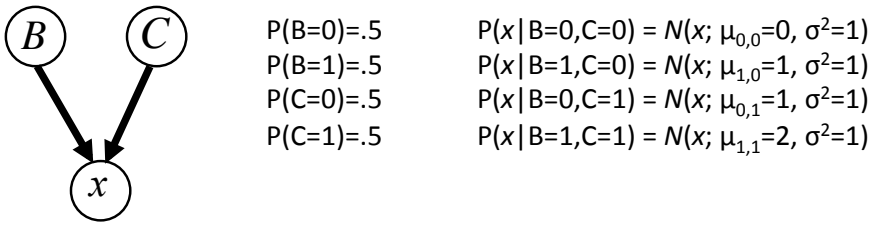


FIGURE 1 – Un réseau bayésien et les probabilités inconditionnelles et conditionnelles associées pour les variables binaires B et C et une variable continue à valeur réelle x. La notation $\mathcal{N}(x; \mu_{b,c}, \sigma^2)$ dénote une distribution normale avec une moyenne donnée par $\mu_{b,c}$ et une variance donnée par σ^2.

Question 1 (version française) (20 points)

Considérons un ensemble de trois variables aléatoires $\{B, C, x\}$, où B et C sont binaires et x est une variable aléatoire continue à valeur réelle. Soit ces états notés 0 ou 1 pour B et C. Le réseau bayésien ci-dessus illustre comment la distribution conjointe est factorisée. Rappelons qu'en une dimension, nous pouvons écrire la distribution gaussienne comme suit :

$$P(x) = \mathcal{N}\big(x; \mu, \sigma\big) \rightarrow P(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\Big]. \tag{1}$$

a) (3 points) Quelle est la distribution de probabilité associée au modèle de la figure ci-dessus et comment se factorise-t-elle ? (Fournir une équation)

b) (2 points) Qu'est-ce que $P(x)$? (Donner une équation et dessinez une figure approximative de la fonction de densité de probabilité, c.-à-d. $p(x)$ vs x.)

Pour les questions suivantes, montrez les équations de probabilité que vous utilisez pour obtenir votre réponse, ainsi que votre réponse numérique finale.

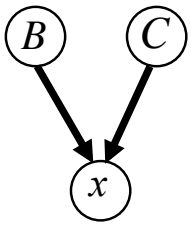
c) (3 points) Qu'est-ce que $P(B = 1|x = 1)$? (Donner une équation et une valeur numérique.)

d) (3 points) Qu'est-ce que $P(B = 1|x = 1, C = 1)$? (Donner une équation et une valeur numérique.)

e) (3 points) Qu'est-ce que $P(B = 1|x = 1, C = 0)$? (Donner une équation et une valeur numérique.)

f) (3 points) Qu'est-ce que $P(B = 1|x = 2)$? (Donner une équation et une valeur numérique.)

g) (3 points) Ce réseau modélise une forme d'addition bruitée en utilisant des variables parentes binaires et une variable enfant continue. Si B et C étaient redéfinis pour être des variables aléatoires continues à valeurs réelles b et c, comment construiriez-vous un modèle analogue qui implémente l'idée de x résultant de l'addition de b et c avec bruit gaussien ?



P(B=0)=.5
P(B=1)=.5
P(C=0)=.5
P(C=1)=.5

$P(x|B=0,C=0) = N(x; \mu_{0,0}=0, \sigma^2=1)$
$P(x|B=1,C=0) = N(x; \mu_{1,0}=1, \sigma^2=1)$
$P(x|B=0,C=1) = N(x; \mu_{0,1}=1, \sigma^2=1)$
$P(x|B=1,C=1) = N(x; \mu_{1,1}=2, \sigma^2=1)$

FIGURE 2 – A Bayesian network and the associated unconditional and conditional probabilities for binary variables B and C and continuous, real valued variable x. The notation $\mathcal{N}(x; \mu_{b,c}, \sigma^2)$ denotes a normal distribution with mean given by $\mu_{b,c}$ and variance given by σ^2.

