Le perceptron

Sylvie Benzoni-Gavage

25 février 2022

1 Chacune son camp

Vous connaissez sans doute des sports de ballon qui se jouent à deux équipes se faisant face : basketball, football, handball, rugby... Avant le coup d'envoi, les deux équipes sont de part et d'autre de la ligne séparant le terrain en deux au milieu.

Pendant le match, les positions respectives des joueurs ou joueuses se compliquent. Si l'on trace, à un instant donné, une ligne imaginaire séparant les deux équipes, elle n'est généralement pas droite. Si par exemple les équipes se plaçaient comme des figurines de baby-foot, la ligne de séparation pourrait ressembler à celle tracée sur la photo.

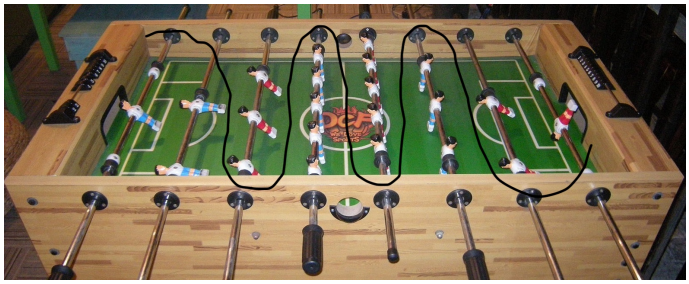

FIGURE 1 – Ligne de séparation des équipes au baby-foot

On peut penser à des situations plus simples, ou plus compliquées, selon l'espace dans lequel les équipes se déplacent. En plus simple il y a le tir à la corde : les deux équipes se déplacent en ligne droite, et elles sont juste séparées par un point sur la corde.

FIGURE 2 – Tir à la corde en Bretagne. Crédit : Jacques Le Letty

Il y a bien sûr les sports de ballon sur un terrain plat mentionnés plus haut. En plus compliqué on peut imaginer le quidditch, sport de l'univers Harry Potter dans lequel les joueurs et joueuses se déplacent en trois dimensions sur leurs balais volants. La séparation des deux équipes à chaque instant peut se faire non par une ligne, mais par une surface (tout aussi imaginaire que la ligne séparant les équipes de foot).

FIGURE 3 – Quidditch en Lego. Crédit : Tim Moreillon

On pourrait encore passer à une dimension supplémentaire en imaginant une bataille dans l'espace-temps, comme dans le film Tenet par exemple : les combattantes des deux camps pourraient alors être séparées par une hypersurface [1]. Mais rassurez-vous si, comme beaucoup

1. Une hypersurface dans un espace à 4 dimensions est l'analogue d'une surface dans un espace à 3

de personnes, vous n'avez rien compris à ce film : dans la suite de cet article, les dimensions supplémentaires devraient vous apparaître plus naturelles [2].

2 Un zeste de jargon

Cette histoire de séparation en deux «camps» est à la base d'une méthode importante en science des données, qui trouve sa source dans un objet étrange appelé perceptron. Plus précisément, le perceptron est avant tout un objet immatériel, un concept élaboré à la fin des années 1950 par un psychologue, Frank Rosenblatt [5]. Le perceptron est un « système nerveux hypothétique », inspiré de ce que l'on savait à l'époque de la manière dont le cerveau humain perçoit le monde extérieur.

Ce système était censé pouvoir reproduire la façon dont notre cerveau distingue des images. Plus précisément, le principe de base de ce système se rapproche de la façon dont notre cerveau distingue la répartition de l'équipe bleue sur le terrain, lorsque nous regardons un match de foot où joue l'équipe de France. Souvenez vous : on peut mentalement séparer les deux équipes par une ligne imaginaire.

Le perceptron fut conçu comme un algorithme, c'est-à-dire une succession d'instructions de logique et de calcul, pour la reconnaissance d'images (bien avant l'avènement de l'imagerie numérique), et implémenté dans l'une des toutes premières machines électroniques : le Mark I perceptron.

---

dimensions, d'une ligne (ou courbe) dans un espace à deux dimensions, et d'un point dans un espace à une dimension seulement.

2. Pour se familiariser avec la notion de dimension, il est toujours conseillé de regarder le film Dimensions ... une promenade mathématique...

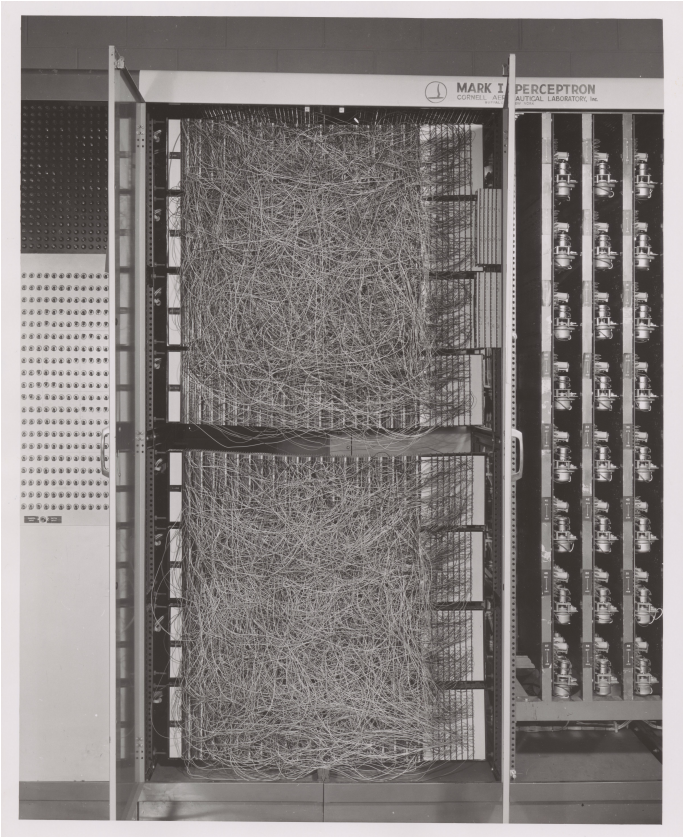

FIGURE 4 – Mark I Perceptron au Cornell Aeronautical Laboratory, © Cornell University Library

Même si cette machine n'a pas tenu toutes ses promesses [4], le perceptron a en quelque sorte signé la naissance de ce que l'on appelle l'apprentissage supervisé, aujourd'hui utilisé à grande échelle sur les quantités faramineuses de données d'Internet (les fameuses «big data»). L'apprentissage supervisé fait partie de l'apprentissage automatique (ou machine learning), qui alimente les systèmes d'intelligence artificielle vous proposant des vidéos à regarder ou des musiques à écouter, par exemple.

Le but ici n'est pas d'entrer dans le détail du perceptron et encore moins de ses successeurs, appelés réseaux de neurones. L'idée est de montrer comment et quelles mathématiques interviennent dans la séparation de données, en particulier pour l'apprentissage supervisé. Pour cela on va remonter à un article de 1965 par le théoricien de l'information Thomas Cover [2]. Comme Rosenblatt, Cover s'intéressait à la reconnaissance d'images. Cependant, ses résultats s'appliquent à toutes sortes de données.

La suite de cet article vise à expliquer le résultat le plus frappant montré par Cover en 1965. Il concerne le seuil sur la quantité de données que l'on peut séparer en fonction de la dimension de l'espace dans lequel elles se trouvent. Formulé ainsi, il reste assez mystérieux. On va y arriver progressivement.

