Contrôle périodique
INF8225 IA : techniques probabilistes et d'apprentissage

Toute documentation est autorisée.
Veuillez travailler individuellement. Le professeur Pal répondra aux questions dans le chat du Zoom.

Question (1)	Points
a	5
b	5
c	5
d	5
e	5
Total	25

Hiver 2021, le 26 févr. 15h au 18h.

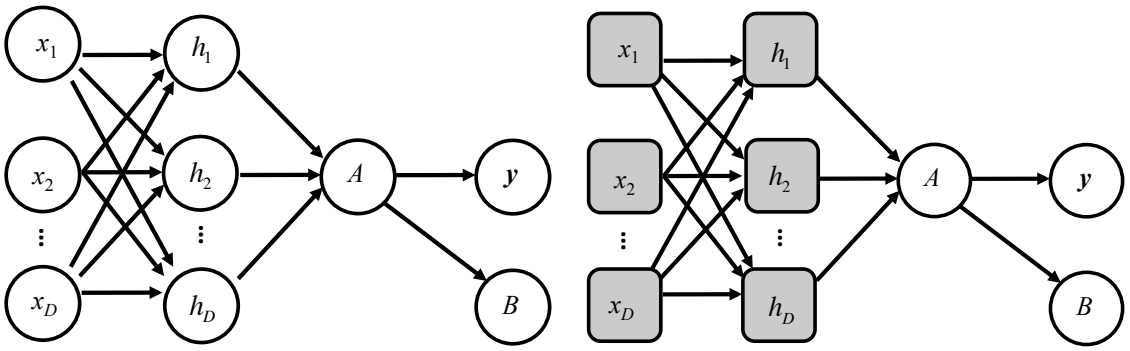

FIGURE 1 – (à gauche) Un réseau bayésien composé de variables binaires $x_1, \ldots, x_D$, $h_1, \ldots, h_D$ et A. La variable B est discrète et la variable $\mathbf{y}$ est une variable aléatoire continue bidimensionnelle. (à droite) Un modèle hybride où $x_1, \ldots, x_D$ sont des observations binaires, $h_1, \ldots, h_D$ sont les unités cachées d'un réseau de neurones et A est une variable aléatoire binaire, B est une variable aléatoire discrète et $\mathbf{y}$ est une variable aléatoire continue bidimensionnelle.

Question 1 (version française) (25 points)

a) (5 points) Pour le modèle de gauche, comment est le modèle de probabilité factorisé ? Donnez une équation. Expliquez comment un modèle de mélange gaussien à covariance diagonale pourrait être utilisé pour modéliser la relation entre A et $\mathbf{y}$.

b) (5 points) Pour le modèle de droite, en utilisant les définitions habituelles d'un réseau de neurones et d'un réseau bayésien, quelle probabilité est modélisée ici et comment le modèle de probabilité se décompose-t-il ? Donnez une équation qui indique également comment le réseau neuronal est construit. Utilisez les fonctions d'activation sigmoïde. Expliquez également avec des équations comment vous pouvez utiliser un modèle de mélange gaussien à covariance complète à deux composants pour capturer la relation entre A et $\mathbf{y}$.

c) (5 points) Considérez le graphique de gauche. Étant donné un ensemble de données où $D = 3$ et vous avez $N = 10000$ exemples constitués d'observations pour toutes les variables du graphique de gauche, expliquez comment vous apprendriez les paramètres pour chaque type de variable dans le graphique. Assurez-vous d'être clair sur ce que vous utilisez pour votre fonction objectif.

d) (5 points) Considérez le graphique de droite. Étant donné un ensemble de données où $D = 3$ et vous avez $N = 10000$ exemples constitués d'observations pour toutes les variables du graphique - à l'exception de $h_1, \ldots, h_D$, expliquez comment vous feriez l'apprentissage en ce modèle.

e) (5 points) Combien de paramètres y a-t-il dans le modèle à gauche par rapport au modèle à droite. Donnez des expressions mathématiques pour les deux en fonction de D et expliquez comment vous avez obtenu votre expression.

