MathAData: l'enseignement des mathématiques en lien avec l'expérimentation sur des challenges d'IA

Stéphane Mallat
Chaire de Sciences des Données
Collège de France

1 Introduction

De nombreuses études montrent que le va-et-vient entre des phases d'expérimentations, d'abstraction et de traitements mathématiques joue un rôle fondamental pour améliorer la compréhension des outils mathématiques [3]. Ce constat n'est plus guère remis en question et apparaît dans l'une des premières propositions du rapport Torrossian-Vilani [1] sur l'enseignement des mathématiques. Cette approche est menée avec succès dans le primaire par certains pays, notamment inspirés par la "méthode de Singapour." Elle permet de faire le lien entre les outils mathématiques que les élèves apprennent (chiffres, addition, multiplications) et leur intuition dans des situations pratiques où ces mathématiques s'appliquent. Cela approfondit considérablement la compréhension des concepts abstraits de mathématiques.

Dans le primaire, l'expérimentation peut se faire par des manipulations d'objets ou avec des jeux. Bien que cette démarche pédagogique reste tout aussi valable au lycée, l'expérimentation devient plus complexe. Elle doit être en lien avec les mathématiques de plus haut niveau qui sont au programme. Il faut aussi que les problèmes posés soient motivants pour les adolescents, et puissent les aider à se projeter un avenir en lien avec les mathématiques. C'est un élément important pour réduire la désaffection des mathématiques entre la seconde et la terminale, que l'on observe aujourd'hui. Il s'agit de montrer que les mathématiques ne sont pas en retrait du monde, et jouent un rôle important pour faire face aux grands enjeux sociétaux, depuis la santé jusqu'au réchauffement climatique.

Les allers-retours entre expérimentation, abstraction et mathématiques sont aussi au cœur de la recherche mathématique [2]. Les questions posées par les sciences et notamment la physique ont toujours été un moteur fondamental pour le développement de nouvelles mathématiques, que l'on retrouve dans les travaux de Newton, Gauss, Fourier, Poincaré... La beauté des mathématiques apparaît alors dans la simplicité et la puissance des concepts abstraits qui permettent de résoudre des problèmes pratiques en apparence très différents. Il n'est cependant pas facile d'enseigner les mathématiques au lycée avec une démarche

d'expérimentation en physique, chimie ou biologie. Cela demande aux élèves de suffisamment bien comprendre ces autres sciences, et le temps consacré à l'expérimentation scientifique réduit celui consacré aux mathématiques. Cet article explique pourquoi les challenges de données permettent d'éviter ces difficultés, tout en posant des questions sur des sujets très variés. Par là même, ils ouvrent une opportunité pour l'enseignement des mathématiques au lycée.

La difficulté n'est pas seulement du côté des élèves. En France, beaucoup de professeurs ont été formés avec une approche relativement formelle des mathématiques. On peut en effet introduire les mathématiques par une axiomatique, qui prend sa source dans la théorie des ensembles. Cependant cette approche s'est révélée être un chemin très difficile pour l'enseignement, qui rend les mathématiques inaccessibles pour de nombreux élèves. Même si l'éducation nationale a fait machine arrière depuis longtemps, cette culture et le prestige du formalisme restent présents dans l'enseignement à l'université. Le lien des mathématiques avec le monde s'enseigne alors à travers quelques "applications" en fin de cours, s'il reste du temps.

Enseigner les mathématiques par un aller-retour avec des questions issues d'expériences, avec des phases de modélisation et de traitements mathématiques, est un changement profond. Travailler sur des problèmes ouverts, qui admettent des solutions multiples, est aussi important pour que les élèves puissent exprimer leur créativité. Cela évite qu'ils soient bloqués ou obnubilés par la recherche de "la réponse", en s'engageant dans un processus moins stressant par essais et erreurs.

On voit ici que de nombreux obstacles s'accumulent pour développer un tel enseignement des mathématiques, dont l'enjeu est pourtant fondamental. Les contraintes peuvent être ainsi résumées :

1. Des problèmes ouverts, dont les enjeux sont importants et adaptés aux intérêts des élèves, mais qui puissent s'expliquer rapidement en termes mathématiques.