3 Données

Mettons de côté le jargon - perceptron, algorithme, apprentissage supervisé / automatique, intelligence artificielle, réseaux de neurones etc. - et posons nous cette question simple : qu'est-ce qu'une donnée ?

Si l'on pense aux données personnelles (supposément protégées par la loi), elles concernent des individus d'une population. L'âge est une donnée personnelle, par exemple.

De manière générale, une donnée peut être une information qualitative, comme le genre d'une personne (si tant est qu'il soit bien défini), ou une information quantitative, c'est-à-dire une valeur numérique (un nombre), exprimée ou non dans une unité de mesure. En l'occurrence, l'âge est une information quantitative, le plus souvent exprimée en années (ou en mois pour les bébés). Une donnée sans unité de mesure pour un individu est par exemple son nombre d'amis : s'il est difficile à mesurer dans la vraie vie, on peut le définir précisément comme son nombre d'abonnés sur un réseau social donné, par exemple.

L'avantage des informations quantitatives est qu'elles peuvent se représenter sur un graphique. Pour ce qui va nous intéresser, on représentera ces informations comme des collections de points : pour une population donnée, à chaque individu correspondra un point. Ce sera un point sur une droite graduée lorsqu'on n'a qu'un type d'information par individu, comme par exemple l'âge. Ce sera un point dans un plan lorsqu'on a deux types d'information par individu : par exemple chaque point aura pour abscisse l'âge de l'individu et pour ordonnée son nombre d'amis.

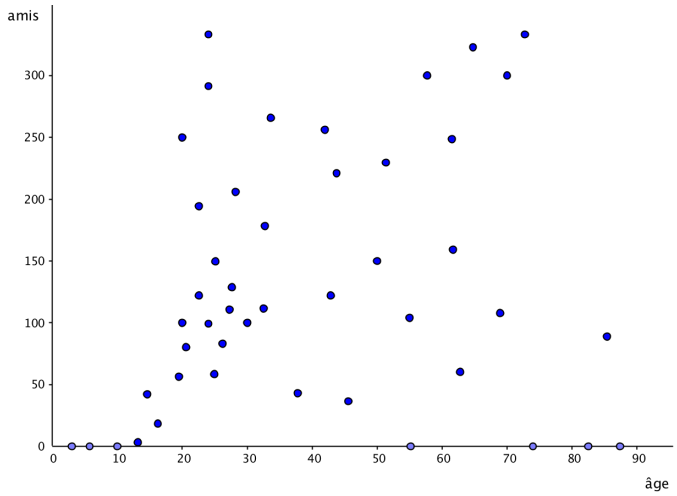

FIGURE 5 – Nuage de 50 points correspondant aux données de 50 individus fictifs

Si l'on dispose d'un troisième type d'information on peut représenter les trois informations quantitatives que l'on connaît pour chaque individu par un point dans l'espace à trois dimensions. Et il n'y a pas de raison de s'arrêter à trois dimensions, même si cela devient plus difficile à se représenter : chaque nouveau type d'information introduit naturellement

une dimension supplémentaire.

Pour éviter toute confusion dans ce qui suit, on considèrera comme des données uniquement des informations quantitatives, autrement dit, des données numériques. De plus et pour simplifier, on appellera donnée au singulier l'ensemble des informations considérées pour chaque individu : par exemple son âge et son nombre d'amis constitueront «sa» donnée, en l'absence d'informations supplémentaires.

En fait, les données peuvent tout aussi bien concerner des individus d'une population que toute sorte d'autres objets d'étude et notamment des images, dont la reconnaissance avait motivé l'élaboration du perceptron. Depuis l'avènement des images numériques, on peut facilement associer à chaque image un certain nombre d'informations quantitatives : pour une image couleur constituée d'un million de pixels, chaque pixel ayant trois niveaux de couleur (rouge, vert et bleu), il y a trois millions d'informations quantitatives ; on peut alors représenter la donnée d'une image comme un point dans l'espace à trois millions de dimensions ! Si on a 100 images on aura 100 points dans l'espace à trois millions de dimensions.

De manière générale, un jeu de données se représente par une collection de points dans un espace d'une certaine dimension. On parle aussi de nuage de points, par analogie avec les gouttes d'eau dans notre ciel à trois dimensions. Que ce soient 50 points dans le plan, de dimension 2, comme sur la figure 5, ou 100 points dans l'espace à trois millions de dimensions représentant 100 images numériques en couleur, un jeu de données peut toujours être représenté par un nuage de points.

FIGURE 6 – Nuage © Khardan CC-BY-SA FR

Dans toute la suite on notera N le nombre de points dans le nuage et d la dimension de l'espace où se trouve ce nuage. Sur la photo d'un vrai nuage on imagine que le nombre N de gouttes d'eau est déjà très grand, puisqu'on ne les distingue même pas à l'œil nu. Mais la dimension d de l'espace dans lequel baigne ce nuage, en l'occurrence le ciel, est juste égale à trois.

En pratique, pour des jeux de données et notamment ceux issus d'Internet, aussi bien N que d pourront être très très grands. Par exemple N se compte en milliards si l'on considère

tous et toutes les internautes de la Terre, en dizaines de milliards si l'on compte les images disponibles sur Internet, tandis que d se compte en millions pour chaque image. Les sciences expérimentales ou numériques produisent elles-mêmes des quantités considérables de données.

La science des données consiste à étudier les propriétés statistiques, mais aussi géométriques, des jeux de données. C'est d'ailleurs par ces mots, «propriétés géométriques et statistiques», que commence le titre de l'article fondateur de Cover dont nous allons parler. Le sujet principal, très important aujourd'hui en science des données, en est la séparation de données. C'est ici qu'on rejoint le début de cet article : séparer un jeu de données en deux, c'est un peu comme séparer deux équipes !

4 Oui ou non

Voyons comment faire parler les données. Imaginons un jeu de données simple, par exemple constitué de l'âge d'une certaine population. On pourrait en faire une première analyse statistique, tracer la « pyramide des âges » par exemple, mais ce n'est pas le sujet. Rappelez vous : on va représenter chaque individu par un point sur une droite, une demi-droite en vérité, puisque les âges sont des nombres positifs. C'est possible sur un graphique de taille raisonnable tant que la population n'est pas trop nombreuse.

Imaginons maintenant qu'on ait une information supplémentaire pour chaque individu, une information qualitative cette fois-ci. Par exemple, on sait si cet individu aime le café ou pas. Alors notre représentation graphique a des chances de ressembler à la figure 7 (on laisse la lectrice imaginer l'échelle, ce sont des données fictives de toute façon). La répartition des amatrices de café selon leur âge dépend sans doute des régions du monde, etc. C'est juste un exemple.

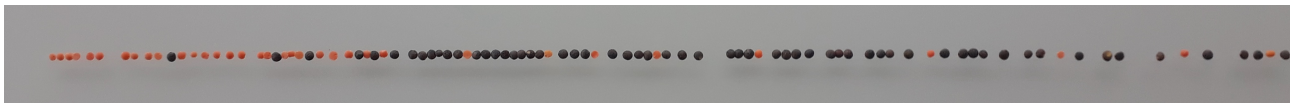

FIGURE 7 – Représentation selon leur âge de personnes aimant le café (en vert foncé) et ne l'aimant pas (couleur corail) dans une population donnée. (On a posé des lentilles sur une table pour donner un peu de relief à ce «nuage» rectiligne.)

