remarques : Pour cet examen les élèves ont le droit à la table des valeurs pour le TCL (il s'agit d'une feuille A4 avec uniquement un tableau et aucun commentaire manuscrit).

exercice 1

Question 1 : $A \in \mathbb{R}^{n\times n}$ est une matrice inversible si et seulement si

- ☑ A est de rang n
- ☑ $det(A) \neq 0$
- ☑ il existe B telle que $AB = I = BA$
- ☑ il existe une séquence de multiplications matricielles inversibles $(M_1, ...M_m)$ telle que $M_m \times \cdots \times M_1 \times A = I$
- ☑ les lignes de A forment une base

Question 2 : La chain rule

- ☑ permet de calculer le gradient d'une fonction composée de deux fonctions
- ☑ permet de calculer le gradient d'une fonction composée de n fonctions
- ☐ pour $h \circ g \circ f$ nécessite le calcul de ∇f, $(\nabla g) \circ (f)$ et $((\nabla h) \circ (\nabla g \circ (\nabla f))$
- ☐ pour $h \circ g \circ f$ nécessite le calcul de ∇f, $(\nabla g) \circ (f)$ et $(\nabla h) \circ (g \circ (\nabla f))$

Question 3 : La méthode des pivots de Gauss permet d'inverser

- ☐ n'importe quelle matrice
- ☑ n'importe quelle matrice inversible
- ☐ n'importe quelle matrice diagonale
- ☐ toutes les reponses précédentes

Question 4 : Quel énoncé du TCL est correct ?

- ☐ $\frac{\sum_{i=1}^n X_i - \mu}{\sigma\sqrt{n}} \to \mathcal{N}(0,1)$
- ☑ $\frac{(\sum_{i=1}^n X_i) - \mu}{\sigma\sqrt{n}} \to \mathcal{N}(0,1)$
- ☐ $\sqrt{n}\frac{\sum_{i=1}^n X_i - \mu}{\sigma} \to \mathcal{N}(0,1)$
- ☐ $\sqrt{n}\frac{(\sum_{i=1}^n X_i) - \mu}{\sigma\sqrt{n}} \to \mathcal{N}(0,1)$

Question 5 : Á quoi sert l'inversion de matrice ?

- ☐ Minimiser une erreur par descente de gradient
- ☑ Résoudre des systèmes linéaires
- ☐ Effectuer des tests d'hypothèses
- ☐ Résoudre la chain rule

Exercice 2 : Régression Logistique

Nous avons un problème de classification. Pour le résoudre, nous allons utiliser une régression logistique. Notre modèle est donc le suivant $F : \mathbb{R}^3 \to \mathbb{R}^2$. Notre modèle prend des informations X et prédit un vecteur de probabilités Y. F fait le calcul suivant

$$F : X \mapsto F(X) = \text{softmax}\,(WX)$$

où W est une matrice et softmax : $(x, y) \mapsto \left(\frac{e^x}{e^x+e^y}, \frac{e^y}{e^x+e^y}\right)$.

Question 1 : Quelles sont les dimensions de W, X et $F(X)$?

La fonction F va de $\mathbb{R}^3$ dans $\mathbb{R}^2$, donc softmax$(WX) \in \mathbb{R}^2$ et $X \in \mathbb{R}^3$. De plus la fonction softmax va de $\mathbb{R}^2$ dans $\mathbb{R}^2$, on en déduit que $WX \in \mathbb{R}^2$. Ainsi, $F(X)$ a deux dimensions, X en a trois et W est une matrice de taille 2×3.

Étant donné un exemple d'entraînement $X = \begin{pmatrix}1 & 0 & 1\end{pmatrix}$ et une vérité terrain associée $Y = \begin{pmatrix}1 & 0\end{pmatrix}$, on va prédire $F(X)$ que l'on va comparer à Y pour mesurer l'erreur du modèle et ainsi pouvoir l'ajuster (modifier les paramètres W). Pour cela, on voit F comme la fonction suivante $F : W \mapsto \text{softmax}\,(WX)$ et X est une constante.