Question 1 (English version) (20 points)

Consider a set of three random variables $\{B, C, x\}$, where B and C are binary and x is a continuous real valued random variable. Let these states be denoted as 0 or 1 for B and C. The Bayesian network above illustrates how the joint distribution is factorized. Recall that in one dimension we can write the Gaussian distribution as :

$$P(x) = \mathcal{N}\big(x; \mu, \sigma\big) \rightarrow P(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big[-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2} \Big]. \tag{2}$$

a) (3 points) What is the probability distribution associated with the model in the Figure above and how does it factorize ? (Provide an equation)

b) (2 points) What is $P(x)$? (Give an equation and draw an approximate figure of the probability density function, i.e. $p(x)$ vs x.)

For the following questions show the equations of probability you use to obtain your answer, as well as your final numerical answer.

c) (3 points) What is $P(B = 1|x = 1)$? (Give an equation and a numerical value.)

d) (3 points) What is $P(B = 1|x = 1, C = 1)$? (Give an equation and a numerical value.)

e) (3 points) What is $P(B = 1|x = 1, C = 0)$? (Give an equation and a numerical value.)

f) (3 points) What is $P(B = 1|x = 2)$? (Give an equation and a numerical value.)

g) (3 points) This network models a form of noisy addition using binary parent variables and a continuous child variable. If B and C were redefined to be continuous real valued random variables b and c, how would you construct an analogous model that implements the idea of x arising from the addition of b and c with Gaussian noise ?

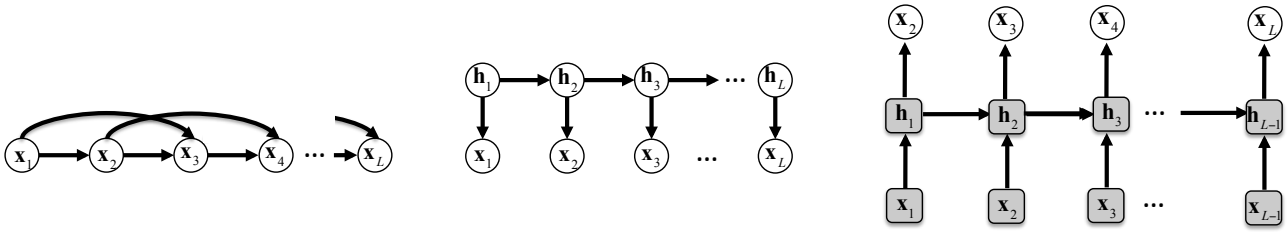


FIGURE 3 – (À gauche) Un modèle de Markov du second ordre. (Au milieu) Un modèle de Markov caché (HMM). (À droite) Un réseau neuronal récurrent (RNN).

Question 2 (version française) (20 points)

Pour la figure ci-dessus, imaginez que pour $\mathbf{x}$ et $\mathbf{y}$ vous avez des vecteurs one-hot représentant des mots d'un vocabulaire de 50000 mots possibles. Dans la figure du milieu, $\mathbf{h}$ représente des variables aléatoires discrètes avec 100 états. Pour la figure de droite, $\mathbf{h}$ est une unité cachée de 100 dimensions. Pour chacun des modèles à gauche, au milieu et à droite, fournissez l'analyse demandée pour chaque modèle pour les questions a), b) et c).

a) (6 points au total, 2 par modèle) i) Écrivez le modèle mathématique associé à chaque graphique. Puis, imaginez maintenant que vous souhaitiez apprendre les paramètres de chaque modèle à partir de $N = 100000$ séquences de la forme $\mathbf{x}_1, \ldots, \mathbf{x}_L$, $L = 100$. ii) Quelle serait votre fonction objectif pour chaque modèle ?

b) (6 points au total, 2 par modèle) Fournissez une expression pour le nombre de paramètres et le nombre de paramètres libres que chacun de vos modèles aurait. Notez que vous devez créer votre expression sous une forme générale, en utilisant L pour la longueur de la séquence, v pour la taille du vocabulaire et d pour la dimensionnalité de la variable cachée. Expliquez également comment vous avez obtenu votre expression.