On n'arrive pas à séparer parfaitement par un point les amatrices de café des autres (on peut comprendre que le seul critère d'âge ne soit pas suffisant). Cependant on aperçoit un seuil, c'est-à-dire un âge en dessous duquel la plupart des personnes n'aiment pas le café et au dessus duquel la plupart des personnes l'aiment.

Cela revient à cela, séparer les données ! Quel est l'intérêt ? Et bien si on prend un individu mystère dont on connaît l'âge, on pourra prédire s'il a des chances d'aimer le café selon qu'il se trouve d'un côté ou de l'autre du seuil.

Cet exemple simpliste se généralise à des jeux de données en toute dimension. C'est l'idée de l'apprentissage supervisé. On entraîne la machine à trouver des séparations entre données pour lesquelles on connaît une information qualitative de type « oui ou non » (aimer ou pas

le café dans l'exemple), puis la machine utilise cette séparation pour prédire si de nouvelles données relèvent du oui ou du non.

5 La séparation idéale

Intéressons nous à la première étape de l'apprentissage supervisé, qui part de données enrichies d'une information qualitative de type « oui ou non » : on dit aussi qu'elles sont classifiées, sous-entendu en deux classes (comme deux équipes), ou étiquetées (l'étiquette correspondant au maillot de l'une ou de l'autre équipe).

Comme on l'a vu, considérer des données en une seule dimension est vraiment réducteur. En deux dimensions, la situation idéale se présente lorsque des données classifiées sont séparées par une droite, comme sur la figure 8 et comme nos joueuses de foot avant le coup d'envoi.

FIGURE 8 – Séparation de données en dimension deux par une droite : les lentilles corail correspondent à une classe de données et les lentilles vertes à l'autre ; une droite de séparation est matérialisée par un spaghetti cru.

On se doute que c'est assez peu probable pour des données « prises au hasard » : nous reviendrons sur le sens précis de cette affirmation plus loin. Une situation plus probable est celle où les données classifiées sont séparées par une courbe, comme sur la figure 9 et comme nos joueuses de foot pendant le match.

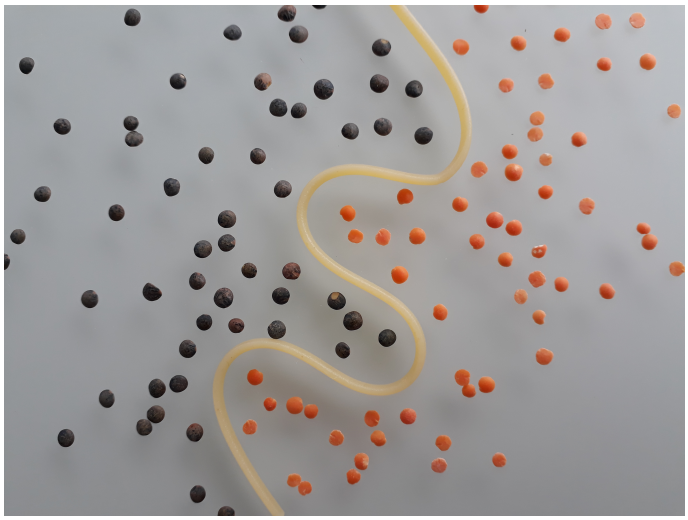

FIGURE 9 – Séparation de données en dimension deux par une courbe : les lentilles corail correspondent à une classe de données et les lentilles vertes à l'autre ; une courbe de séparation est matérialisée par un spaghetti cuit.

On comprend néanmoins que cela devient vite compliqué, de séparer les données. Mais compliqué comment ? Possible ou impossible ? Probable ou improbable ? C'est l'objet de la suite.

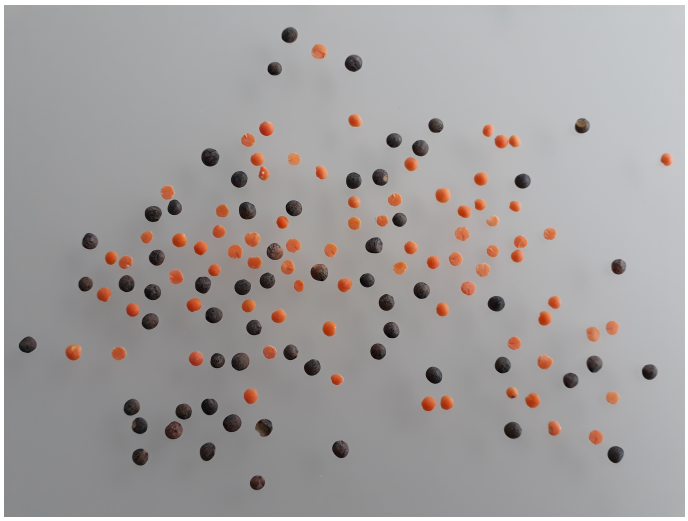

FIGURE 10 – Données en dimension deux plus difficiles à séparer

6 Les chances de séparation

Même pour un petit nombre de données, la séparation n'est pas évidente. Regardez la figure 11. Les cinq points sont grosso modo situés de la même manière dans les deux cas,

mais leurs étiquettes (couleur vert foncé ou corail) sont différentes. Dans le cas de droite, on peut séparer la classe de couleur verte par un cercle, on ne peut pas la séparer par une droite.

FIGURE 11 – Séparation par une droite ou par un cercle (approximativement représenté par un bout de spaghetti cuit)

Autrement dit, lorsqu'on change les étiquettes, des données qui étaient séparables par une droite peuvent cesser de l'être, mais rester séparables par une courbe «simple» telles qu'un cercle.

6.1 Séparation en dimension un

Pour estimer nos chances de pouvoir séparer des données étiquetées, revenons à notre exemple le plus simple : celui de données en une dimension. Si l'on repense aux lentilles, on se doute qu'il y a peu de chances qu'un mélange de lentilles vertes et de lentilles corail alignées sur une droite soit bien séparé, avec une couleur d'un côté et l'autre couleur de l'autre. Mais ce n'est pas exactement la question, car les lentilles ont une couleur fixée à l'avance.

Pour poser plus correctement la question, on peut plutôt penser aux joueuses du tir à la corde : si elles sont alignées le long de la corde sans savoir a priori qui est dans quelle équipe et si on distribue à chaque joueuse un maillot bleu ou rouge au hasard, il y a peu de chances qu'on se retrouve avec toutes les joueuses rouges d'un côté et toutes les bleues de l'autre.

On peut en fait calculer exactement la probabilité que cela arrive. Disons qu'on a N joueuses. On distribue vraiment les maillots au hasard, en piochant dans un tas de N maillots bleus et N maillots rouges, sans se soucier de faire deux équipes de taille égale. Chaque joueuse ayant une chance sur deux de recevoir un maillot bleu, il y a $\underbrace{2 \times \cdots \times 2}_{N fois} = 2^N$ façons de distribuer les maillots : $2 = 2^1$ possibilités pour une joueuse, $4 = 2^2$ pour deux joueuses, $8 = 2^3$ pour trois joueuses, etc. Parmi ces 2^N distributions possibles, combien y en a-t-il pour lesquelles toutes les joueuses recevant un maillot bleu soient les unes à côté des autres, sans aucune joueuse au maillot rouge entre elles ?