Question 2 : Calculer le gradient de la fonction $W \mapsto WX$

Ici, il faut faire attention ! La fonction (que l'on va noter g) qui nous intéresse va de $\mathbb{R}^{2\times 3}$ vers $\mathbb{R}^2$, nous avons donc 12 dérivées partielles à calculer.

$$\nabla g = \begin{pmatrix} \frac{\partial g_1}{\partial W_{1,1}} & \frac{\partial g_1}{\partial W_{1,2}} & \frac{\partial g_1}{\partial W_{1,3}} & \frac{\partial g_1}{\partial W_{2,1}} & \frac{\partial g_1}{\partial W_{2,2}} & \frac{\partial g_1}{\partial W_{2,3}} \\ \frac{\partial g_2}{\partial W_{1,1}} & \frac{\partial g_2}{\partial W_{1,2}} & \frac{\partial g_2}{\partial W_{1,3}} & \frac{\partial g_2}{\partial W_{2,1}} & \frac{\partial g_2}{\partial W_{2,2}} & \frac{\partial g_2}{\partial W_{2,3}} \end{pmatrix}^T = \begin{pmatrix} X_1 & X_2 & X_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & X_1 & X_2 & X_3 \end{pmatrix}^T$$

Question 3 : Calculer le gradient de la fonction $f : (x, y) \mapsto \text{softmax}(x, y) = \left(\frac{e^x}{e^x+e^y}, \frac{e^y}{e^x+e^y}\right)$

On a une fonction de $\mathbb{R}^2$ dans $\mathbb{R}^2$ et donc q4 dérivées partielles à calculer.

$$\nabla \text{softmax} = \begin{pmatrix} \frac{\partial \text{softmax}_1}{\partial x} & \frac{\partial \text{softmax}_1}{\partial y} \\ \frac{\partial \text{softmax}_2}{\partial x} & \frac{\partial \text{softmax}_2}{\partial y} \end{pmatrix} = \begin{pmatrix} \frac{e^{x+y}}{(e^x+e^y)^2} & -\frac{e^{x+y}}{(e^x+e^y)^2} \\ -\frac{e^{x+y}}{(e^x+e^y)^2} & \frac{e^{x+y}}{(e^x+e^y)^2} \end{pmatrix}$$

La fonction de coût $\mathcal{L}$ est la cross-entropy, de formule $\mathcal{L}(\vec{u}, \vec{v}) = -\sum_i u_i \log(v_i)$

Question 4 : Calculer le gradient de $\mathcal{L}$ par rapport à v

Ici, v est un vecteur de taille quelconque, par exemple n et la valeur $\mathcal{L}(\vec{u}, \vec{v})$ est un scalaire. Nous avons donc n dérivées partielles à calculer.

$$\nabla \mathcal{L} = \begin{pmatrix} \frac{\partial \mathcal{L}}{\partial v_1} \\ \vdots \\ \frac{\partial \mathcal{L}}{\partial v_n} \end{pmatrix} = \begin{pmatrix} -\frac{u_1}{v_1} \\ \vdots \\ -\frac{u_n}{v_n} \end{pmatrix}$$

Question 5 : Calculer le gradient de $\mathcal{L}(Y, F(W))$ par rapport à W (n'oubliez pas de remplacer par les valeurs de X et Y pour simplifier les calculs)

Cette valeur permet de faire une étape de descente de gradient. Pour optimiser F il faudra répéter le processus avec plusieurs exemples différents (différentes valeurs de X et Y).

