

# IFT 603-712 : Devoir 1

## Travail par équipe de 3

Remettez votre solution aux numéros 1 et 2 en format pdf ou manuscrit (et scanné) via [turninweb](#). Même chose pour le code.

1. [1 point] Démontrez la propriété de l'entropie suivante :

$$H[x,y] = H[y|x] + H[x]$$

Vous pouvez faire la démonstration pour le cas discret ou continu.

2. [1.5 points] Démontrez la propriété de l'information mutuelle suivante :

$$I[x,y] = H[x] - H[x|y]$$

Vous pouvez faire la démonstration pour le cas discret ou continu.

3. [1 point] Démontrez la propriété suivante :

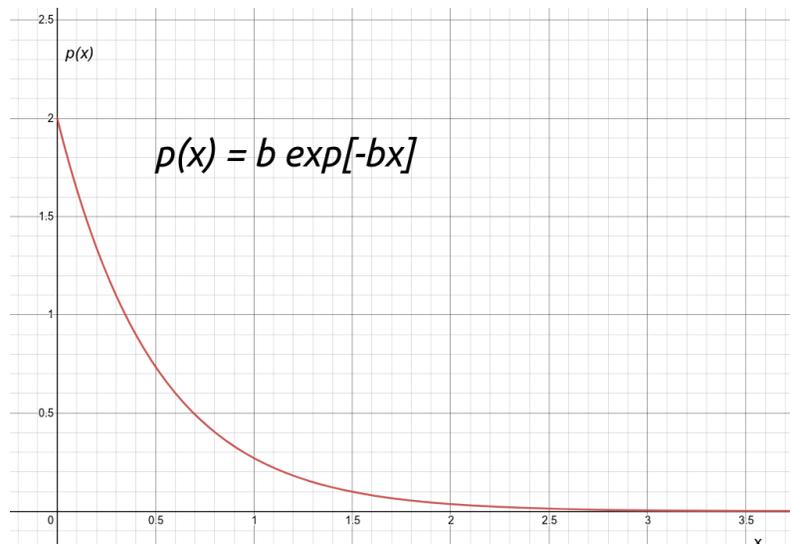
$$\text{Cov}[x,y] = E_{xy}[xy] - E[x]E[y]$$

où Cov est la fonction de covariance et E est l'espérance mathématique. Vous pouvez faire la démonstration pour le cas discret ou continu.

4. [0.5 point] Soit une variable aléatoire binaire  $x$  ayant produit la séquence d'observations suivante :  $\{0,0,0,1,0,1,0,0,1\}$ . Partant de ces observations, donnez (a) la distribution  $P(x)$  également appelée « l'évidence » (b) l'espérance mathématique de  $x$ , (c) la variance de  $x$  et (d) l'entropie de  $x$ . Précisez vos calculs.

5- [1 point] Une entreprise d'assainissement des eaux vous engage pour l'aider à mieux analyser ses échantillons d'eau. Chaque échantillon possède une fiche signalétique indiquant une concentration en plomb. Dans les échantillons d'eau propre à la consommation, on mesure en moyenne 1 microgramme de plomb par litre. Cependant, les échantillons impropres à la consommation ont des concentrations en plomb pouvant aller jusqu'à 5 à 15 microgrammes par litre. Cependant, comme la compagnie n'a aucune compétence en analyse de données, elle s'appuie sur l'avis d'un expert pour déterminer si un échantillon d'eau est potable ou non.

- À l'aide d'une **base de données** annotée, d'une distribution **exponentielle** (voir figure suivante) et d'une distribution **gaussienne**, dites comment vous procéderiez pour obtenir la distribution de vraisemblance de chaque classe (indice: l'espérance mathématique d'une distribution exponentielle est  $1/b$  où  $b$  est le paramètre mentionné à la figure suivante).
- À l'aide de ces deux distributions, donnez un algorithme de maximum de vraisemblance vous permettant de classer un nouvel échantillon sur la base de sa concentration en plomb.



**6. [5 points]** Programmez les algorithmes de régression linéaire et non linéaire polynomial vus aux chapitres 1 et 3 ainsi que la recherche d'hyperparamètre « k-fold cross-validation » vu au chapitre 1. Pour ce faire, vous devez télécharger le fichier **tp1.zip** du site web du cours.

L'algorithme doit être implémenté via la classe **Regression**. Votre implémentation de cette classe doit être placée dans le fichier **solution\_regression.py** qui contient déjà une ébauche. Veuillez vous référer aux commentaires sous la signature de chaque méthode de la classe **Regression** afin de savoir quoi implémenter. Votre implémentation doit être efficace et utiliser les fonctionnalités des bibliothèques Numpy et Sklearn. Les fonctions qu'il vous faut compléter sont identifiées par le commentaire « AJOUTER CODE ICI ».

De plus, dans le fichier pdf que vous remettrez, veuillez SVP inclure **un résultat graphique** illustrant un cas de sur-apprentissage ainsi qu'un cas de sous-apprentissage. Dans les deux cas, veuillez préciser les paramètres utilisés.

**Note 1**: bien que vide, le code de la classe **Regression** s'exécute déjà. Pour vous en convaincre, vous n'avez qu'à taper la commande suivante dans un terminal linux :

```
python regression.py tanh 100 20 0.0 1 0.000001 --sklearn
```

En tapant :

```
python3 regression.py -h
```

vous verrez à quoi correspondent les arguments de ce programme.

**Note 2** : le code des devoirs a été testé avec des versions de python 3.7 et 3.8. Pour faire fonctionner votre code sur les ordinateurs du département, vous devez

1. démarrer une session linux (ubuntu)
2. démarrer un terminal
3. activer (ou créer) un environnement virtuel python et y installer les packages requis (voir site web du cours si vous ne savez pas comment)
3. normalement, si vous démarrez une session ipython (taper ipython dans le terminal) vous verrez « Python 3.8 » (ou une version plus récente).

**Note 3** : il est recommandé de rédiger son code dans un ide comme spyder ou pycharm.

**Note 4** : pour exécuter le code des notebooks disponibles sur le site web du cours, vous devez taper la commande « jupyter notebook » dans un terminal.

**Note 5** : en plus de votre code, veuillez soumettre un fichier "**gitlab.txt**" dans lequel vous donnez le lien vers votre dépôt gitlab. Il est obligatoire d'utiliser le service gitLab de l'université (gitHub, Bitbucket, etc. sont interdits.)

**Note 6** : en plus de vos documents de travail, vous devez remplir, signer et joindre à votre travail le **formulaire d'intégrité** disponible sur le site web du cours.

**Note 7**: en plus de vos documents de travail, vous devez remplir et joindre à votre travail votre **agenda d'équipe** disponible sur le site web du cours.