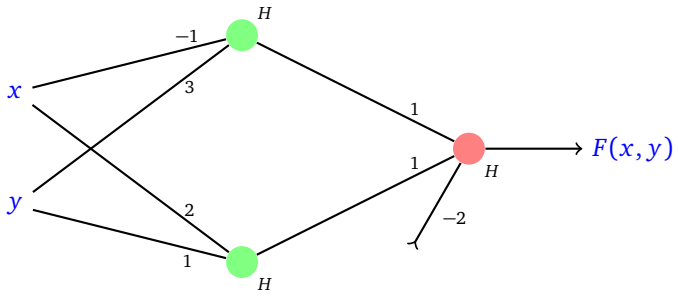
Examen

Durée: 1h30.
Le sujet compte 2 pages.
Sont autorisés pour cet examen :

- une antisèche de format A4,
- un ordinateur strictement réservé à l'utilisation de Python et de Moodle (section "Examen" du cours),
- vos scripts des TD.

Le squelette du script Python `examen_MDS_NOM_Prenom.py` se trouve sur Moodle. Il est à rendre complété ainsi que tout autre document numérique pertinent (image etc...).
Le barème est donné à titre indicatif.

Exercice 1 (5 points). On considère le réseau de neurones suivant.



1. Détailler le raisonnement permettant de connaître la valeur de $F(x, y)$ selon l'entrée (x, y).
2. Compléter le script Python afin d'implémenter ce réseau et de retrouver le résultat.

Exercice 2 (5 points). Les deux questions de cet exercice sont indépendantes.

1. Un match de tennis classique (hors tournois du grand chelem), opposant un joueur A à un joueur B, est constitué d'au plus trois sets. Le vainqueur du match est le premier des deux joueurs à gagner deux sets. Soit X la variable aléatoire donnant le résultat des sets d'un match entre les joueurs A et B : par exemple, X prend la valeur AA si A a gagné les deux premiers sets (et alors le match s'arrête), ou, autre exemple, X prend la valeur BAB si le joueur B a gagné les sets 1 et 3, et le joueur A a gagné le set 2. Soit Y le nombre total de sets effectués lors d'un match. On suppose que les deux joueurs gagnent le set avec équiprobabilité.
   (a) Déterminer la loi jointe du couple (X, Y).
   (b) Calculer l'entropie de X et l'entropie de Y. Quelle variable aléatoire est la plus sujette au hasard ?
2. Rappeler le lien entre la divergence KL et l'entropie croisée. Quel est l'intérêt de minimiser la divergence KL ? l'entropie croisée ?

Exercice 3 (10 points). On considère les données du prix de logements en Californie (regroupés par groupes) [1], dans le fichier `housing.csv` à télécharger sur Moodle.

group	MedInc	HouseAge	...	Longitude	Price
1	8.3252	41.	...	-122.23	4.526
2	8.3014	21.	...	-122.22	3.585
⋮	⋮	⋮	⋮	⋮	
20640	2.3886	16.	...	-121.24	0.894

[1] `https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_california_housing.html#sklearn.datasets.fetch_california_housing`

Détail des variables :

- `MedInc` : revenu médian
- `HouseAge` : age médian
- etc... [2]
- `Longitude` : longitude
- `Price` : prix (en centaines de milliers de dollars)

Le but de cet exercice est de construire un modèle de régression linéaire pour expliquer la variable `Price` à partir des autres variables explicatives `MedInc`, `HouseAge`, ..., et `Longitude`. On demandera d'arrondir les résultats au centième.

1. Effectuer une standardisation des données. Donner la valeur du revenu médian du premier groupe suite à cette standardisation.
2. Donner la solution au problème de régression linéaire au sens des moindres carrés, sans régularisation.
3. Quelle valeur de la variable `Price` renvoie ce modèle si l'on donne en entrée les valeurs des variables explicatives du premier groupe ? Commenter.
4. On souhaite dans cette question étudier l'effet de l'ajout d'une régularisation Lasso.
   (a) Pour une valeur du paramètre de régularisation au choix, donner la solution du problème de régression linéaire au sens des moindres carrés, avec régularisation Lasso.
   (b) Donner la représentation du chemin de régularisation des différents coefficients en fonction de la valeur de paramètre de régularisation. Quel ordre de grandeur semble pertinent pour ce paramètre ?
   (c) En déduire la variable explicative semblant avoir le plus d'effet sur la variable expliquée ?

[2] `https://scikit-learn.org/stable/datasets/real_world.html#california-housing-dataset`