Pour compter les distributions correspondant à cette situation, on peut s'imaginer remonter la file de joueuses par exemple de gauche à droite. À la joueuse la plus à gauche je donne au hasard un maillot bleu ou un maillot rouge. Cela fait deux choix possibles, mais disons qu'il est bleu. Alors je remonte la file vers la droite en distribuant un certain nombre de maillots bleus. Disons k maillots bleus : ce nombre k est pris au hasard entre 1 et N ;

autrement dit, cela fait N possibilités. Le reste de la distribution est ensuite complètement fixé car je ne peux plus changer de couleur, pour que les équipes soient bien séparées. Si k est égal à N je n'aurai en fait pas distribué de maillot rouge et il n'y aura qu'une équipe : tant pis pour le sport, cette possibilité compte quand même. Si k est strictement plus petit que N, je distribue des maillots rouges à toutes les joueuses se trouvant à droite des joueuses en bleu.

Finalement, puisqu'on a deux choix possibles au départ (le bleu ou le rouge à la première joueuse), il y a en tout $2N$ distributions possibles pour avoir des équipes bien séparées de part et d'autre d'un point. La probabilité de tomber sur l'une d'elle est $p = \frac{2N}{2^N}$. On s'aperçoit que cette probabilité p devient vite petite quand N augmente : pour $N = 10$ par exemple, on a $p = \frac{20}{1024}$ soit environ 2%. Pour $N = 100$, on a $p = \frac{200}{1024^{10}}$, ce qui est bien plus petit que

$$\frac{200}{1000^{10}} = \frac{2}{10^{28}} = 0,00000000000000000000000000002.$$

Autrement dit, il est tout à fait improbable, pour ne pas dire impossible, que 100 joueuses alignées à qui l'on distribue un maillot bleu ou rouge au hasard se retrouvent avec toutes les bleues d'un côté et toutes les rouges de l'autre.

6.2 Séparation en dimension deux

Comme on vient de le voir, il est très peu probable de pouvoir séparer un grand nombre de données unidimensionnelles. On peut se poser la question analogue en ajoutant une dimension. Imaginons des joueuses placées sur un terrain sur lequel on n'a pas encore tracé la ligne de séparation au coup d'envoi, et qu'on leur distribue au hasard un maillot bleu ou un maillot rouge. Quelle chance y a-t-il qu'on puisse tracer une ligne droite séparant les bleues des rouges ?

C'est là encore une question de combinatoire, c'est-à-dire qu'on va devoir compter les possibilités. Mais on se doute qu'il va falloir faire aussi un peu de géométrie : on dit que cette question relève de la géométrie combinatoire.

On remarque qu'on aura un petit problème pour compter les distributions possibles si jamais trois joueuses sont alignées entre elles. Car alors, si elles reçoivent des maillots par exemple comme sur la figure 12 à gauche, il n'y a aucun moyen de les séparer par une droite, mais cela ne tient à pas grand chose : si elles étaient placées un tout petit peu différemment, comme sur la figure 12 à droite, on pourrait bel et bien les séparer par une droite.

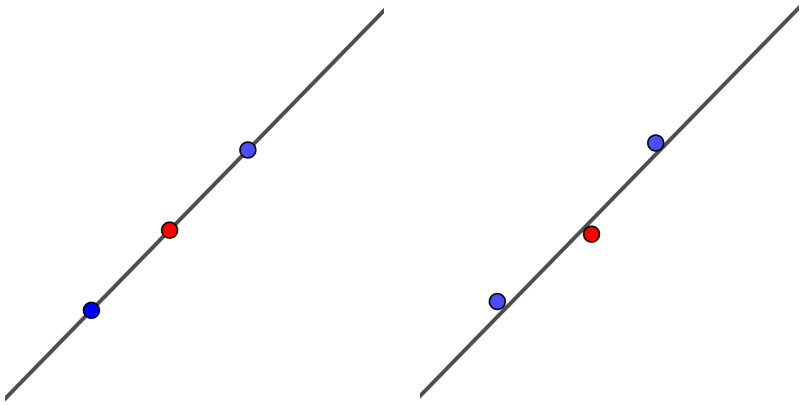

FIGURE 12 – Dispositions possibles de deux joueuses bleues et une joueuse rouge

Aussi, pour compter sans problème les possibilités, on va exclure le cas de gauche. On supposera qu'il n'y a aucun groupe de trois joueuses qui soit aligné. Mathématiquement parlant, on dira qu'elle sont placées sur le terrain en *position générale*.

Commençons par remarquer que pour un groupe d'au maximum 3 joueuses en position générale et pour n'importe quelle distribution de maillots, on peut toujours séparer par une ligne droite les bleues des rouges. Autrement dit, pour N joueuses avec N au plus égal à 3, le nombre de distributions séparables par une ligne droite est égal au nombre total de distributions, à savoir 2^N. Ceci se voit par inspection des distributions possibles, qui ne sont pas trop nombreuses.

Notons que les distributions où il n'y a que des bleues ou que des rouges se traitent facilement : il suffit de positionner une droite telle que toutes les joueuses soient dans un même demi-plan délimité par cette droite. C'est notamment ce qui arrive avec une seule joueuse (qui peut bien sûr n'être que d'une couleur). Le cas $N = 2$ se fait encore de tête, car parmi les 4 distributions possibles, 2 correspondent à une seule couleur, et pour les deux autres, on a exactement un point bleu et un point rouge, qu'il suffit par exemple de séparer par leur médiatrice.

Le cas de 3 joueuses mérite de faire un dessin. Le fait qu'elles soient en position générale indique que leurs positions dans le plan forme un «vrai» triangle, non aplati. La figure 13 montre les huit distributions possibles, toutes séparées par une ligne droite.

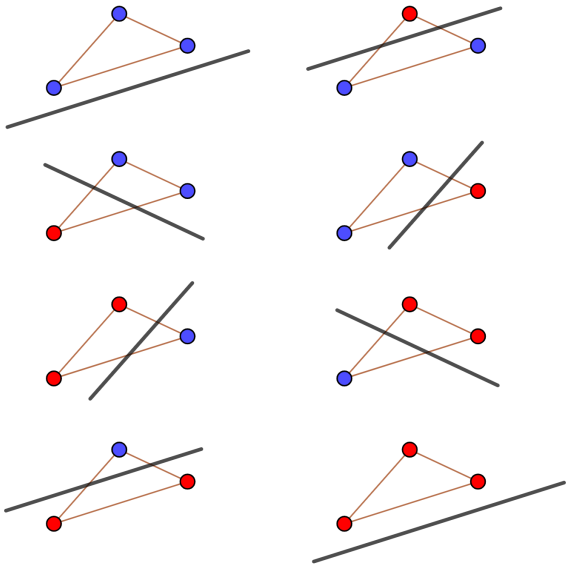

FIGURE 13 – Distributions possibles de maillots entre trois joueuses en position générale

C'est à partir de 4 joueuses que cela se complique. Pour 4 exactement, on peut encore faire un dessin même si cela devient fastidieux. Si on repart des 8 distributions possibles pour 3 des joueuses, on peut représenter toutes les distributions possibles avec une joueuse de plus et observer celles qui sont séparables par une ligne droite et éventuellement celles qui ne le sont pas. Cette approche doit inclure deux cas de figure, selon que la 4ème joueuse est située à l'intérieur (Figure 14) ou à l'extérieur (Figure 15) du triangle formé par les 3 premières. Notons qu'elle ne peut pas être sur le triangle si l'on suppose les 4 joueuses en position générale. Sur les deux figures sont représentées des droites passant par le 4ème point et qui séparent les trois premiers selon leur couleur, lorsque c'est possible. On verra plus loin l'importance de ces droites pour comprendre le décompte des distributions séparables.