On commence par remplacer dans nos gradients, X et Y par leurs valeurs respectives.

$$\nabla g(W) = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \end{pmatrix}^T$$

Attention, il ne faut pas se tromper entre les petits x et y du softmax et les grand X et grand Y. Il faut calculer WX,

$$WX = \begin{pmatrix} W_{1,1} + W_{1,3} \\ W_{2,1} + W_{2,3} \end{pmatrix}$$

puis on met à jour le softmax :

$$\nabla \text{softmax}(WX) = \begin{pmatrix} \frac{e^{W_{1,1}+W_{1,3}+W_{2,1}+W_{2,3}}}{(e^{W_{1,1}+W_{1,3}}+e^{W_{2,1}+W_{2,3}})^2} & -\frac{e^{W_{1,1}+W_{1,3}+W_{2,1}+W_{2,3}}}{(e^{W_{1,1}+W_{1,3}}+e^{W_{2,1}+W_{2,3}})^2} \\ -\frac{e^{W_{1,1}+W_{1,3}+W_{2,1}+W_{2,3}}}{(e^{W_{1,1}+W_{1,3}}+e^{W_{2,1}+W_{2,3}})^2} & \frac{e^{W_{1,1}+W_{1,3}+W_{2,1}+W_{2,3}}}{(e^{W}_{1,1}+W_{1,3}+e^{W_{2,1}+W_{2,3}})^2} \end{pmatrix}$$

Et enfin la loss

$$\nabla \mathcal{L}(F(W)) = \begin{pmatrix} -\frac{1}{F(W)_1} \\ 0 \end{pmatrix}$$

Nous allons devoir utiliser la chain-rule.

$$\nabla \mathcal{L}(Y, F(W)) = \mathcal{L}(F(W)) \times \nabla \text{softmax}(WX) \times \nabla g(W)$$

Ici les choses se simplifient assez bien entre la cross-entropy et le softmax. En effet, si on fait le produit matriciel entre les gradients ∇softmax et $\nabla \mathcal{L}$.

$$\nabla \text{softmax}(WX) \times \nabla \mathcal{L}(F(W)) = \begin{pmatrix} F(W)_1 - 1 \\ F(W)_2 \end{pmatrix}$$

Il ne reste plus qu'à faire la multiplication par $\nabla g(W)$ et ainsi

$$\nabla \mathcal{L}(Y, F(W)) = \begin{pmatrix} F(W)_1 - 1 & 0 & F(W)_1 - 1 & F(W)_2 & 0 & F(W)_2 \end{pmatrix}^T$$

Exercice 3 : Classification de Loutres

Dans cet exercice, nous allons étudier les performances du modèle précédent (il n'est pas nécessaire d'avoir fait l'exercice précédent mais d'en avoir lu l'énoncé). Nous sommes dans un problème de classifications à deux classes : **loutre de mer** ou **loutre de rivière**. Il y a un total de 1200 exemples et les classes sont équilibrées (même nombre de loutres de mer que de loutres de rivière). Notre modèle classifie correctement 1050 loutres. Nous prédisons un total de 600 loutres de mer.

Question 1 : Dresser le tableau de contingence du problème.

Pour rappel, voici la définition du tableau de contingence

$$\begin{matrix} VP & FP \\ FN & VN \end{matrix}$$

D'après l'énoncé, on a 1200 exemples et donc $VP+FP+VN+FN = 1200$. De plus, les classes sont équilibrées donc $VP + FN = VN + FP$. Ensuite notre modèle classifie correctement 1050 loutres, on en déduit $VP + VN = 1050$ et enfin, notre modèle prédit un total de 600 loutres de mer, ce que l'on traduit par $VP + FP = 600$. On a donc 4 équations pour 4 inconnues. On en déduit le tableau de contingence suivant

$$\begin{matrix} 525 & 75 \\ 75 & 525 \end{matrix}$$

Question 2 : Calculer les paramètres de l'évaluation.

$$\begin{cases} \text{sensibilité} & = \frac{VP}{VP+FN} = \frac{525}{600} = 0,875 \\ \text{spécificité} & = \frac{VN}{FP+VN} = \frac{525}{600} = 0,875 \\ \text{VPP} & = \frac{VP}{VP+FP} = \frac{525}{600} = 0,875 \\ \text{VPN} & = \frac{VN}{FN+VN} = \frac{525}{600} = 0,875 \end{cases}$$

Supposons maintenant que la probabilité de prédire correctement la classe loutre de mer sachant qu'on analyse une image de loutre de mer est de 95%. Par ailleurs, on ne change ni le dataset ni le nombre de prédictions "loutres de mer" du modèle.