2. Des problèmes qui mettent en jeu une grande partie des mathématiques au programme, avec une phase de modélisation, et qui ouvrent une perspective sur l'importance de mathématiques plus sophistiquées pour résoudre des questions importantes.

3. Des expérimentations qui s'inscrivent dans un aller-retour équilibré avec la démarche d'abstraction mathématique et de traitements mathématiques.

4. Pour passer à l'échelle nationale, il faut s'adapter à la diversité des niveaux des élèves et former les professeurs.

L'objectif de cet article est de montrer que les challenges de données ouvrent de nouvelles perspectives pour s'attaquer de front à tous les aspects de ce problème, avec des expérimentations numériques.

Co-développement MathAData et Éducation Nationale MathAData regroupe une équipe au Collège de France et à l'École Normale Supérieure, en partenariat avec l'Institut Louis Bachelier, qui organise des challenges de données en intelligence artificielle sur la plateforme web *challengedata.ens.fr*. L'ambition de cette équipe est de proposer une approche pragmatique pour développer un enseignement des mathématiques par l'expérimentation, avec des challenges de données, au lycée et à l'université. Pour adapter une telle approche au lycée, MathAData fait un important travail pédagogique et didactique avec les professeurs dans les classes. Ce travail est mené en co-développement avec les professeurs des labos-maths de l'Académie de Lille [5].

L'intelligence artificielle est déjà présente au lycée, où de plus en plus d'élèves utilisent ChatGPT pour faire leurs devoirs à la maison. L'étude de ces challenges permet aussi de démystifier l'intelligence artificielle en montrant ses racines mathématiques, comme l'explique la section 2. La section 3 montre que la résolution de challenges couvre l'essentiel des mathématiques enseignées au lycée: statistiques, probabilités, algèbre, analyse et géométrie. Cela permet d'aborder ces mathématiques par des expérimentations numériques. La section 4 aborde les difficultés d'un tel enseignement dans les classes, et les solutions pédagogiques proposées.

2 Les challenges de données en intelligence artificielle

L'objectif de cette première section est de montrer que les challenges de données sont des problèmes ouverts, avec des enjeux adaptés aux intérêts des élèves, et qui peuvent s'expliquer rapidement en termes mathématiques.

Challenges de données Un challenge d'intelligence artificielle consiste à développer un algorithme qui peut apprendre à répondre à une question posée sur des données, à partir d'exemples d'entraînement qui sont mis à disposition. La formalisation mathématique est identique pour tous les challenges. La description ci-dessous est faite pour des professeurs de mathématiques, et non pas pour leurs élèves.

On note d la donnée, par exemple une image, qui est spécifiée par une liste de nombres. On note r la valeur de la réponse, par exemple 0 si c'est une image de chat et 1 si c'est une image de chien. À partir de la donnée d, le but est de calculer une estimation $\hat{r}$ de la réponse r, avec le moins d'erreur possible. Les problèmes de classification, régression et génération correspondent à des réponses r de types différents. En classification on doit identifier la catégorie de chaque donnée d, par exemple la classe des images de chats versus celle des images de chiens, et r est l'index de la classe que l'on code par un entier. Une régression estime un nombre réel r, comme la taille d'une personne qui apparaît dans une image. Un problème de génération estime une nouvelle donnée r qui peut être de grande dimension, par exemple une nouvelle image, ou un texte r généré à partir d'un texte d qui pose une question. Bien que des challenges de classification, régression et génération soient

en apparence très différents, ils se résolvent dans le même cadre mathématique introduit dans la section suivante.

Dans tous les challenges, la donnée d est une liste de d nombres réels $d = (d_1, d_2...)$. Si d est une image noir et blanc, chaque d_i spécifie l'intensité lumineuse d'un point (pixel) de l'image, codée par un entier entre 0 (pixel noir) et 255 (pixel blanc). La dimension de d est grande, typiquement de cent à plusieurs millions de variables pour une image.

La Figure 1 montre deux challenges de classification. Pour le challenge appelé MNIST, la donnée d est une petite image sans couleur d'un chiffre manuscrit, avec 28×28 pixels. Le challenge est d'identifier la valeur $r \in \{0, ..., 9\}$ du chiffre qui apparaît dans l'image. Pour le challenge appelé CIFAR, la donnée est une petite image couleur de 32×32 pixels, qu'il faut répartir en 10 classes : chats, chevaux, grenouilles, voitures, bateaux... qui sont aussi indexées par un entier $r \in \{0, ..., 9\}$. Ce challenge est nettement plus difficile que la reconnaissance de chiffres de MNIST, mais il reste accessible aux élèves de lycées, en utilisant les mêmes algorithmes d'apprentissage.