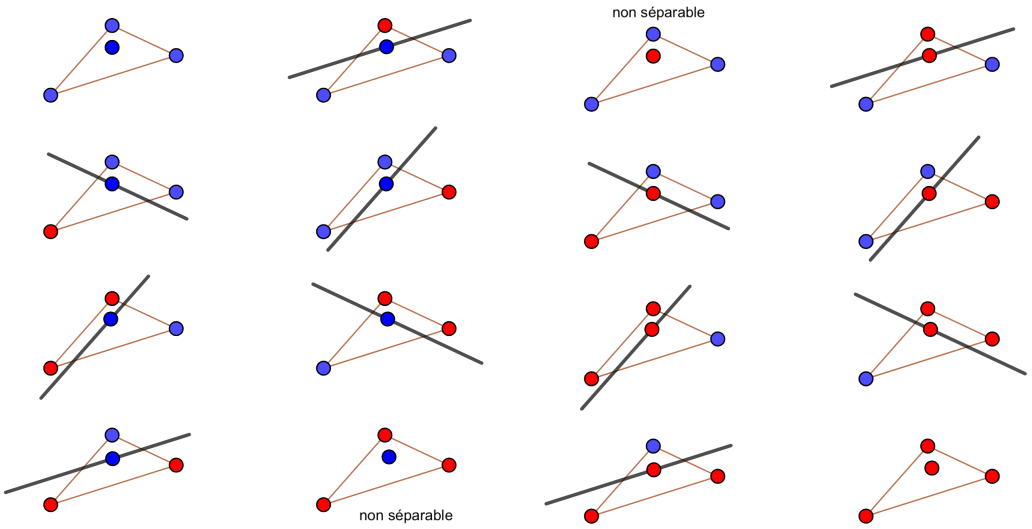

FIGURE 14 – Distributions possibles de maillots entre quatre joueuses en position générale, lorsque la 4ème joueuse est à l'intérieur du triangle formé par les trois premières.

Plus précisément, lorsque la 4ème joueuse est à l'extérieur du triangle, on peut supposer qu'elle n'est pas dans les secteurs extérieurs délimités par le prolongement des côtés des triangles : si c'était le cas on serait ramené au cas où l'une des joueuses est à l'intérieur du triangle formé par les 3 autres, c'est-à-dire aux distributions de la figure 14.

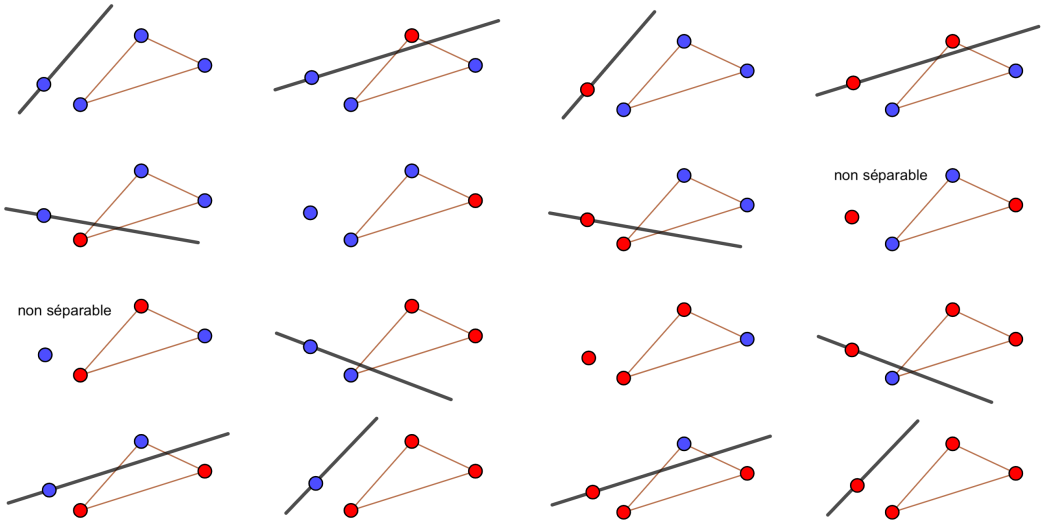

FIGURE 15 – Distributions possibles de maillots entre quatre joueuses en position générale, lorsque la 4ème joueuse est à l'intérieur du triangle formé par les trois premières, et qu'aucune joueuse n'est à l'intérieur du triangle formé par les trois autres.

On observe dans les deux cas de figure que deux des distributions ne sont pas séparables par une ligne droite. Autrement dit, quelles que soient les positions respectives des quatre joueuses, pourvu qu'elles soient en position générale, il y a exactement 14 distributions séparables sur les 16 distributions possibles. Ce nombre 14 est indépendant des positions des joueuses, même si le détail des distributions non séparables est différent selon les cas de figure : on remarquera sur la figure 14 que les deux distributions non séparables correspondent à trois maillots de la même couleur et un de l'autre couleur, tandis que sur la figure 15 les deux distributions non séparables correspondent à deux maillots bleus et deux maillots rouges.

Lorsque le nombre N de points augmente indéfiniment, on ne peut plus se reposer sur un décompte exhaustif. Néanmoins, si on comprend bien comment on est passé de 3 à 4, on comprendra comment passer de N à $N+1$.

Deux façons de compter On peut déjà comprendre ce qui caractérise les distributions séparables dans le cas $N = 4$. Dans les deux cas de figure (Figure 14 et Figure 15), on observe qu'il y a deux types de distributions séparables. D'une part il y a celles pour lesquelles on peut faire passer par le 4ème point une droite de séparation des trois premiers points selon leur couleur : elles sont au nombre de 12. D'autre part il y a celles pour lesquelles la couleur du 4ème point est imposée par sa position et la couleur des autres, pour que la distribution soit séparable : cela en fait deux. Et donc 14 au total.

Le fait de pouvoir faire passer une droite de séparation des trois premiers points par le 4ème est la clé principale de compréhension de notre dénombrement, que l'on peut voir comme $14 = 12 + 2$ mais aussi comme $14 = 8 + 6$. En effet, on peut d'abord compter, dans les deux cas de figure, le nombre de distributions sur les trois premiers points pour lesquelles on peut faire passer une droite de séparation par le 4ème. On en trouve 6. Pour chacune de ces 6 distributions, quitte à bouger un tout petit peu ladite droite, on peut séparer les quatre points, que le 4ème soit bleu ou rouge : en partant d'une telle distribution sur trois points, on en obtient deux encore séparables sur quatre points. Pour les deux autres distributions sur les trois premiers points, la couleur du 4ème est imposée, si l'on veut une distribution séparable, Autrement dit, le nombre de distributions séparables pour les quatre joueuses est celui des distributions (forcément séparables) sur 3 joueuses, c'est-à-dire 8, augmenté du nombre de distributions sur les trois premières joueuses qui donnent deux distributions séparables sur les quatre joueuses, c'est-à-dire 6. Au final on en trouve donc bien $8 + 6 = 14$. Une dernière remarque numérologique à ce stade est que $6 = 2 \times 3$.

