DM – Méthodologie en IA

Tutorat 2020-2021 : **20 QCMs**

Ok les gars. Je vous explique la situation : c'est un DM de Santé Num sur le cours Méthodologie en IA de David Chardin. ***TOUS LES QCMs ONT ETE RELUS ET MODIFIES/CORRIGES PAR LE PROF****. Je vous avais fait un beau petit DM avec les QCMs modifiés en QRUs, les items, énoncés et corrections qui n'allaient pas avaient été modifiés/enlevés/remplacés, etc… mais j'ai très malheureusement perdu mon fichier, et j'ai absolument pas le temps de reprendre toute la mise en page (que ce soit sur le fond ou la forme) de ce DM, par rapport à moi mais aussi et surtout par rapport à vous. Je suis désolée dans tous les cas, parce que j'avais déjà pris du retard sur la sortie de ce DM, mais encore plus désolée maintenant vu que ça n'aura pas été rectifié et mis en page… Je tenais quand même à vous sortir tout ça avec les commentaires du prof surtout, parce qu'il donne quelques indications du genre « ça je ne le demanderai pas, ça ne m'intéresse pas » ou quoi, donc c'est toujours bon à prendre ! Ne vous cassez pas la tête dessus du coup, si vous avez le temps je pense que ça peut être une bonne chose de se prendre 1 petite heure et de le lire tranquille et essayer de répondre (en faisant attention pcq ya la correction juste en-dessous des QCMs sorry) et lire les remarques du prof. Si vous êtes vraiment à la bourre et que vous avez autre chose à foutre de plus important, pas de panique, comme je dis toujours, vaut mieux faire des choix stratégiques. Breef, j'arrête de vous faire perdre votre temps. Je vous souhaite PLEIN DE COURAGE pour ces derniers jours, vous en êtes totalement capables donc croyez en vous.*
Et une mention spéciale à tous ces gentils P1 du MC qui sont beaucoup trop chous et qui m'ont conseillé pendant que j'étais en PLS pour mon fichier perdu, gros bisous à vous : doral'exploiteuse, JPdentiras, anisbenkanoun, Paulinepome et marie.plagnet

QRU 1 : A propos du cours sur la méthodologie en intelligence artificielle, indiquez la proposition exacte :

A) La fonction de coût permet de déterminer la meilleure approximation linéaire des données disponibles
B) La fonction de perte représente la différence entre ce qu'on a obtenu et ce qu'on souhaite obtenir comme résultat
C) On utilise la descente de gradient pour maximiser l'erreur
D) Lors de l'apprentissage d'un modèle, l'objectif est de trouver un taux d'apprentissage le plus élevé possible
E) Les propositions A, B, C et D sont fausses

QRU 2 : A propos du cours sur la méthodologie en intelligence artificielle, indiquez la proposition FAUSSE :

A) Si on crée une fonction qui permet de relier un label Y à une variable explicative X, et que cette fonction prend la forme d'une droite d'équation Y=a+bX, on aura réalisé une régression linéaire
B) Dans les problèmes de classification, on peut utiliser une fonction prenant une forme de « S » : on l'appelle fonction sigmoïde ou fonction logistique
C) Si les données x1 et x2 ont la même dimension, la descente de gradient va fonctionner correctement et l'algorithme va rapidement trouver un minimum local
D) Si les données x1 et x2 présentent des dimensions très différentes, on va devoir faire un scaling des données pour faire en sorte qu'elles aient une dimension similaire
E) Les propositions A, B, C et D sont fausses

QRU 3 : A propos des différents types d'apprentissage, indiquez la proposition exacte :

A) Dans le modèle d'apprentissage supervisé, on va vouloir créer des modèles qui permettent de mettre en évidence des tendances présentes dans des bases de données, sans savoir à l'avance ce que sont ces tendances
B) L'apprentissage non supervisé est le plus fréquent dans des applications en médecine
C) L'apprentissage supervisé peut être utilisé pour prédire des prix d'immobilier ou pour le triage de catégories de mails par exemple
D) L'apprentissage supervisé s'utilise pour regrouper des sites similaires sur le web, ou pour distinguer des sous-types de tumeurs à partir de données génétiques
E) Les propositions A, B, C et D sont fausses