Question 3. Calculer les paramètres de l'évaluation.

L'énoncé nous donne $\mathbb{P}(S|M) = 0.95$. Ceci nous donne directement la sensibilité. Les prévalences ne changent pas et donc $\mathbb{P}(M) = 0.5$ et $\mathbb{P}(S) = 0.5$. On utilise alors un peut de théorie des probabilités conditionnelles :

$$\mathbb{P}(S) = \mathbb{P}(S|M)\mathbb{P}(M) + \mathbb{P}(S|\bar{M})\mathbb{P}(\bar{M}) \Leftrightarrow \frac{1}{2} = \frac{0.95}{2} + \frac{\mathbb{P}(P|\bar{O})}{2} \Leftrightarrow \mathbb{P}(S|\bar{M}) = 0.05$$

Ainsi la spécificité vaut $1 - 0.05 = 0.95$. Pour la VPP et la VPN, on va passer par l'intersection des probabilités

$$\mathbb{P}(M|S) = \frac{\mathbb{P}(S|M)\mathbb{P}(M)}{\mathbb{P}(S)} = \mathbb{P}(S|M) = 0.95$$

et

$$\mathbb{P}(\bar{M}|\bar{S}) = \frac{\mathbb{P}(\bar{S}|\bar{M})\mathbb{P}(\bar{M})}{\mathbb{P}(\bar{S})} = \mathbb{P}(\bar{S}|\bar{M}) = 1 - \mathbb{P}(S|\bar{M}) = 0.95$$

En conservant ces nouvelles hypothèses,

Question 4. Donner le nombre de prédictions correctes estimé.

L'espérance du nombre de prédictions B correctes est donnée par le nombre d'exemples multiplié par la probabilité d'avoir une bonne prédiction. Formellement,

$$B = 1200 \times (\mathbb{P}(S \cap M) + \mathbb{P}(\bar{S} \cap \bar{M})) = 1200 \times (0,475 + 0,475) = 1140$$

En pratique on observe que le modèle trouve 1157 loutres correctement.

Question 5. faites un test d'hypothèse à % pour savoir si notre hypothèse (de la question 3) est vraisemblable.

On note X_i les variables aléatoires décrivant le phénomène aléatoire : la prédiction sur l'exemple i est correcte. Ainsi ces variables suivent des lois de Bernoulli de paramètres p égal à la probabilité de faire une prédiction correcte. Formellement, les X_i sont i.i.d. et $X_i \sim \beta(0.95)$. La question 4 nous dit

$$\mathbb{E}(X_1 + \cdots + X_{1200}) = 1140$$

Pour effectuer un test d'hypothèse il nous faut deux choses : 1) vérifier les hypothèses du TCL et 2) l'écart-type de notre variable aléatoire. Pour le premier point, on se rappelle que l'hypothèse à vérifier dans le cas de variables de Bernoulli est $pN > 20$ et $(1-p)N > 20$. En remplaçant, on trouve $pN = 1140 > 20$ et $(1-p)N = 60 > 20$. On pourra donc appliquer le TCL. Pour ce qui est de l'écart-type, on utilise la propriété sur la variance de variables aléatoires indépendantes :

$$\mathbb{V}(X_1 + \cdots + X_{1200}) = 1200\mathbb{V}(X_1) = 1200p(1-p) = 57$$

On en déduit un écart-type de 7.55. On applique alors le TCL et les dérivations vues en classe, en posant $S = X_1 + \cdots + X_{1200}$:

$$\begin{aligned}\mathbb{P}(S \in [1140 - a; 1140 + a]) &= \mathbb{P}(S - 1140 \in [-a; +a]) \\ &= \mathbb{P}\left(\frac{S - 1140}{7.55} \in \left[-\frac{a}{7.55}; \frac{a}{7.55}\right]\right) \\ &= \mathbb{P}\left(Z \in \left[-\frac{a}{7.55}; \frac{a}{7.55}\right]\right) \\ &= 0.98\end{aligned}$$