Figure 1: Un challenge de classification demande d'estimer l'index r de la classe d'une donnée d. Haut: exemples d'images du challenge MNIST. Différentes classes correspondant à différents chiffres manuscrits. Bas: exemples d'images du challenge CIFAR. Les classes d'images correspondent à des avions, voitures, oiseaux, chats...

Plateforme web La plateforme *challengedata.ens.fr* met à disposition plus de 80 challenges, de types et de niveaux de difficultés différents, pour l'enseignement au lycée, à l'université et pour la recherche. Cela peut être un problème de reconnaissance d'images, de diagnostic médical, de prédiction climatique, de calcul de pollution, d'analyse de textes... Les données sont alors des images, des séries temporelles, des textes, des sons... Les challenges proposés sont utilisés pour l'enseignement universitaire depuis la licence jusqu'au doctorat, dans tous les domaines des sciences et sciences humaines. Les plus faciles sont adaptés à l'enseignement au lycée.

Pour chaque challenge, la plateforme web fournit des exemples d'entraînement avec les données d et les réponses r. Par exemple, la classe de l'animal qui apparaît dans l'image. Nous verrons que les algorithmes de l'algorithme sont optimisés sur ces exemples d'entraînement. Comme un professeur, la plateforme propose une procédure de test sur de nouvelles données d qui sont différentes des exemples d'entraînement, et pour lesquels on ne fournit pas la réponse r. Un participant soumet à la plateforme les estimations $\hat{r}$ calculées par son algorithme sur ces exemples de tests. La plateforme, qui connaît les bonnes réponses, renvoie un score (une note), qui est reliée à l'erreur moyenne sur les exemples de test. Cette note est affichée sur un "leader board" pour comparer l'efficacité des différents algorithmes d'apprentissage. Cet aspect donne une composante ludique à la recherche de solutions, qui peut être pratiquée en groupe. Une phase de collaboration de toute la classe permet souvent d'obtenir une meilleure solution que les solutions trouvées par chacun des groupes.

Un challenge s'explique rapidement. Il suffit d'introduire la donnée d et la réponse r. On n'explique pas le lien entre la réponse r et les données d, comme on le ferait en physique. Ce sera aux élèves de le découvrir lors de leur modélisation, avec des algorithmes d'apprentissage que l'on va maintenant expliquer.

3 Les mathématiques de l'intelligence artificielle

On pourrait penser que la résolution d'un challenge de données est essentiellement un problème de statistiques. Ce n'est pas du tout le cas. Nous allons montrer que la résolution de ces challenges met en jeu une grande partie des mathématiques au programme du lycée. Cet exposé rapide introduit à la fois les principes de l'intelligence artificielle et les domaines mathématiques mis en jeu. La présentation pédagogique pour les élèves de lycée simplifie et détaille cette introduction, en l'intégrant dans les présentations des chapitres mathématiques du programme.

Le but est de trouver un algorithme qui calcule la réponse r à partir de la donnée d. Par exemple, r peut être l'index de la classe de l'image d. Pour cela, on définit un algorithme qui calcule une estimation $\hat{r}$ de r, en utilisant une liste de paramètres $x = (x_1, x_2, ...)$ qu'il faudra apprendre.

Une des idées importantes est que les données sont représentées dans un espace euclidien

dont la distance est une mesure de similarité entre les données. Comme l'explique la section 3.1, cela devient un problème de géométrie. Au lycée, cet espace est une droite, un plan ou l'espace tridimensionnel. Calculer la réponse associée à chaque donnée revient à séparer cet espace euclidien en différentes zones. Cette séparation est apprise en fixant les paramètres x de la frontière entre ces zones. La section 3.2 explique que ces paramètres sont optimisés sur des exemples d'entraînement, afin que l'estimation $\hat{r}$ de la réponse r ait le moins d'erreurs possible. Cela ressemble à l'apprentissage d'un élève qui s'entraîne sur des exercices avec les réponses fournies par son professeur, afin qu'il apprenne de ses erreurs.