QRU 4 : A propos de l'algorithme de descente de gradient, indiquez la proposition exacte :

A) Si on considère une fonction Y=a+bX, on va essayer de trouver les valeurs de Y et de X
B) Pour mesurer l'erreur engendrée en ayant pris des valeurs d'essai pour a et b, on utilise la fonction de coût, qui peut prendre tout un tas de forme, comme par exemple la moyenne de la différence absolue
C) Notre objectif est que la courbe se rapproche du résultat attendu, dans ce cas-là l'erreur augmente
D) On veut faire en sorte que l'erreur diminue jusqu'à ce qu'on atteigne un minimum, on pourra ainsi toujours trouver les meilleurs paramètres a et b possibles
E) Les propositions A, B, C et D sont fausses

QRU 5 : A propos de la descente de gradient, indiquez la proposition exacte :

A) La descente de gradient est un algorithme de base en Machine Learning, qu'on peut utiliser pour arriver à une erreur minimale
B) Si on part d'une fonction Y=a+bX, la descente de gradient démarre en prenant des paramètres arbitraires qu'elle met à jour jusqu'à atteindre la convergence
C) Si augmenter la valeur du paramètre permet de diminuer l'erreur, alors la dérivée de la fonction de coût sera positive
D) Si on arrive sur un minimum local, la dérivée de la fonction d'erreur sera nulle et le paramètre ne changera plus de valeur
E) Les propositions A, B, C et D sont fausses

QCM 6 : A propos de la descente de gradient, indiquez la (les) proposition(s) exacte(s) :

A) L'algorithme « descente de gradient » est utilisé seulement dans le cas de la régression linéaire
B) La descente de gradient ne sera pas applicable dans une situation où on a plusieurs minimums locaux.
C) Il existe toujours une solution unique aux problèmes qui utilisent la descente de gradient
D) Pour utiliser la descente de gradient comme algorithme d'apprentissage pour répondre à un problème donné, il faut savoir ajuster correctement le taux d'apprentissage
E) Les propositions A, B, C et D sont fausses

Correction QCM 6 : D

A) Faux : il peut être utilisé pour un grand nombre de fonctions, et pas seulement pour la régression linéaire
B) Faux : si, il existe des situations où on peut très bien avoir plusieurs minimums locaux possibles, et donc en fonction des paramètres qu'on utilise au départ, le résultat ne sera pas forcément le même
C) Faux : justement non, il n'existe pas toujours une solution unique aux problèmes (notamment dans la cas où on est en présence de plusieurs minimums locaux possibles)
D) Vrai
E) Faux

QCM 7 : A propos du taux d'apprentissage, indiquez la (les) proposition(s) exacte(s) :

A) Le taux d'apprentissage est un facteur qui multiplie la dérivée de la fonction de coût
B) Si on prend un taux d'apprentissage très petit, à chaque mise à jour on va avoir une grande variation des paramètres θ.
C) Si on prend un taux d'apprentissage trop élevé, on risque de dépasser voire même de s'éloigner du seuil minimal recherché
D) Le taux d'apprentissage n'est pas une valeur fixe, il varie à chaque mise à jour par l'algorithme
E) Les propositions A, B, C et D sont fausses

Correction QCM 7 : AC

A) Vrai
B) Faux : une toute petite variation de θ
C) Vrai
D) Faux : le taux d'apprentissage est une valeur fixe !
E) Faux

QCM 8 : A propos du taux d'apprentissage, indiquez la (les) proposition(s) exacte(s) :