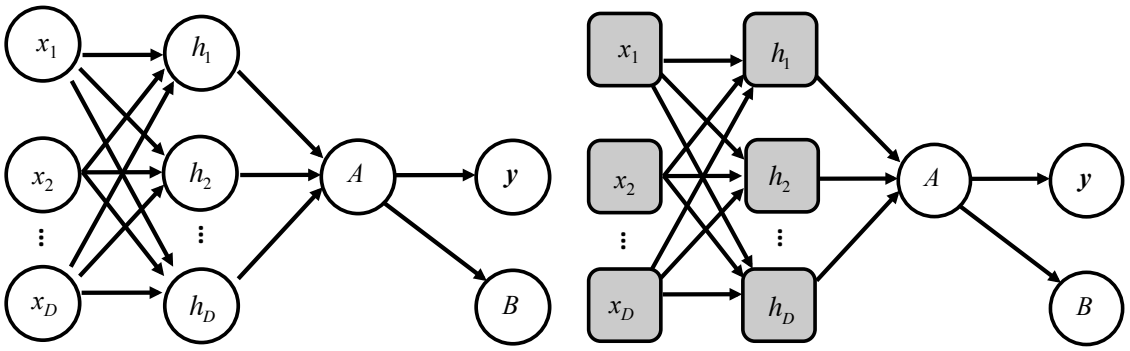

FIGURE 2 – (left) A Bayesian network consisting of binary variables $x_1, \ldots, x_D$, $h_1, \ldots, h_D$ and A. The variable B is discrete, and the variable $\mathbf{y}$ is a two dimensional continuous random variable. (right) A hybrid model where $x_1, \ldots, x_D$ are binary observations, $h_1, \ldots, h_D$ are the hidden units of a neural network, and A is a binary random variable, B is a discrete random variable and $\mathbf{y}$ is a two dimensional continuous random variable.

Question 1 (English version) (25 points)

a) (5 points) For the model on the left, how does the probability model factorize ? Give an equation. Explain how a diagonal covariance Gaussian mixture model could be used to model the relationship between A and $\mathbf{y}$. Use the definition of a Bayesian network from class 1b.

$$P(x_1, \ldots, x_D, h_1, \ldots, h_D, A, B, \mathbf{y}) = \prod_{i=1}^{D} \Big[P(x_i) P(h_i | x_1, \ldots, x_D) \Big] P(A | h_1, \ldots, h_D) P(\mathbf{y} | A) P(B | A) \tag{1}$$

If $P(\mathbf{y}|A)$, is a diagonal covariance Gaussian mixture, we could think of there as being two different mean vectors $\boldsymbol{\mu}_{a=1}$, and $\boldsymbol{\mu}_{a=2}$ and two different covariance matrices $\boldsymbol{\Sigma}_a = \text{diag}(\sigma_{1,a}^2, \sigma_{2,a}^2)$ and we could write the model as :

$$P(\mathbf{y}|A) = \mathcal{N}\big(\mathbf{y}; \boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a\big) = \frac{1}{\sqrt{2\pi}\sigma_{1,a}} \exp\Big[- \frac{(y_1 - \mu_{1,a})^2}{2\sigma_{1,a}^2} \Big] \frac{1}{\sqrt{2\pi}\sigma_{2,a}} \exp\Big[- \frac{(y_2 - \mu_{2,a})^2}{2\sigma_{2,a}^2} \Big] \tag{2}$$

b) (5 points) For the model on the right, using the usual definitions of a neural network and a Bayesian network, what probability is modeled here and how does the probability model decompose ? Give an equation which also indicates how the neural network is constructed. Use sigmoid activation functions. Also explain with equations how you could use a two component, full covariance Gaussian mixture model to capture the relationship between A and $\mathbf{y}$.

$$P(A, B, \mathbf{y} | x_1, \ldots, x_D) = P(A|\mathbf{x}) P(\mathbf{y}|A) P(B|A), \tag{3}$$

where

$$P(A|\mathbf{x}) = \text{Bernoulli}(A, f(\mathbf{h}(\mathbf{x}))) = \text{Bern}(A, \text{sigmoid}(\mathbf{w}^T \text{sigmoid}(\mathbf{W}\mathbf{x} + \mathbf{b}) + b)), \tag{4}$$

and where

$$P(\mathbf{y}|A) = \mathcal{N}\big(\mathbf{y}; \boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a\big) = (2\pi)^{-1} |\boldsymbol{\Sigma}_a|^{-\frac{1}{2}} \exp\Big[- \frac{1}{2} (\mathbf{y} - \boldsymbol{\mu}_a)^T \boldsymbol{\Sigma}_a (\mathbf{y} - \boldsymbol{\mu}_a) \Big] \tag{5}$$

c) (5 points) Consider the graph on the left. Given a data set where $D = 3$ and you have $N = 10,000$ examples consisting of observations for all of the variables in the graph on the left, explain how you would learn the parameters for each type of variable in the graph. Be sure to be clear about what you are using for your objective function.

Our objective function would be :

$$\begin{aligned} &-\log P(x_1,\ldots,x_D,h_1,\ldots,h_D,A,B,\mathbf{y}) = \\ &-\sum_{i=1}^{D}\log P(x_i) - \sum_{i=1}^{D}\log P(h_i|x_1,\ldots,x_D) - \log P(A|h_1,\ldots,h_D) - \log P(\mathbf{y}|A) - \log P(B|A) \end{aligned} \tag{6}$$

summed over each of the $j = 1 \ldots N$ examples $\{x_1,\ldots,x_D,h_1,\ldots,h_D,A,B,\mathbf{y}\}_j$ in our data set. In terms of parameter updates the problem decouples for each term such that for the parameters of each $P(x_i)$ we have : (note see lesson 1b)

$$P(x_i = x) = \frac{1}{N}\sum_{j=1}^{N}\mathbf{1}(\tilde{x}_{i,j} = x). \tag{7}$$

