approx 20min$ [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:

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

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?

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

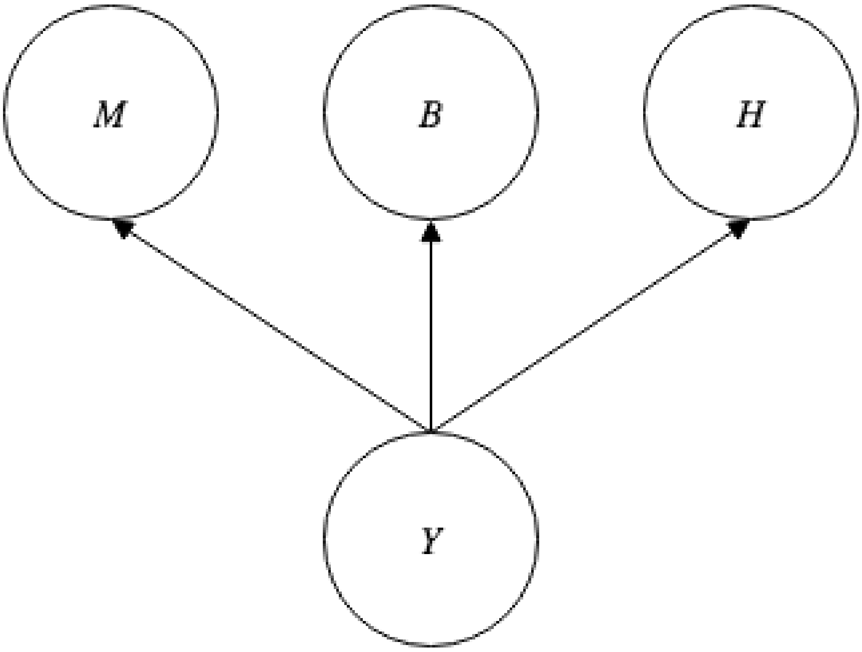

Figure 1: Naive Bayes network

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.

b Predict your grade ($\approx 30min$)

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:

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

   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:

$$\begin{aligned} y^\star &= \arg\max_y P(y|x_1, x_2, x_3) && (3) \\ &= \arg\max_y \mathcal{N}(a_2 a_4 x_1 + (a_3 a_4 + a_1 a_5) x_2 + a_4 \mu_C + a_5 \mu_3 + \mu_y, a_4^2 \sigma_C^2 + a_5^2 \sigma_3^2 + \sigma_y^2) && (4) \\ &= a_2 a_4 x_1 + a_3 a_4 x_2 + a_5 x_3 + a_4 \mu_C + \mu_y && (5) \end{aligned}$$

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 $p(Y|x_1, x_2, x_3) = \mathcal{N}(w_1x_1 + w_2x_2 + w_3x_3 + b, \sigma)$.
The decision rule for the input triplet (x_1, x_2, x_3) is given by:

$$\begin{aligned}
y^\star &= \arg\max_y P(y|x_1, x_2, x_3) && (6)\\
&= \arg\max_y \mathcal{N}(w_1x_1 + w_2x_2 + w_3x_3 + b, \sigma) && (7)\\
&= w_1x_1 + w_2x_2 + w_3x_3 + b && (8)
\end{aligned}$$

3. Can you recognise a model seen in class?
   Yes it is just a linear regression.

4. From a data set of points $d = \{(x_{1,1}, x_{2,1}, x_{3,1}, y_1), \ldots, (x_{1,N}, x_{2,N}, x_{3,N}, y_N)\}$ (that you assume to be i.i.d) explain how you would compute the parameters of the simplest model.
   We have to solve the following optimisation problem:

$$\begin{aligned}
W &= \arg\max_W p(d|W) && (9)\\
&= \arg\max_W p(y_{1:N}|X_{1:N}, W)p(X_{1:N}) && (10)\\
&= \arg\max_W \log(p(y_{1:N}|X_{1:N}, W)) && (11)\\
&= \arg\max_W \sum_{i=1}^{N} \log(p(y_i|X_i, W) && (12)\\
&= \arg\max_W \sum_{i=1}^{N} \frac{-1}{2}\log(2\pi\sigma) - \frac{(W^TX_i - y_i)^2}{2\sigma^2} && (13)\\
&= \arg\min_W \sum_{i=1}^{N} (W^TX_i - y_i)^2 && (14)
\end{aligned}$$

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 maximised by finding the value of W which cancel the gradient:

$$\begin{aligned}
&\nabla_W \sum_{i=1}^{N} (W^TX_i - y_i)^2 = 0 && (15)\\
&\sum_{i=1}^{N} \nabla_W (W^TX_i - y_i)^2 = 0 && (16)\\
&\sum_{i=1}^{N} 2X_i(W^TX_i - y_i) = 0 && (17)\\
&X^T(XW - y) = 0 && (18)\\
&X^TXW - X^Ty = 0 && (19)\\
&W - (X^TX)^{-1}X^Ty = 0 && (20)\\
&W = (X^TX)^{-1}X^Ty && (21)
\end{aligned}$$

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

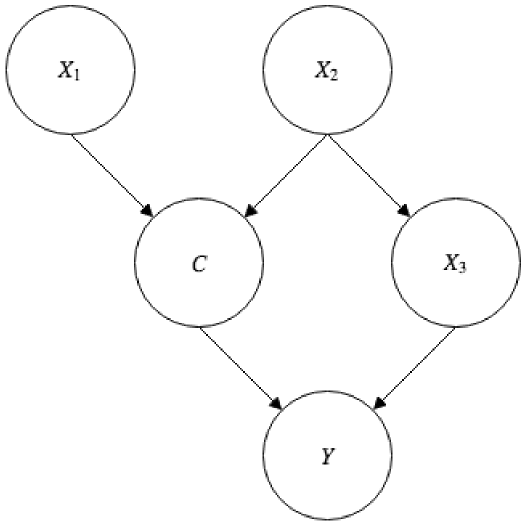

Figure 2: Bayesian network of the final grade calculation.