A) Pour converger, il faut que les modifications apportées aux paramètres θ au fur et à mesure qu'on se rapproche du minimum soient de plus en plus fines, mais cela ne peut pas se faire sans avoir à modifier le taux d'apprentissage
B) Au fur et à mesure qu'on se rapproche du minimum, le taux d'apprentissage diminue petit à petit
C) Si on choisit un taux d'apprentissage trop petit, il va falloir un nombre important d'itérations afin d'obtenir le résultat
D) Alors que si on choisit un taux d'apprentissage trop élevé, on risque de ne jamais converger
E) Les propositions A, B, C et D sont fausses

Correction QCM 8 : CD

A) Faux : le taux d'apprentissage est une valeur fixe. Cela peut donc se faire sans avoir à le modifier
B) Faux : c'est le paramètre de la dérivée de la fonction de coût, et donc la pente de la courbe qui diminue +++
C) Vrai
D) Vrai : L'objectif est donc de trouver un taux d'apprentissage qui permet la convergence, sans utiliser trop de calculs +++
E) Faux

QCM 9 : A propos du scaling des données, indiquez la (les) proposition(s) exacte(s) :

A) Le scaling des données (ou mise à l'échelle) concerne les données de sortie qu'on va utiliser (c'est-à-dire Y)
B) On utilise le scaling des données lorsque nos données présentent des dimensions trop différentes
C) Si les données ont des dimensions similaires, la fonction de perte sur une vue 2D présente une forme de cible
D) Si les données ont des dimensions différentes, la pente menant au minimum sera beaucoup plus faible, et l'algorithme va mettre beaucoup de temps pour trouver le minimum
E) Les propositions A, B, C et D sont fausses

Commenté [DC1]: En réalité on peut aussi faire un scaling sur Y…

Commenté [DC2]: C'est une illustration du cours, mais la fonction de perte va être continue, c'est un peu bancal comme item

Commenté [DC3]: Ca peut être vrai, mais ça dépend des paramètres initiaux (la pente ne sera pas faible partout)

Correction QCM 9 : BCD

A) Faux : il concerne les données d'**entrée** (donc **X**)
B) Vrai
C) Vrai
D) Vrai
E) Faux

QCM 10 : A propos du scaling des données, indiquez la (les) proposition(s) exacte(s) :

A) Le centrage des données a pour but que les valeurs des données se répartissent sur la même fourchette, et donc qu'elles aient la même dimension
B) La réduction des données a pour but que la moyenne des deux données soit égales
C) Il existe différentes mesures de la dispersion des données : variance, écart-type, range
D) Le scaling permet de diminuer la dispersion des données, et donc d'augmenter leur dimension afin de faciliter la réalisation de la descente de gradient
E) Les propositions A, B, C et D sont fausses

Correction QCM 10 : C

A) Faux : ça c'est la **réduction** des données
B) Faux : ça c'est le **centrage** des données
C) Vrai
D) Faux : on **diminue** la dimension
E) Faux

QCM 11 : A propos du cours sur la méthodologie en intelligence artificielle, indiquez la (les) proposition(s) exacte(s) :

A) Si on prend un exemple de prédiction du prix des maisons en fonction de leur surface en mètres carrés, on se trouve dans un modèle de régression à 1 seule variable
B) En Machine Learning, si notre modèle de régression est imparfait, on va vouloir l'affiner en supprimant les variables explicatives en trop
C) Lorsqu'on prend en compte des variables supplémentaires, on se place dans un modèle de régression linéaire multiple
D) L'objectif est d'arriver à un modèle complexe qui passe par l'ensembles des points des données existantes
E) Les propositions A, B, C et D sont fausses

Commenté [DC4]: On n'est pas obligé de faire de la régression linéaire

Correction QCM 11 : A

A) Vrai
B) Faux : affiner le modèle se fait en **ajoutant** de nouvelles variables explicatives
C) Faux : on ne parle plus de régression linéaire, mais de **régression polynomiale**
D) Faux : un modèle trop complexe comme ça peut être délétère, il faut arriver à un bon compromis
E) Faux

Commenté [DC5]: On peut très bien faire de la régression linéaire multivariée, il faut changer la proposition C.