L'objectif est que l'algorithme fasse peu d'erreurs sur des données différentes de celles de l'entraînement mais du même type. On dit alors que l'algorithme "généralise". La section 3.3 montre que cela s'évalue en calculant l'erreur faite par l'algorithme sur des données de test, différentes des données d'entraînement. C'est la même approche que celle d'un enseignant qui évalue l'apprentissage de ses élèves en donnant un examen, avec des exercices similaires mais différents de ceux étudiés en classe.

3.1 Estimation paramétrique et géométrie euclidienne

Caractéristiques Pour calculer la réponse r associée à une donnée d, on associe à chaque donnée d une liste de caractéristiques $k = (k_1, k_2, ...)$. Si on prend 2 caractéristiques alors k est considéré comme un point de coordonnées (k_1, k_2) dans un plan. Plus généralement, k est un point d'un espace euclidien dont la dimension est égale au nombre de caractéristiques, qui sont les coordonnées de k.

On va ajuster le calcul de k pour que des données d ayant la même réponse r aient des caractéristiques k qui soient proches. Au contraire, les caractéristiques doivent être éloignées lorsque les réponses sont différentes. Il faut donc choisir des caractéristiques qui discriminent les données correspondant à des réponses différentes. C'est la difficulté principale de chaque challenge.

À titre d'illustration, on simplifie le challenge MNIST en classifiant seulement deux nombres manuscrits: les images de 2 et de 7. On peut définir une caractéristique k_1 qui est l'intensité moyenne des pixels $(d_1, d_2, ...)$ de l'image. Les images de 2 ont typiquement plus de pixels clairs (grandes valeurs) que les images de 7. Elles ont donc une valeur moyenne $k = k_1$ qui est souvent plus grande, ce qui permet de les discriminer. On peut cependant trouver une caractéristique plus discriminante, par exemple avec une moyenne localisée dans certaines parties de l'image. De nombreuses idées peuvent être essayées, ce qui permet aux élèves de trouver des solutions créatives.

On peut aussi raffiner le résultat en prenant deux plutôt qu'une caractéristique. Par exemple, k_1 peut être la moyenne sur la moitié supérieure de l'image et k_2 la moyenne de l'image sur la moitié inférieure. La Figure 2 montre l'ensemble de ces caractéristiques pour des images de 2 (en bleu) et des images de 7 (en orange) dans une base de données d'entraînement. On voit que ces deux nuages de points sont assez bien séparés, mais pas parfaitement. On peut selectionner une troisième caractéristique k_3 pour mieux séparer

ces images, auquel cas chaque caractéristique est représentée par un point dans un espace tridimensionnel. Plus on a de caractéristiques mieux on peut séparer les données. La recherche des caractéristiques fait partie du travail créatif des élèves, qui leur permet d'effectuer une première modélisation du problème.

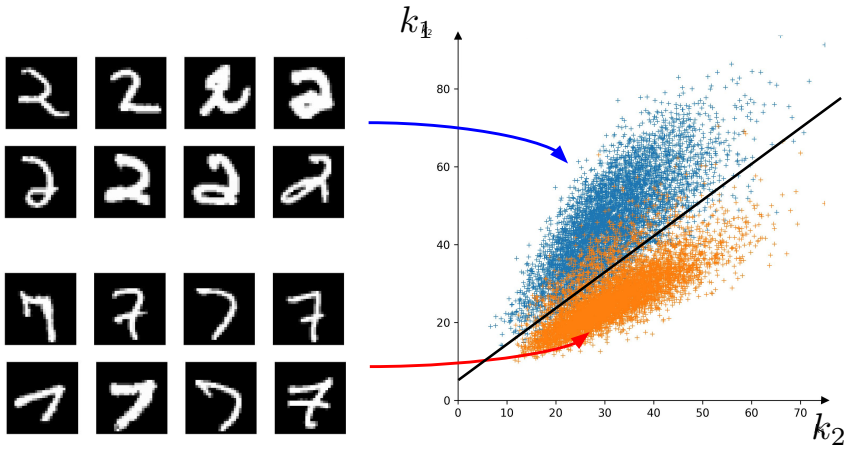

Figure 2: À chaque image on associe ici deux caractéristiques (k_1, k_2) qui sont les coordonnées d'un point dans un plan. Les caractéristiques des images de 2 sont des points bleus et celles des images de 7 des points oranges. Un classificateur linéaire sépare le nuage des points bleus et celui des points oranges avec une droite. Les erreurs correspondent aux points qui sont du mauvais côté de la droite.