Décompte pour les plus aguerries On va généraliser ce raisonnement pour passer de N joueuses à $N+1$ joueuses, avec N quelconque. Ceci permettra de montrer par récurrence que le nombre de distributions séparables à N joueuses en position générale ne dépend pas de leurs positions respectives et de calculer ce nombre.

Prenons donc un entier N supérieur ou égal à 3 et notons $c(N)$ le nombre de distributions séparables à N joueuses en position générale. On suppose que ce nombre est bien défini, en ce sens qu'il ne dépend pas des positions respectives des joueuses en position générale. On va montrer que le nombre de distributions séparables à $N+1$ joueuses en position générale ne dépend pas non plus de leurs positions respectives et qu'en outre ce nombre vaut $c(N+1) = c(N) + 2N$ (comme ce que l'on a vu pour $N = 3$: rappelez vous, $14 = 8 + 2 \times 3$).

La seule subtilité nouvelle par rapport à ce qu'on a fait dans le cas $N = 3$, c'est qu'on se donne une position générale arbitraire des $N+1$ joueuses, et qu'on va en choisir une parmi elles dont la position permet de se ramener à ce que l'on sait du cas de N joueuses. Cela nous évitera des considérations géométriques comme celles qui nous ont fait distinguer deux cas de figures pour $N = 3$, et qui seraient ingérables pour N quelconque.

Choisissons donc une joueuse séparée de toutes les autres par une droite. C'est toujours possible pour un nombre fini de points. Appelons A la position de cette joueuse et D une droite qui la sépare des N autres. Pour la même raison que dans le cas $N = 3$, le nombre total de distributions séparables sur l'ensemble des $N+1$ joueuses est égal à celui sur les N joueuses augmenté du nombre de distributions sur ces N joueuses pour lesquelles on peut faire passer une droite de séparation par la $(N+1)$ème. Or ce dernier nombre est exactement $2N$. On peut s'en convaincre en faisant passer par A une sorte de canne à pêche qui ramène tous les autres points sur la droite D, comme sur la figure 16.

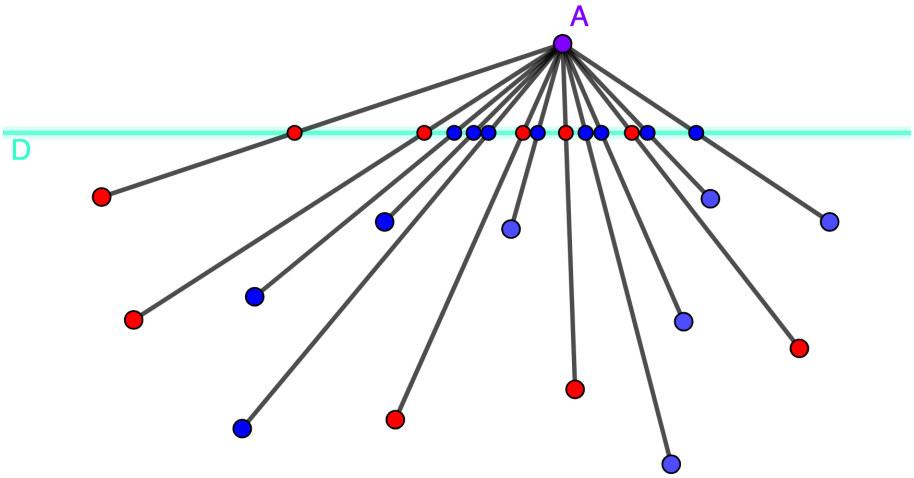

FIGURE 16 – Pêche aux N joueuses séparées de celle en A par la droite D.

Séparer les N joueuses autres que celle placée en A par une droite passant par A revient en effet exactement à séparer les points de la droite D obtenus comme intersections entre D et les droites reliant A aux autres joueuses : ces droites contiennent exactement une autre joueuse en plus de A, du fait de l'hypothèse de position générale. Comme on l'a vu dans le cas monodimensionnel, il y a exactement $2N$ distributions séparables de ces N points. On a donc bien $c(N+1) = c(N) + 2N$ comme annoncé.

On peut déduire de la formule de récurrence $c(N+1) = c(N) + 2N$ pour $N \geqslant 3$ et de $c(3) = 8$ une expression explicite de $c(N)$ pour tout N :

$$c(N) = 8 + \sum_{n=3}^{N-1} 2n = 2 + N(N-1)\,.$$

Ainsi, la probabilité qu'une distribution de maillots à N joueuses dans le plan soit séparée par une ligne droite est $p = p(N)$ avec :

- pour N au plus égal à 3, $p(N) = 1$,
- pour $N \geqslant 4$,

$$p(N) = \frac{2 + N(N-1)}{2^N}\,.$$

Celle-ci est plus élevée que la probabilité $p = 2N/2^N$ trouvée en dimension 1. On a par exemple

$$p(10) = \frac{2+90}{1024} \simeq 9\%\,.$$

Cependant $p(N)$ devient encore petite rapidement quand N augmente : $p(N)$ passe en dessous de 2% dès $N = 13$ (contre $N = 10$ en dimension 1). La lectrice le vérifiera facilement sur sa calculatrice. La lectrice pourra aussi voir à la main que

$$p(100) = \frac{2+9900}{1024^{10}} < \frac{10\ 000}{1000^{10}} = \frac{1}{10^{26}} = 0,00000000000000000000000001$$

est encore extrêmement petite, même si elle est de l'ordre de 100 fois celle en dimension 1.

6.3 Séparation en dimension quelconque

Il n'y a pas de raison de s'arrêter à la dimension deux ! Imaginons des individus vivant dans un espace de dimension d et à qui l'on distribue aléatoirement des maillots bleus ou rouges. Vous pouvez penser par exemple aux joueuses de quidditch, en dimension 3, ou aux protagonistes dans le film Tenet, en dimension 4. Le problème de leur séparation est tout à fait analogue à celui qu'on a vu en dimension 2.

Les droites que l'on avait pour séparer deux équipes sur un terrain en dimension 2 deviennent des plans pour séparer deux équipes (par exemple de quidditch) en dimension 3. À partir de la dimension 4 on appelle *hyperplan* tout sous-espace de dimension un de moins que l'espace ambiant. Les hyperplans ont une caractéristique très importante. Tel un coup de couteau, ils séparent l'espace ambiant en deux. Autrement dit, chaque point de l'espace est forcément d'un côté ou de l'autre de l'hyperplan, à moins qu'il soit carrément dessus. De plus, pour décider si un point est d'un côté ou de l'autre, il suffit de regarder le signe d'une quantité vraiment simple à écrire. C'est la raison majeure pour laquelle on préfère séparer les points par des hyperplans que par des frontières plus compliquées.

On peut par ailleurs généraliser la notion de position générale aux espaces de dimension d aussi grande qu'on veut. Un nuage de points dans un espace de dimension d est dit en

position générale s'il ne comprend pas 3 points alignés, ni 4 points coplanaires (c'est-à-dire appartenant à un même plan), ni k points dans un espace de dimension $k-2$, quel que soit k entre 4 et d.

Munis de notre outil de coupe, les hyperplans, et de cette notion de position générale, nous pouvons généraliser le résultat vu en dimension 2. En utilisant des arguments analogues, on démontre le résultat suivant.