c) (6 points au total, 2 par modèle) Imaginez que vous devez faire une prédiction sur un ensemble de test. i) Fournissez une équation expliquant comment calculeriez-vous la probabilité suivante pour chaque modèle : $P(\mathbf{x}_3|\mathbf{x}_1, \mathbf{x}_2)$. ii) Compte tenu des observations pour $\mathbf{x}_1$ et $\mathbf{x}_2$, comment calculeriez-vous la configuration la plus probable des deux mots suivants, c'est-à-dire de $\mathbf{x}_3$ et $\mathbf{x}_4$ pour chaque modèle ?

d) (2 points) Considérons maintenant uniquement le modèle de gauche. Si ce modèle a été étendu pour devenir un modèle de Markov d'ordre k, fournissez une expression pour le nombre de paramètres nécessaires pour créer le modèle en fonction de k.

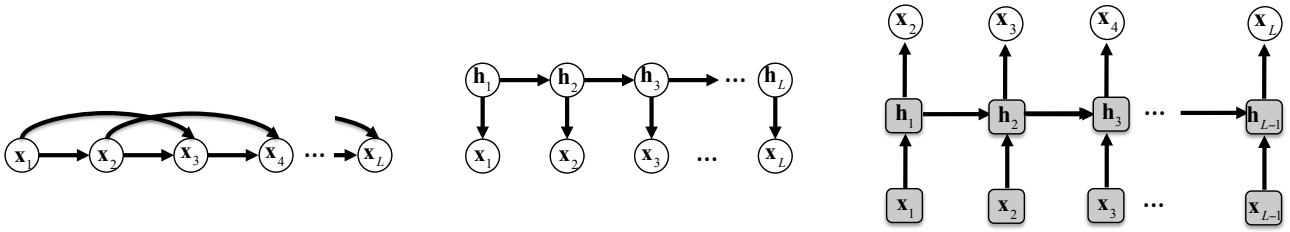


FIGURE 4 – (Left) A second order Markov model. (Middle) A Hidden Markov Model (HMM). (Right) A Recurrent Neural Network (RNN)

Question 2 (English version) (20 points)

For the figure above imagine that for $\mathbf{x}$ and $\mathbf{y}$ you have one-hot vectors representing words from a vocabulary of $50,000$ possible words. In the middle figure $\mathbf{h}$ represents discrete random variables with 100 states. For the figure on the right $\mathbf{h}$ is a hidden unit with 100 dimensions. For each of the models on the left, middle and right, provide the requested analysis for each model for questions a), b) and c).

a) (6 points in total, 2 per model) i) Write the mathematical model associated with each graph. Then imagine that you want to learn the parameters of each model from $N = 100,000$ sequences of the form $\mathbf{x}_1, \ldots, \mathbf{x}_L$, $L = 100$. ii) What would be your objective function for each model ?

b) (6 points in total, 2 per model) Provide an expression for the number of parameters and the number of free parameters that each of your models would have. Note that you should create your expression in a general form, using L for the length of the sequence, v for the vocabulary size and d for the dimensionality of the hidden variable. Also provide some explanation of how you obtained your expression.

c) (6 points in total, 2 per model) Imagine you need to make a prediction on a test set. i) Provide an equation explaining how would you compute the following probability for each model : $P(\mathbf{x}_3|\mathbf{x}_1, \mathbf{x}_2)$. ii) Given observations for $\mathbf{x}_1$ and $\mathbf{x}_2$, how would you compute the most probable configuration of the next two words, i.e. of $\mathbf{x}_3$ and $\mathbf{x}_4$ for each model ?

d) (2 points) Consider now just the model on the left. If this model was extended to be a k^{th} order Markov model, provide an expression for the number of parameters needed to create the model as a function of k.