Séparation géométrique Construire des classificateurs dans l'espace euclidien des caractéristiques met en jeu l'essentiel du programme de géométrie de la seconde, première et terminale, dans le plan et dans l'espace. Les caractéristiques ont été choisies pour discriminer les données d qui correspondent à des classes r différentes. On peut donc séparer l'espace des caractéristiques en zones où les données appartiennent le plus souvent à une même classe. On va séparer ces zones avec une frontière linéaire qui est paramétrée par x.

Supposons que l'on ait que deux classes différentes, par exemple des images de 2 et de 7. Il faut alors séparer l'espace des caractéristiques en deux zones. Si on a une seule caractéristique, on compare $k = k_1$ à la valeur d'un seuil x pour définir ces deux zones. Si $k = (k_1, k_2)$ est dans un plan, alors on divise ce plan en deux parties avec une droite séparatrice. L'équation d'une droite non-horizontale peut s'écrire $k_1 + ak_2 = b$. Elle est donc paramétrée par $x = (a, b)$. Si on utilise trois caractéristiques $k = (k_1, k_2, k_3)$ qui définissent un point dans l'espace alors la séparation se fait avec un plan d'équation $k_1 + ak_2 + ck_3 = b$. Il est paramétré par $x = (a, b, c)$. Les paramètres de x modifient la position et l'orientation de la droite ou du plan séparateur. Nous verrons qu'ils sont optimisés lors de l'apprentissage, afin de minimiser l'erreur de classification calculée sur les exemples d'entraînement.

Ces algorithmes sont des classificateurs linéaires. Ils séparent l'espace des caractéristiques

k en deux parties avec un plan ou une droite, dont l'équation peut se récrire avec un produit scalaire :

$$w.k - b = 0 \ .$$

Le vecteur w est orthogonal à la droite ou au plan. Un point est d'un côté ou de l'autre du plan ou de la droite suivant le signe de $w.k - b$, qui définit donc l'estimation $\hat{r}$ de la classe.

Réseaux de neurones Trouver des caractéristiques qui séparent linéairement toutes les classes est l'aspect le plus difficile d'un challenge de données. Pour des problèmes complexes, comme le challenge CIFAR de la figure 1, il est très difficile de trouver "à la main" des caractéristiques qui ne produisent pas beaucoup d'erreurs. Un réseau de neurones peut faire mieux en optimisant aussi le choix de la liste de caractéristiques k. Un algorithme de réseau de neurones reste basé sur le calcul de produits scalaires mais il fait aussi intervenir une fonction d'activation, qui complique beaucoup l'analyse mathématique. L'étude d'un réseau de neurones apporte aux élèves une ouverture sur les approches algorithmiques les plus récentes de l'intelligence artificielle. Cependant leur analyse mathématique est trop compliquée au lycée et reste mal comprise.

Un réseau de neurone à deux couches calcule d'abord chaque coordonnée k_i d'une caractéristique k, avec un vecteur de poids $\bar{w}_i$ et un biais b_i :

$$k_i = s(w_i \, . \, x - b_i).$$

La "fonction d'activation" $s(t)$ du neurone peut être une sigmoïde, ou un rectificateur $s(t) = \max(t, 0)$. La seconde couche fait une seconde séparation linéaire $w.k - b$, où w est un autre vecteur de poids et b est un biais. S'il faut séparer deux classes alors on ne garde que le signe de $w.k - b$. Les paramètres des deux couches du réseau sont les poids w_i et w et les biais b_i et b. Ils sont optimisés au cours de l'apprentissage, afin de réduire l'erreur de l'estimation sur les exemples d'entraînement.

L'utilisation de plus de 2 couches peut améliorer le calcul des caractéristiques. Les réseaux de neurones les plus performants peuvent avoir plusieurs centaines de couches, suivant les applications. Pour la reconnaissance d'images, les grands réseaux de neurones incluent plusieurs milliards de paramètres, qui correspondent aux différents poids et biais qui calculent chaque couche à partir de la précédente. Pour ChatGPT, le nombre de paramètres se compte en trillions.