Théorème 1. *Le nombre de distributions de maillots bleus ou rouges pour lesquelles on peut séparer les maillots bleus des rouges par un hyperplan ne dépend que de la dimension d de l'espace ambiant et du nombre d'individus N, pourvu qu'ils soient disposés en position générale. De plus, si on note $C(N,d)$ ce nombre, on a la formule de récurrence suivante :*

$$C(N+1,d) = C(N,d) + C(N,d-1)\,.$$

Si vous avez déjà rencontré le triangle de Pascal, la formule du théorème doit vous dire quelque chose. Car elle est aussi satisfaite par les coefficients binomiaux [3] $\binom{N}{d}$:

$$\binom{N+1}{d} = \binom{N}{d} + \binom{N}{d-1}.$$

Pourtant, lorsqu'on calcule effectivement nos nombres $C(N,d)$, on ne trouve pas du tout les coefficients binomiaux. La raison en est que les valeurs de départ ne sont pas les mêmes [4]. On a en effet $C(1,d) = 2$ pour tout d. Alors que les coefficients binomiaux $\binom{1}{d}$ sont tels que $\binom{1}{1} = 1$ tandis que tous les autres sont nuls, pour $d \geqslant 2$.

La connaissance de $C(1,d)$ pour tout d et de $C(2,1) = 2^2 = 4$ (vue précédemment aussi) et la formule de récurrence suffisent à remplir de ligne en ligne le tableau de la Figure 17, pour $N \geqslant 1$ et $d \geqslant 1$.

Par comparaison, le triangle de Pascal peut aussi s'écrire sous forme de tableau «plein», mais les termes au dessus de la diagonale sont nuls. Alors que dans notre tableau on a $C(N,d) = 2^N$ pour $N \leqslant d+1$.

D'après ce qu'on a vu en dimension $d = 1$, $C(N,1) = 2N$, ce qui se lit aussi dans le tableau (Figure 17), de sorte que la formule de récurrence

$$C(N+1,d) = C(N,d) + C(N,d-1)$$

redonne bien celle qu'on avait vue pour $C(N,2) = c(N)$:

$$C(N+1,2) = C(N,2) + 2N\,.$$

3. Le coefficient binomial $\binom{N}{d}$ se lit «d parmi N» et se définit comme le nombre de parties à d éléments d'un ensemble à N éléments, ou encore, comme on le fait au lycée, comme le nombre de chemins conduisant à d succès sur l'arbre représentant l'expérience de N tirages à pile ou face (en choisissant à l'avance si c'est pile ou si c'est face qui est un succès et en utilisant une pièce non truquée, de sorte que pile et face ont autant de chance de sortir l'un que l'autre).

4. En langage lycéen, on dirait que c'est l'initialisation de la récurrence qui change, entre le calcul des $C(N,d)$ et celui des $\binom{N}{d}$.

H

N									
1	2	2	2	2	2	2	2	2	...
2	4	4	4	4	4	4	4	4	...
3	6	8	8	8	8	8	8	8	...
4	8	14	16	16	16	16	16	16	...
5	10	22	30	32	32	32	32	32	...
6	12	32	52	62	64	64	64	64	...
⋮	⋮					⋱			
N	$2N$	...					2^N	2^N	...
d	1	2	3	4	5	...	$N-1$	N	...

FIGURE 17 – Tableau de valeurs donnant le nombre de distributions séparables pour N données en position générale en dimension d.

H

N									
1	1	0	0	0	0	0	0	0	...
2	2	1	0	0	0	0	0	0	...
3	3	3	1	0	0	0	0	0	...
4	4	6	4	1	0	0	0	0	...
5	5	10	10	5	1	0	0	0	...
6	6	15	20	15	6	1	0	0	...
⋮	⋮					⋱			
N	N	...					N	1	...
d	1	2	3	4	5	...	$N-1$	N	...

FIGURE 18 – Tableau de valeurs des coefficients binonomiaux (le triangle de Pascal se trouve en dessous de la diagonale, sans sa première colonne).

Cette formule avait permis de trouver la formule explicite : $C(N,2) = 2 + N(N-1)$ pour $N \geqslant 3$.

Pour celles et ceux qui veulent voir ce qu'il y a derrière les points de suspension du tableau des $C(N,d)$. On peut trouver une formule «explicite» générale pour $C(N,d)$, grâce à la connaissance de $C(1,d)$ pour tout d, de $C(2,1) = 4$ et à la formule de récurrence. Cette formule générale s'exprime ainsi à l'aide de coefficients binomiaux :

$$C(N,d) = 2\sum_{k=0}^{d} \binom{N-1}{k}.$$

La lectrice pourra vérifier que ceci redonne bien $C(N,2) = 2 + N(N-1)$ lorsque $N \geqslant 3$. En outre, cette formule se réduit pour $N \leqslant d+1$ à

$$C(N,d) = 2\sum_{k=0}^{N-1} \binom{N-1}{k} = 2(1+1)^{N-1} = 2^N$$

d'après la formule du binôme. Ceci est cohérent, heureusement, avec ce que l'on a observé dans le tableau (Figure 17).

La formule générale se démontre par récurrence sur N, en utilisant les «valeurs initales» $C(1,d)$ et $C(2,1)$ et les deux relations de récurrence que l'on connaît : celle sur les $C(N,d)$ et celle sur les $\binom{N}{d}$.

Si l'on revient à notre question de départ, la probabilité qu'une distribution de maillots à N joueuses dans un espace de dimension d soit séparée par un hyperplan est égale à $p = P(N,d)$ donnée par

$$P(N,d) = \frac{C(N,d)}{2^N},$$

où l'on sait exprimer $C(N,d)$, donc aussi $P(N,d)$ en fonction de N et d à l'aide de coefficients binomiaux.

La formule cachée

$$P(N,d) = \frac{1}{2^{N-1}} \sum_{k=0}^{d} \binom{N-1}{k}.$$

On peut aussi lire les valeurs de $C(N,d)$ en complétant la figure 17 dans un tableur. On peut par exemple en extraire

$$P(10,3) = \frac{260}{1024} \simeq 0,254, \quad P(10,4) = \frac{512}{1024} = \frac{1}{2}.$$

Tiens tiens, une chance sur deux de séparer 10 données en dimension 4 : est-ce un «hasard» ? Non, nous allons en reparler ! On peut aussi extraire

$$P(100,3) = \frac{323\ 600}{1024^{10}} < \frac{400\ 000}{10^{30}} = \frac{4}{10^{25}} = 0,0000000000000000000000004,$$

$$P(100,4) = \frac{7\ 852\ 352}{1024^{10}} < \frac{8\ 000\ 000}{10^{30}} = \frac{8}{10^{24}} = 0,000000000000000000000008.$$

Augmenter la dimension d permet d'augmenter la probabilité $P(N,d)$. On a même vu que $P(N,d) = 1$ pour $N \leqslant d+1$. La question qui va nous occuper maintenant est d'estimer $P(N,d)$ lorsque N et d sont grands avec $N \geqslant d+2$.

7 Bénédiction ou malédiction de la dimension ?

On peut faire une première observation qui commence avec d et N encore petits. Dans le cas particulier $d = 1$ et $N = 4$, on n'a pas besoin d'utiliser la formule générale pour s'apercevoir, en dessinant tous les cas possibles, qu'il y a exactement la moitié de distributions qui sont séparables et la moitié qui ne le sont pas.