En utilisant la table, on trouve $\frac{a}{7.55} = 2.326$ et donc $a \approx 17.5$.

$$\mathbb{P}(S \in [1122,5; 1157,5 + a]) = 0.98$$

Et effectivement la valeur observée 1157 tombe dans cet intervalle. On ne peut donc pas rejeter notre hypothèse avec un test à 98%.
rappel : $\sqrt{57} \approx 7,55$ je dis ça...

Exercice 4 : Un peu de théorie...

On va faire un peu de théorie de la mesure (théorie à la racine de la théorie moderne des probabilités).
On appelle tribu $\mathcal{T}$ sur un espace E, tout ensemble de sous-ensembles de E vérifiant :

- $\mathcal{T} \neq \emptyset$
- $\forall T \in \mathcal{T}, \quad T^C \in \mathcal{T}$ (stabilité par passage au complémentaire)
- $\forall (T_i)_{i\in\mathbb{N}} \in \mathcal{T}, \quad \cup_{i\in\mathbb{N}} T_i \in \mathcal{T}$ (stabilité par union dénombrable).

Question 1. Montrer que la tribu grossière $\mathcal{T} = \{\emptyset, E\}$ est effectivement une tribu
La tribu grossière n'est pas vide puisqu'elle contient deux éléments. En suite pour chaque élément la tribu grossière contient effectivement son complémentaire puis que le complémentaire du vide est bien E. Enfin les unions d'éléments de $\mathcal{T}$ sont soit égales à E si E est dans la suite des T_i et vide sinon. Or E et le vide appartiennent bien à la tribu ce qui montre sa stabilité par union dénombrable.
Question 2. En supposant $E = \{1, 2, 3, 4\}$, dire si $\mathcal{T} = \{\{1\}, \{2, 3, 4\}, \{1, 2\}\{3, 4\}\}$ est une tribu sur E.
$\mathcal{T}$ n'est pas une tribu. En effet, $\mathcal{T}$ contient $\{1, 2\}$ et $\{3, 4\}$ mais pas leur union qui vaut E.
Question 3. Montrer que toute tribu contient nécessairement l'ensemble vide
Soit $\mathcal{T}$ une tribu. Comme $\mathcal{T}$ n'est pas vide, on considère T un élément dans $\mathcal{T}$. Comme $\mathcal{T}$ est stable par passage au complémentaire, T^C appartient également à $\mathcal{T}$. Comme $\mathcal{T}$ est stable par union dénombrable, $T \cup T^C = E$ appartient à $\mathcal{T}$. Enfin, on applique de nouveau le passage au complémentaire pour trouver $E^C = \emptyset \in \mathcal{T}$.
Question 4. Montrer que toute tribu est stable par intersection finie
Soient $\mathcal{T}$ une tribu et $T_1, \ldots, T_n$ une suite finie d'éléments de $\mathcal{T}$. On utilise les propriétés de $\mathcal{T}$ pour avoir

$$\cap_{i=1}^n T_i = \left(\cup_{i=1}^n T_i^C\right)^C \in \mathcal{T}$$

Une topologie sur un espace E est un ensemble de sous-ensembles de E contenant, l'ensemble vide, E et stable par union quelconque et intersection dénombrable.
Question 5. Donner un exemple de topologie sur $E = \{1, 2, 3, 4\}$ qui n'est pas une tribu

$$\{\emptyset, \{1\}, E\}$$