3.2 Apprentissage par essais et erreurs: optimisation et analyse de fonctions

La phase d'apprentissage optimise les paramètres x de l'algorithme d'estimation, pour minimiser l'erreur entre $\hat{r}$ et r. Cela met en jeu à la fois les statistiques et l'analyse de fonctions.

Dans un challenge on fournit des données d'entraînement d avec les réponses r. L'erreur $f(x)$ est égale à la moyenne des erreurs sur tous les exemples d'entraînement entre $\hat{r}$ et r.

C'est une fonction des paramètres de x. Le cas le plus simple que l'on étudie au lycée est celui où il n'y a qu'un paramètre si bien que $f(x)$ est une fonction d'une seule variable. Pour un problème de classification d'images, l'erreur est égale au pourcentage d'exemples pour lesquels $\hat{r} = r$, pour un x fixé :

$$f(x) = \frac{\text{Nombre d'images d'entraînement mal classées}}{\text{Nombre d'images d'entraînement}}.$$

L'apprentissage cherche donc un paramètre $\hat{x}$ qui minimise la valeur de $f(x)$. Au cœur de l'apprentissage apparaît un problème mathématique et algorithmique d'optimisation, qu'il faudra résoudre efficacement avec un ordinateur. Un apprentissage par essais et erreurs revient à modifier x afin de réduire l'erreur, jusqu'à éventuellement atteindre le minimum. Cela peut se faire simplement en calculant l'erreur pour un grand nombre de x et en sélectionnant $\hat{x}$ qui minimise $f(x)$. Cela revient à tracer la fonction $f(x)$ et à identifier son minimum. Cependant, cette approche est relativement inefficace et souvent trop coûteuse en calcul lorsque l'on a beaucoup de données. Une autre approche est de modifier progressivement x afin de progressivement réduire l'erreur $f(x)$. Cela se fait avec un calcul de dérivées.

L'apprentissage en intelligence artificielle met en jeu de nombreux outils d'analyse de fonctions avec le calcul de dérivées. En terminale cela peut aller jusqu'aux notions de convexité et de convergence d'algorithmes itératifs.

3.3 Généralisation et tests: probabilités et statistiques

Le danger de la minimisation de l'erreur d'entraînement est d'apprendre un $\hat{x}$ qui s'adapte de façon trop spécifique aux données d'entraînement. Le pire serait de simplement mémoriser les réponses, comme un élève qui apprend "par cœur" les réponses d'exercices mathématiques. Cela donne une erreur d'entraînement qui est nulle mais qui devient très grande sur des nouvelles données de test. Comme on l'a déjà vu, l'objectif de l'apprentissage est de généraliser et donc de commettre peu d'erreurs sur des exemples que l'on ne connaît pas à l'avance.

Les exemples de tests sont fournis par la plateforme web du challenge. On calcule l'erreur de test pour le meilleur paramètre $\hat{x}$ appris lors de l'entraînement. Pour un problème de classification, l'erreur de test est égale au pourcentage d'exemples de tests pour lesquels $\hat{r} = r$, pour un $x = \hat{x}$:

$$f_{\text{test}}(\hat{x}) = \frac{\text{Nombre d'images de test mal classées}}{\text{Nombre d'images de test}}.$$

On ne peut pas calculer soi-même cette erreur de test car la plateforme ne nous fournit pas les réponses associées aux données de test. On obtient cette erreur en soumettant nos estimations $\hat{r}$ sur les données de test et la plateforme (qui a mémorisé les bonnes réponses)

affiche l'erreur (le score) sur un leader-board. Si les erreurs de test et d'entraînement restent du même ordre

$$f_{\text{test}}(\hat{x}) \approx f(\hat{x})$$

alors cela indique que l'on a une bonne généralisation sur des exemples inconnus.