QCM 12 : A propos de l'underfitting et de l'overfitting, indiquez la (les) proposition(s) exacte(s) :

A) Si le modèle est trop simple pour représenter la distribution des données, on dit qu'il y a underfitting
B) Si le modèle est trop complexe, avec une erreur quasi nulle, il risque d'être inutilisable pour de nouvelles données, et on parle alors d'overfitting
C) Dans les problèmes de régression et de classification on peut se retrouver avec un modèle trop simple ou trop complexe, et il va falloir trouver un bon intermédiaire qui sépare suffisamment les données d'entraînement tout en étant applicable pour de nouvelles données
D) Pour pallier à de l'underfitting, une solution est de prendre en compte un plus grand nombre de variables explicatives
E) Les propositions A, B, C et D sont fausses

Correction QCM 12 : ABCD

A) Vrai
B) Vrai
C) Vrai : item long, mais totalement vrai ++
D) Vrai
E) Faux

QCM 13 : A propos de l'underfitting et de l'overfitting, indiquez la (les) proposition(s) exacte(s) :

A) Rajouter trop de variables explicatives au modèle rajoute un risque d'underfitting
B) Dans des applications où l'on ne connaît pas à l'avance les données qui vont être pertinentes, on utilise des méthodes qui réduisent le nombre de données prises en compte en essayant d'éliminer les variables qui comportent une information redondante
C) On peut également utiliser la méthode de régularisation qui permet de limiter la magnitude des valeurs que peuvent prendre les paramètres θ.
D) L'underfitting et l'overfitting peuvent limiter la qualité des modèles, donc il est important de travailler avec des données de bonne qualité
E) Les propositions A, B, C et D sont fausses

Commenté [DC6]: Le problème n'est pas tant l'underfitting et l'overfitting ici, mais plutôt la qualité de la mesure des variables explicatives.

Correction QCM 13 : BCD

A) Faux : risque d'overfitting
B) Vrai
C) Vrai
D) Vrai
E) Faux

QCM 14 : A propos de l'échantillonnage en apprentissage supervisé, indiquez la (les) proposition(s) exacte(s) :

A) En Machine Learning, l'objectif est de créer un modèle à partir d'un échantillon d'une population, afin d'utiliser ce modèle pour l'appliquer à l'ensemble de la population
B) A partir de la population, on extrait une cohorte qui va servir pour entraîner le modèle et l'évaluer
C) Il est absolument nécessaire que les cohortes soient représentatives de la population générale
D) Si on prend une cohorte non représentative pour entraîner le modèle, on est dans le cas d'un underfitting
E) Les propositions A, B, C et D sont fausses

Commenté [DC7]: Un peu ambiguë car on va effectivement entrainer et évaluer sur la cohorte d'entrainement, mais ca ne suffira pas, il faudra aussi testé sur une 2nd cohorte.

Correction QCM 14 : AC

A) Vrai
B) Faux : on extrait 2 cohortes différentes : la cohorte d'entraînement pour entraîner le modèle, puis une deuxième cohorte appelée cohorte de test pour évaluer le modèle +++++
C) Vrai +++
D) Faux : on sera plutôt dans le cas d'un overfitting ++
E) Faux

Commenté [DC8]: C'est un peu plus compliqué que ça, même si en partie vrai…

QCM 15 : A propos de la cross validation, indiquez la (les) proposition(s) exacte(s) :

A) Si on utilise la même cohorte pour faire l'entraînement et l'évaluation du modèle, on risque de surestimer les performances du modèle
B) Pour être au must, un modèle doit avoir été entraîné et testé sur un nombre suffisant de patients
C) En utilisant la méthode de cross validation, on aura évalué le modèle sur la totalité de la population
D) Avec cette méthode, il ne sera plus nécessaire d'évaluer le modèle sur une 2ème population
E) Les propositions A, B, C et D sont fausses

Correction QCM 15 : ABC

A) Vrai : c'est la raison pour laquelle on va utiliser la méthode de cross validation
B) Vrai
C) Vrai
D) Faux : il faudra tout de même évaluer le modèle sur une 2ème population (la population test), car cette dernière est indépendante de la population d'entraînement, et cela nous permettra d'affirmer que le modèle est bel et bien reproductible
E) Faux