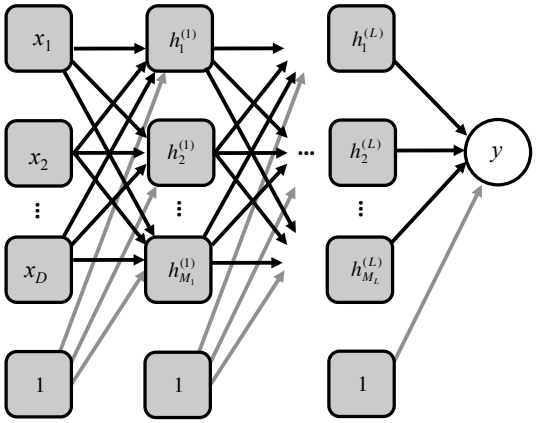


FIGURE 5 – Un réseau de neurones.

Question 3 (version française) (10 points)

Considérons un réseau de neurones avec une seule couche d'entrée avec $D = 3$ unités, L couches cachées, chacun avec 3 unités et un seul neurone de sortie binaire. Vous avez $i = 1, .., N$ exemples $y_i \in \{0, 1\}$, $\mathbf{x}_i \in \mathbb{R}^3$ dans un ensemble d'apprentissages. On a une fonction de perte donnée par

$$L = -\sum_{i=1}^{N} \big(y_i \log(f(\mathbf{x}_i)) + (1 - y_i) \log(1 - f(\mathbf{x}_i)\big), f(\mathbf{x}_i) = f(a^{(L+1)}(\mathbf{h}^{(L)}(\mathbf{a}^{(L)}(\ldots \mathbf{h}^{(1)}(\mathbf{a}^{(1)}(\mathbf{x}_i)))))) \quad (3)$$

où $\mathbf{x}_i$ est un exemple d'entrée, y_i un cible binaire. La fonction d'activation f de la couche finale ayant la forme d'une fonction sigmoïde, ainsi que toutes les couches intermédiaires, de sorte que

$$f(a^{(L+1)}(\mathbf{x}_i)) = \frac{1}{1 + \exp(-a^{(L+1)}(\mathbf{h}^{(L)}(\mathbf{x}_i)))}, \quad h_k^{(l)}(a_k{}^{(l)}(\mathbf{x}_i)) = \text{sigmoid}(a_k{}^{(l)}(\mathbf{x}_i)), \quad (4)$$

où c est une constante scalaire, $\mathbf{a}^{(l)} = [a_1^{(l)} ... \ a_K^{(l)}]^T$ est le vecteur résultant du calcul de la préactivation habituelle $\mathbf{a}^{(l)} = \mathbf{W}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}$, qui pourrait être simplifié à $\boldsymbol{\theta}^{(l)}\hat{\mathbf{h}}^{(l-1)}$ en utilisant l'astuce de définir $\hat{\mathbf{h}}$ comme $\mathbf{h}$ avec un 1 concaténé à la fin du vecteur.

a) (2 points) Qu'est-ce que $\frac{\partial L}{\partial f}$?

b) (2 points) Qu'est-ce que $\frac{\partial a^{(L+1)}}{\partial \mathbf{h}^{(L)}}$ et $\frac{\partial \mathbf{a}^{(l+1)}}{\partial \mathbf{h}^{(l)}}, l \neq L + 1$?

c) (2 points) Imaginez maintenant que le premier exemple dans les données d'apprentissage donne un vecteur de préactivation pour la première couche consistant en $\mathbf{a}^{(1)} = [-2.0, 0, 16.5]^T$, qu'est-ce que $\frac{\partial \mathbf{h}^{(1)}}{\partial \mathbf{a}^{(1)}}$?

d) (2 points) Vous voulez apprendre un modèle avec L=50 couches. Expliquez comment vous pouvez modifier le modèle en utilisant l'idée de connexion résiduelle de ResNets et réécrivez $\mathbf{h}^l(\mathbf{x}_i)$ pour montrer votre modification. Expliquez comment cette modification pourrait vous permettre d'apprendre un modèle avec L=50 couches ?

e) (2 points) Si vous aviez besoin d'utiliser ce réseau pour effectuer une classification multi-classes, expliquez comment vous pourriez modifier la perte et la couche finale.