Les erreurs de test et d'entraînement ont des fluctuations car elles sont calculées par des moyennes empiriques sur des données que l'on suppose être prises au hasard, et donc indépendantes avec la même distribution. L'amplitude de ces fluctuations est mesurée par la variance de ces moyennes empiriques. Si les exemples sont pris au hasard alors cette variance est inversement proportionnelle au nombre d'images d'entraînement et de test. Pour s'assurer que l'apprentissage sur les exemples d'entraînement ne souffre pas de ces fluctuations il faut que le nombre de données d'entraînement soit suffisamment grand relativement au nombre de paramètres que l'on veut apprendre. Ces questions permettent d'étudier des problèmes statistiques de niveaux différents, sur la moyenne, la variance et la loi des grands nombres, qui est au cœur de l'apprentissage.

4 Enseignements et pédagogie

En pratique, une difficulté de l'enseignement des mathématiques par l'expérimentation est de garantir que l'expérimentation se fasse sur un temps limité et s'inscrive dans un aller-retour équilibré avec la démarche de modélisation et de traitements mathématiques. Le temps réservé à l'enseignement des mathématiques au lycée ayant été beaucoup réduit, il faut laisser toute sa place à l'abstraction et à la compréhension des mathématiques au programme. Ces équilibres sont difficiles à mettre en place et sont le résultat d'un important travail pédagogique et didactique dans les classes, mené en co-développement avec les professeurs des labos-maths [5] de l'Éducation Nationale de l'Académie de Lille et de Paris.

Modules d'enseignement L'enseignement proposé par l'équipe de MathAData est organisé en modules dédiés à différentes parties du programme de mathématiques, en statistiques, probabilités, géométrie, analyse de fonctions, et tous incluent de l'algèbre. Ces modules peuvent être utilisés sur différents challenges de données, suivant les intérêts des professeurs et de leurs élèves.

Le premier module intègre une introduction à l'IA. Une première abstraction mathématique introduit les phases d'estimation, d'apprentissage et de test, expliquées dans la section 3. Cette présentation est suivie d'une expérimentation numérique sur un challenge, par exemple de classification d'images. Outre la prise en main du notebook informatique et des données, les élèves doivent chercher une façon de discriminer les classes avec une ou deux caractéristiques qu'ils peuvent calculer, ce qui est une première étape de modélisation mathématiques.

Cette introduction est intégrée dans un module thématique afin que les élèves fassent immédiatement le lien avec les mathématiques qu'ils apprennent. Des modules supplémen-

taires permettent d'approfondir les différents aspects du programme. Les modules couvrent les domaines suivants:

- *Statistiques et probabilités* pour l'analyse de l'erreur d'entraînement et la généralisation, avec une caractéristique: histogrammes, moyennes, médianes, variance, loi des grands nombres, probabilités conditionnelles.

- *Géométrie* dans le plan et l'espace pour la classification linéaire: équations de droites dans le plan, intersection de droites, bissectrices, produit scalaire, distance à une droite, extension à l'espace avec les équations de plans et la distance entre un point et un plan.

- *Analyse de fonctions et dérivées* pour la minimisation de l'erreur d'apprentissage par descente de dérivée: fonctions, minimum, dérivée, suite, convergence d'une suite, convexité.

Un module d'ouverture sur les réseaux de neurones à 2 couches sera aussi proposé pour les élèves qui ont déjà suivi le module d'optimisation et de classification linéaire. Celui-ci aura une forte composante d'expérimentation numérique. Ces modules peuvent aussi être l'occasion pour les élèves de faire un projet plus poussé qui peut être intégré à leur grand oral du Bac.

Allers-retours Fluidifier et bien cadrer l'aller-retour entre l'expérimentation informatique et l'enseignement mathématique est fondamental. L'expérimentation doit être une source de questions, d'intuitions, et d'idées mathématiques plutôt que de simples illustrations a posteriori. Par souci de simplicité, les modules d'enseignement suivent une même séquence pédagogique.

On commence par expliciter la motivation du problème posé par le challenge, qui a typiquement une composante qui va bien au-delà de ce challenge particulier, en lien avec des problèmes plus généraux. Cela peut être de la reconnaissance d'images, ou de chants de baleines, ou faire un diagnostic médical. On propose ensuite une approche qui donne l'idée principale qui guide la résolution du problème en lien avec le chapitre de mathématique étudié en statistiques, ou probabilité, ou géométrie, ou analyse. L'expérimentation sur ordinateur permet aux élèves de manipuler et de se familiariser avec les concepts mathématiques de façon intuitive, dans le cadre de la résolution du challenge. Cette expérience leur permet aussi de tester leurs idées, et se poser de nouvelles questions qui sont ensuite abordées par les mathématiques. Ceci est suivi d'une présentation plus formelle des concepts mathématiques sous-jacents et d'un entraînement avec des exercices sur papier, proposés par les professeurs.