QCM 16 : A propos de l'évaluation des performances d'un modèle, indiquez la (les) proposition(s) exacte(s) :

A) On va utiliser un tableau de contingence pour l'évaluation des performances d'un modèle de régression linéaire
B) La sensibilité correspond à la capacité du modèle à prédire un résultat positif lorsqu'il doit prédire ce résultat positif
C) La valeur prédictive négative correspond à la probabilité qu'un résultat soit effectivement négatif lorsque le modèle l'a prédit en tant que tel
D) Lorsque le label vrai est positif mais a été prédit en négatif, on aura un faux négatif
E) Les propositions A, B, C et D sont fausses

Correction QCM 16 : BCD

A) Faux : dans le modèle de classification
B) Vrai :
C) Vrai
D) Vrai
E) Faux

QCM 17 : A propos de l'évaluation des performances d'un modèle, indiquez la (les) proposition(s) exacte(s) :

A) L'accuracy correspond au pourcentage de fois où le modèle a correctement fait la prédiction
B) La précision est une métrique qui correspond au nombre de patients correctement prédits sur l'ensemble de la cohorte évaluée
C) L'accuracy permet de représenter l'ensemble du modèle, et pas forcément que 2 classes
D) On utilise le coefficient de détermination pour prédire l'erreur réalisée par un modèle de régression, qui correspond à la distance entre le label prédit et le label vrai
E) Les propositions A, B, C et D sont fausses

Correction QCM 17 : ABCD

A) Vrai
B) Vrai : accuracy = précision
C) Vrai
D) Faux : Le coefficient de détermination c'est le carré du coefficient de corrélation de Pearson
E) Faux

QCM 18 : A propos de l'évaluation des modèles en apprentissage supervisé, indiquez la (les) proposition(s) exacte(s) :

A) On effectue 2 évaluations : la première sur la cohorte test, et la 2ème sur la cohorte d'entraînement
B) La 1ère évaluation donne une indication sur le fit du modèle
C) La 2ème évaluation permet d'évaluer la reproductibilité du modèle pour le reste de la population
D) On utilise la méthode de cross-validation pour évaluer le modèle sur une population test, indépendante de la population d'entraînement
E) Les propositions A, B, C et D sont fausses

Correction QCM 18 : BC

A) Faux : c'est l'inverse : d'abord sur la cohorte d'entraînement, puis la cohorte test
B) Vrai
C) Vrai
D) Faux : on utilise la méthode de cross-validation pour la 1ère évaluation sur la cohorte d'entraînement
E) Faux

QCM 19 : A propos de de la méthodologie en Machine Learning, indiquez la (les) proposition(s) exacte(s) :

A) Pour appliquer les méthodes de Machine Learning, il est nécessaire que le problème soit bien défini et que les données soient bien sélectionnées et de bonne qualité
B) A partir de la définition du problème, on pourra alors choisir de quel type d'apprentissage on va avoir besoin, et quelle méthode (classification ou régression) utilisée
C) Après analyse des données d'entrée, un scaling peut être nécessaire si celles-ci ont des dimensions trop similaires
D) Afin de bien sélectionner les données, il est important d'avoir réalisé un bon échantillonnage, et d'avoir bien défini le label X à prédire et les variables explicatives Y
E) Les propositions A, B, C et D sont fausses

Commenté [DC9]: X et Y sont des conventions, mais intervertir les noms n'empêche pas de faire du machine Learning… Je ne ferai pas d'item comme ça

Correction QCM 19 : AB

A) Vrai
B) Vrai
C) Faux : si elles ont des dimensions trop différentes
D) Faux : X et Y ont été inversé dans l'item
E) Faux

QCM 20 : A propos de l'évaluation des modèles en apprentissage supervisé, indiquez la (les) proposition(s) exacte(s) :