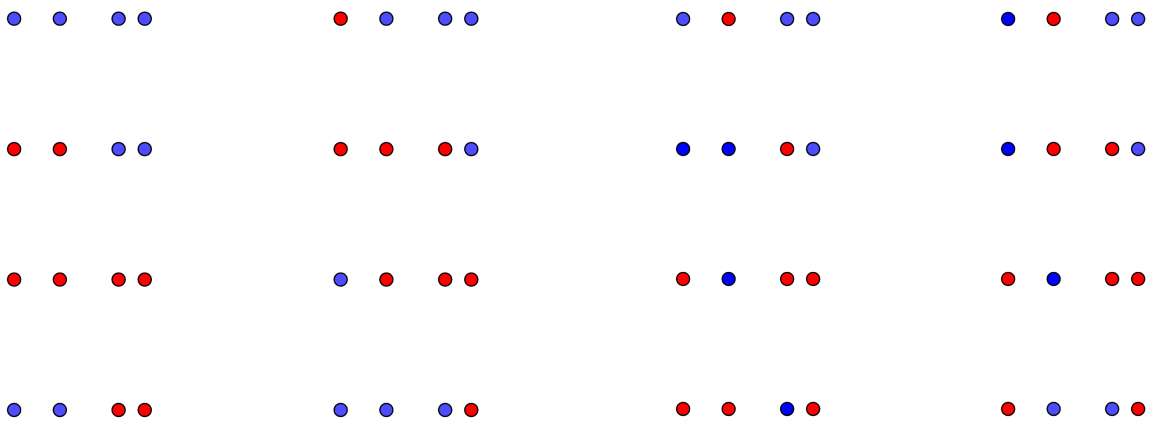

FIGURE 19 – Distributions de maillots en dimension un entre quatre joueuses : les huit séparables à gauche, les huit non séparables à droite.

Ce fait n'est pas un hasard, il est général pour $N = 2d+2$! On peut en effet voir par le calcul que

$$P(2d+2,d) = \frac{1}{2}.$$

Nous l'avons d'ailleurs déjà observé pour $d = 4$ et $N = 2 \times 4 + 2 = 10$.

Calcul de $P(2d+2,d)$. De par leur définition, et comme cela se lit dans le triangle de Pascal (Figure 18), les coefficients binomiaux sont symétriques :

$$\binom{2d+1}{k} = \binom{2d+1}{2d+1-k} \quad \text{pour } 0 \leqslant k \leqslant 2d+1\,.$$

Par suite

$$\sum_{k=0}^{d} \binom{2d+1}{k} = \frac{1}{2} \sum_{k=0}^{2d+1} \binom{2d+1}{k}\,.$$

Or on a par la formule du binôme

$$\sum_{k=0}^{2d+1} \binom{2d+1}{k} = (1+1)^{2d+1} = 2^{2d+1}\,.$$

Comme d'après la formule générale

$$P(2p+2,d) = \frac{1}{2^{2d+1}} \sum_{k=0}^{d} \binom{2d+1}{k}\,,$$

on en déduit,

$$P(2d+2,d) = \frac{1}{2}\,.$$

Cette observation est le témoin le plus criant d'un phénomène mis en évidence par Cover en 1965. À savoir que, pour un grand nombre de données N, la dimension $d = N/2 - 1$ est une valeur critique pour la capacité de l'espace à permettre de séparer ces données à coup d'hyperplans. S'agissant de grandes valeurs de N et de d, la présence de -1 dans la valeur critique de d ne joue pas de rôle et on peut très bien l'oublier. Ce qui compte c'est de savoir comment $2d/N$ se compare à 1. Pour fixer les idées, on va énoncer le résultat de Cover avec des nombres explicitement proches de 1, en l'occurrence 1.0001 et 0.9999, mais on pourrait choisir n'importe quels nombres de part et d'autre de 1.

Théorème 2 (Cover)**.** *Lorsque N et d tendent vers l'infini, la probabilité $P(N,d)$ qu'une distribution au hasard de maillots bleus ou rouges à N joueuses dans un espace de dimension d soit séparée par un hyperplan vérifie :*

- *si N et d tendent vers l'infini avec $2d/N \geqslant 1.0001$, alors $P(N,d)$ tend vers 1 ;*
- *si N et d tendent vers l'infini avec $2d/N \leqslant 0.9999$, alors $P(N,d)$ tend vers 0.*

La démonstration de ce résultat repose sur le théorème de Moivre–Laplace que l'on apprend au lycée. On ne la détaillera pas ici car elle est un peu technique, mais elle est accessible à des étudiants et étudiantes de première année post-bac : elle est expliquée ici dans le mini-cours [1].

Au vu de ce résultat, le fait que la dimension de l'espace soit suffisamment grande peut être vue comme une bénédiction en science des données et plus particulièrement en apprentissage supervisé. Car on sait qu'on a de grandes chances, avec une probabilité de plus en plus proche de 1 lorsque N et d augmentent ensemble de telle sorte que $2d/N \geqslant 1.0001$, de pouvoir séparer N données en dimension d (pourvu qu'elles soient en position générale, ce qu'on ne cherche pas à discuter ici).

Du côté du verre à moitié vide, on sait aussi que les chances de les séparer si $2d/N \leqslant 0.9999$ sont très faibles. On l'a vu sur nos premiers exemples même quand N est «seulement» de l'ordre de la dizaine. Il est donc bon d'avoir en tête cette valeur critique de $d = N/2$.

Cependant, on ne maîtrise pas forcément les valeurs de N et d dans les applications. Une façon d'augmenter d est d'avoir plus d'informations sur l'échantillon considéré. Par exemple, si en plus de l'âge et du nombre d'amis des individus de la population étudiée on connaît la pointure, le salaire et plein d'autres informations quantitatives encore, on augmente la dimension de l'espace des données. Ce n'est pas toujours possible. Sur les données médicales par exemple, il arrive que la population étudiée soit de petite taille et qu'on ait beaucoup d'informations sur chaque individu : c'est le cas pour des affections rares suivies de près par des équipes spécialisées. Mais il arrive aussi qu'on ait des informations limitées sur une très grande population, comme dans le cas du Covid-19 avec les tests effectués sur des milliards de personnes.

Une autre façon d'augmenter la dimension, plus théorique, consiste à se rappeler qu'on peut essayer de séparer les données non seulement par des hyperplans mais aussi par des frontières plus compliquées, comme on l'a fait avec nos lentilles et nos pâtes au début !

Cette idée est aussi dans l'article de Cover. Il s'agit d'envoyer les données dans un espace de dimension plus grande où les frontières compliquées de l'espace départ deviennent des hyperplans. Pour en savoir plus, on pourra lire l'article original, ou les explications données dans [1].

FIGURE 20 – The Magic Circle par John William Waterhouse

Pour terminer, signalons que la science des données, ainsi d'ailleurs que d'autres domaines des mathématiques, sont confrontés à ce qui est appelé la malédiction de la dimension et cherchent plutôt à la réduire qu'à l'augmenter. Mais c'est un autre sujet, sur lequel on pourra consulter par exemple [3] [5], ne serait-ce que pour ses belles images.

Références

[1] S. Benzoni-Gavage. Les bases mathématiques du perceptron. 2022.

[2] T. M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14(3) :326–334, 1965.

[3] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313.5786(DOI : 10.1126/science.1127647) :504–507, 2006.

[4] Mikel Olazaran. A sociological study of the official history of the perceptrons controversy. *Social Studies of Science*, 26(3) :611–659, 1996.

[5] Frank Rosenblatt. The perceptron : A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6) :386–408, 1958.

5. Disponible ici : https://www.cs.toronto.edu/ hinton/science.pdf