Pour chaque module, l'équipe de MathAData propose une fiche de référence avec différentes séquences possibles, ainsi que des diapositives modifiables par les professeurs. Les notebooks pour l'expérimentation informatique sont aussi divisés en sous-parties, adaptables pour l'enseignement de chaque professeur. Cela ne demande pas de programmer.

Des fiches d'exercices mathématiques en lien avec le challenge sont aussi proposées. Tous ces supports pédagogiques sont faits en co-développement avec les professeurs des labos-maths [5]. Ces documents sont mis à disposition sur notre site *mathadata.fr*. Nous organisons aussi des formations pour faciliter la prise en main de ces outils.

Expérimentation numérique Pour des challenges de données, l'expérimentation se fait dans un environnement informatique qui permet d'interagir avec les données et avec les paramètres d'apprentissage, sans programmer. Cela permet aux élèves de travailler de façon plus autonome en classes de mathématiques.

Pour des classes de SNT ou de NSI, il est possible d'introduire des éléments de programmation en Python, ce qui ouvre plus de flexibilité. Cette expérimentation permet alors d'approfondir leur maîtrise de l'informatique et du langage Python. Des interfaces avec la plateforme *challengedata.ens.fr* sont proposées pour automatiser l'accès aux données et facilement soumettre les résultats sur des données de tests.

L'environnement informatique est dans un Notebook Python, intégré dans le logiciel *Basthon* de l'éducation nationale. Cela permet de travailler sur le browser web, sans installer de logiciels. On veut aussi éviter que les élèves, qui ont souvent très peu d'expérience de programmation, ne soient pas bloqués par des bugs de langage Python. Pour cela, la programmation est divisée en petites cellules qui implémentent des fonctions de quelques lignes, dont nous donnons des exemples. Ceux-ci peuvent être simplement modifiés.

Passage à l'échelle Pour passer à l'échelle nationale, il faut aider les professeurs à se former, développer et mettre à disposition des supports pédagogiques et informatiques qui sont adaptés à la diversité des élèves. Ce passage à l'échelle est une difficulté bien connue des nouvelles initiatives pédagogiques. C'est l'enjeu de notre collaboration avec les labos-maths. Cela passera aussi par la construction d'une communauté de professeurs qui partagent leurs supports pédagogiques. Ce travail est initié en collaboration avec la plateforme Capytale *https://capytale2.ac-paris.fr/web/accueil* de l'éducation nationale.

Notre objectif est aussi de faciliter les ponts entre les lycées et les universités, avec qui nous travaillons pour l'enseignement de l'IA par des challenges de données. C'est particulièrement important pour ouvrir des perspectives pour les élèves qui ignorent le plus souvent les débouchés considérables qui s'ouvrent dans ce domaine, par les mathématiques.

5 Conclusion

Pour les lycées, on voit toutes les difficultés pour réussir à mettre en œuvre l'idée pourtant naturelle et classique de l'enseignement des mathématiques à partir de questions posées par l'expérimentation. L'enjeu est considérable pour aider à mieux faire comprendre les mathématiques à une plus large population d'élèves, et de leur montrer que les mathématiques leur ouvrent des métiers d'avenir. MathAData poursuit ce travail pédagogique,

informatique et d'expérimentation dans les classes, en co-développement avec les labos maths de l'Éducation Nationale, car nous pensons que l'enjeu en vaut la chandelle.

6 Bibliographie

1. *21 mesures pour l'enseignement des mathématiques*, rapport de Cédric Villani et de Charles Torossian, Février 2018.
2. *L'expérimentation en mathématiques*, Colloque de la Ciprelem, Juin 2006.
3. *Démarche expérimentale et apprentissages mathématiques*, Rapport de l'INRP, Avril 2007.
4. *Laboratoires de Mathématiques dans le réseau de l'AEFE*, Rapport de l'éducation nationale, 2021.