A) Lors de l'évaluation de la reproductibilité du modèle, si les performances sont très élevées, l'accuracy est proche de 100%, donc le modèle est très bon, mais il peut y avoir un risque d'overfitting
B) Après évaluation en utilisant la méthode de cross-validation, si les performances sont trop basses, le modèle ne sera pas reproductible
C) A l'issue de l'évaluation sur la cohorte d'entraînement, si les performances sont au moins équivalentes à celles mesurées sur la cohorte test, les résultats seront reproductibles
D) Si les performances du modèle sont moins bonnes à l'issue de la 2ème évaluation, cela peut signifier que le modèle est underfitté ou qu'il y a un problème d'échantillonnage
E) Les propositions A, B, C et D sont fausses

Correction QCM 20 : E

A) Faux : On est pas sur de l'évaluation de la reproductibilité du modèle, mais sur l'évaluation du fit du modèle (donc 1ère évaluation, sur la cohorte d'entraînement)
B) Faux : Après évaluation en utilisant la méthode de cross-validation, si les performances sont trop basses, le modèle est **underfitté**
C) Faux : j'ai inversé cohorte d'entraînement et cohorte test dans l'item
D) Faux : **overfitté**
E) Vrai

Commenté [DC10]: Enfin il sera aussi probablement pas très reproductible

CORRECTION des 5 QRUs du début

QRU 1 : B

A) Faux : la fonction de coût (ou fonction de perte) ne permet que de mesurer l'erreur, « approximation linéaire » ne veut rien dire on peut mesurer l'erreur de n'importe quelle approximation / modèle (*modifié par le prof, mais du coup l'item devient un peu ambigu et pas vraiment très intéressant…*)
B) Vrai
C) Faux : pour minimiser l'erreur
D) Faux : l'objectif est de trouver un taux d'apprentissage qui permet la convergence sans utiliser trop de calculs. Si on prend un taux d'apprentissage trop élevé, on risque de ne jamais converger, voire de diverger
E) Faux

QRU 2 : E *(désolée pour la formulation de ce QRU, mais c'était parce que de base j'avais mis ABCD vraies car c'était sous forme de QCM initialement..)*

A) Faux : **Il faut bien comprendre qu'on a décidé que cette fonction prendrait la forme d'une droite** (*correction du prof*)
B) Faux : **Il existe plusieurs façons de résoudre les problèmes de classification, l'utilisation de la fonction logistique n'est qu'une de ces façons (on fait alors une « régression logistique » (« *terme que j'ai évité car il n'est pas facile de comprendre que la régression logistique permet de faire de la classification…* »)** (*correction du prof*)
C) Faux
D) Faux
E) Vrai : Les propositions A, B, C et D sont bien VRAIES, donc la proposition E est la proposition FAUSSE (*voir énoncé*)

QCM 3 : C *(Le prof m'a indiqué qu'il n'interrogerait pas sur l'apprentissage SEMI-supervisé, car il ne l'a pas développé en cours ++)*

A) Faux : c'est dans l'apprentissage **non** supervisé
B) Faux : l'apprentissage supervisé, lorsque l'on travaille sur des bases de données où le label pour les patients est bien connu
C) Vrai

D) Faux : l'apprentissage **non** supervisé
E) Faux

QRU 4 : B

A) Faux : on veut trouver a et b +++
B) Vrai +++
C) Faux : dans ce cas-là, l'erreur diminue
D) Faux : **s'il existe plusieurs minimums locaux de l'erreur, on ne trouvera pas forcément les meilleurs paramètres a et b** (*correction du prof*)
E) Faux

QRU 5 : ABD

A) Vrai : Les paramètres a et b sont utilisés pour notre exemple Y=a+bX, mais on peut, comme expliqué dans le cours utilisé considérer des fonctions beaucoup plus complexes avec plus de 2 paramètres (**correction du prof**)
B) Vrai
C) Faux : elle sera négative, et appliquer la formule reviendra à augmenter un petit peu la valeur du paramètre
D) Vrai ++
E) Faux