For $P(B|A)$ we have

$$P(B = b|A = a) = \frac{\sum_{j=1}^{N}\mathbf{1}(\tilde{B}_j = b, \tilde{A}_j = a)}{\sum_{j=1}^{N}\mathbf{1}(\tilde{A}_j = a)} \tag{8}$$

For each $P(h_i|x_1,\ldots,x_D)$ we have

$$P(h_i = h|x_1 = X_1,\ldots,x_D = X_D) = \frac{\sum_{j=1}^{N}\mathbf{1}(\tilde{h}_{i,j} = h, \tilde{x}_{1,j} = X_1,\ldots,\tilde{x}_{D,j} = X_D)}{\sum_{j=1}^{N}\mathbf{1}(\tilde{x}_{1,j} = X_1,\ldots,\tilde{x}_{D,j} = X_D)} \tag{9}$$

For the parameters of $P(\mathbf{y}|A)$ we have

$$\boldsymbol{\mu}_a = \frac{\sum_{j=1}^{N}\mathbf{1}(\tilde{A}_j = a)\mathbf{y}_j}{\sum_{j=1}^{N}\mathbf{1}(\tilde{A}_j = a)} \tag{10}$$

$$\sigma^2_{1,a} = \frac{\sum_{j=1}^{N}\mathbf{1}(\tilde{A}_j = a)(y_{1,j} - \mu_{1,a})^2}{\sum_{j=1}^{N}\mathbf{1}(\tilde{A}_j = a)} \tag{11}$$

d) (5 points) Consider the graph on the right. Given a data set where $D = 3$ and you have $N = 10,000$ examples consisting of observations for all of the variables in the graph – except for $h_1,\ldots,h_D$, explain how you would perform learning in this model. For the graph on the right, our objective function - for a single example - would be :

$$-\log P(\mathbf{y},B,A|x_1,\ldots,x_D) = -\log P(A|x_1,\ldots,x_D) - \log P(\mathbf{y}|A) - \log P(B|A) \tag{12}$$

For $P(B|A)$, we proceed in the same way as above - we have :

$$P(B = b|A = a) = \frac{\sum_{j=1}^{N}\mathbf{1}(\tilde{B}_j = b, \tilde{A}_j = a)}{\sum_{j=1}^{N}\mathbf{1}(\tilde{A}_j = a)} \tag{13}$$

For the parameters of $P(\mathbf{y}|A)$ the mean is estimated in the same way as above, but we need to use the matrix form to express how the covariance would be estimated

$$\boldsymbol{\mu}_a = \frac{\sum_{j=1}^{N}\mathbf{1}(\tilde{A}_j = a)\mathbf{y}_j}{\sum_{j=1}^{N}\mathbf{1}(\tilde{A}_j = a)} \tag{14}$$

$$\boldsymbol{\Sigma}^2_a = \frac{\sum_{j=1}^{N}\mathbf{1}(\tilde{A}_j = a)(\mathbf{y}_j - \boldsymbol{\mu}_a)(\mathbf{y}_j - \boldsymbol{\mu}_a)^T}{\sum_{j=1}^{N}\mathbf{1}(\tilde{A}_j = a)} \tag{15}$$

For $-\log P(A|x_1, \ldots, x_D)$, this is a neural network with a fixed hidden layer the same size as the input, just learn it with SGD using the data for $x_1, \ldots, x_D$ and A.

e) (5 points) How many parameters are in the model on the left versus the model on the right. Give mathematical expressions for both as a function of D and explain how you obtained your expression.

Consider each term from the loss for the left model

$$-\sum_{i=1}^{D} \log P(x_i) \Rightarrow 2D \tag{16}$$

$$-\sum_{i=1}^{D} \log P(h_i|x_1, \ldots, x_D) \Rightarrow D \cdot 2^{D+1} \tag{17}$$

$$-\log P(A|h_1, \ldots, h_D) \Rightarrow 2^{D+1} \tag{18}$$

$$-\log P(\mathbf{y}|A) \Rightarrow 2 \cdot 4 = 8 \tag{19}$$

$$-\log P(B|A) \Rightarrow 2s \tag{20}$$

$$\tag{21}$$

where discrete variable B has s states, which gives us : $(D+1)2^{D+1} + 2D + 2s + 8$ parameters. We assume there are two means and two diagonal covariance matrices, in two dimensions. If $s = 2$ (it was a two state or binary distribution) we would have : $(D+1)2^{D+1} + 2D + 12$

Consider each term from the loss for the right model, where the neural network is of the form :
$P(A|\mathbf{x}) = \text{Bern}(A, \text{sigmoid}(\mathbf{w}^T \text{sigmoid}(\mathbf{W}\mathbf{x} + \mathbf{b}) + b))$

$$-\log P(A|x_1, \ldots, x_D) \Rightarrow D + D^2 + D + 1 = 2D + D^2 + 1 \tag{22}$$

$$-\log P(\mathbf{y}|A) \Rightarrow 2 \cdot 6 = 12 \tag{23}$$

$$-\log P(B|A) \Rightarrow 2s \tag{24}$$

$$\tag{25}$$

which gives us : $D^2 + 2D + 12 + 2s$