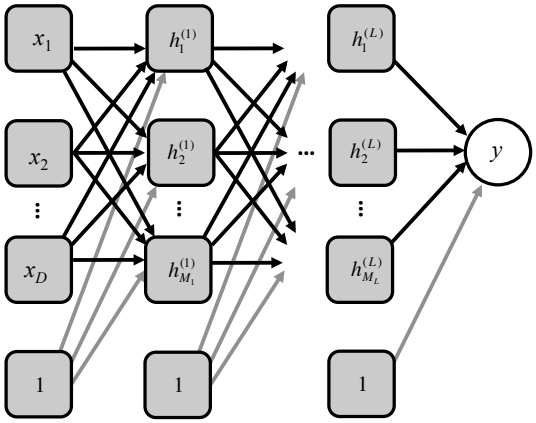


FIGURE 6 – A neural network.

Question 3 (English version) (10 points)

Consider a neural network with a single input layer having $D = 3$ units, L hidden layers, each having 3 units and a single binary output. You have $i = 1, .., N$ examples in a training set where the labels are binary $y_i \in \{0, 1\}$, the input is a vector of continuous real values $\mathbf{x}_i \in \mathbb{R}^3$. The loss function is given by the binary cross entropy

$$L = -\sum_{i=1}^{N} \big(y_i \log(f(\mathbf{x}_i)) + (1 - y_i) \log(1 - f(\mathbf{x}_i)\big), f(\mathbf{x}_i) = f(a^{(L+1)}(\mathbf{h}^{(L)}(\mathbf{a}^{(L)}(\ldots \mathbf{h}^{(1)}(\mathbf{a}^{(1)}(\mathbf{x}_i)))))) \quad (5)$$

where $\mathbf{x}_i$ is an input example, y_i is a continuous scalar value which is the prediction target. The output activation function f has the form of a sigmoid and so do all the intermediate layers, such that

$$f(a^{(L+1)}(\mathbf{x}_i)) = \frac{1}{1 + \exp(-a^{(L+1)}(\mathbf{h}^{(L)}(\mathbf{x}_i)))}, \quad h_k^{(l)}(a_k{}^{(l)}(\mathbf{x}_i)) = \text{sigmoid}(a_k{}^{(l)}(\mathbf{x}_i)), \quad (6)$$

where $\mathbf{a}^{(l)} = [a_1^{(l)} ... \ a_K^{(l)}]^T$ is the vector resulting from the calculation of the usual preactivation function $\mathbf{a}^{(l)} = \mathbf{W}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}$, which could be simplified to $\boldsymbol{\theta}^{(l)}\hat{\mathbf{h}}^{(l-1)}$ using the trick of defining $\hat{\mathbf{h}}$ as $\mathbf{h}$ with a 1 concatenated at the end.

a) (2 points) What is $\frac{\partial L}{\partial f}$?

b) (2 points) What is $\frac{\partial a^{(L+1)}}{\partial \mathbf{h}^{(L)}}$ and $\frac{\partial \mathbf{a}^{(l+1)}}{\partial \mathbf{h}^{(l)}}, l \neq L + 1$?

c) (2 points) Imagine now that the first example in the training data yields a preactivation vector for the first layer consisting of $\mathbf{a}^{(1)} = [-2.0, 0, 16.5]^T$, what is $\frac{\partial \mathbf{h}^{(1)}}{\partial \mathbf{a}^{(1)}}$?

d) (2 points) You want to learn a model with L=50 layers. Explain how you could modify the model using the residual connection idea from ResNets, and rewrite $\mathbf{h}^l(\mathbf{x}_i)$ to show your modification. Explain how this modification might allow you to learn a model with L=50 layers ?

e) (2 points) If you needed to use this network to perform a multi-class classification, explain how you could modify the loss and the final layer.