Distance

Critère discret

Problème

Soit un ensemble d'étudiants, ayant à choisir parmis une liste les activitées qui les attire.
On souhaite trouver la distance entre chaque étudiants.

Formulation mathématique

Soit un vecteurs $V = [v_0, ..., v_n]$.
On a $\forall i \in [0, n] : v_i \in \{0, 1\}$ tq $v_i = 0$ si l'activité n'est pas selectionné et $v_i = 1$ sinon.
On peut alors représenter 2 étudiants par deux vecteurs V, W.
Il nous faut alors trouver une distance entre ces deux vecteurs.

Simplification du problème

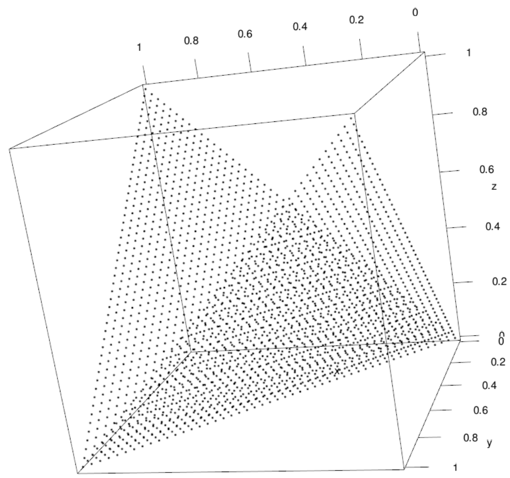
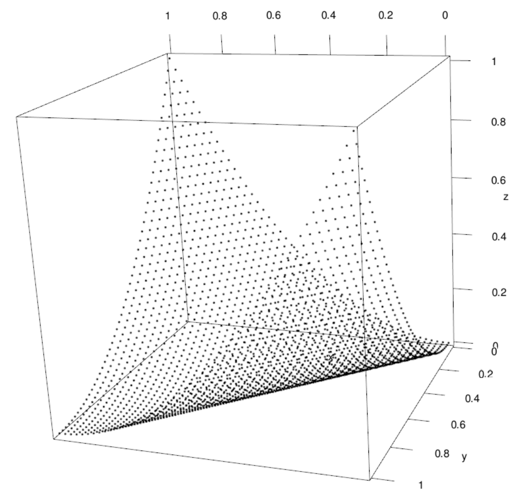
On suppose chaque critères indépendant. On peut ainsi traiter chaque critères (ie chaque couple v_i, w_i) indépendamment.
Il suffit alors de trouver une fonction $D : \mathbb{R}^2 \to \mathbb{R}$ qui nous donne la distance entre v_i et w_i puis de sommer $\forall i \in [0, n]$.

Dans la suite on notera $\forall (V = [v_0, ..., v_n], W = [w_0, ..., w_n]) \in (\mathbb{R}^n)^2$:

$$D(V, W) = \sum_{i=0}^{n} D(v_i, w_i)$$

Solution classique

Une implémentation classique de ce problème serait de prendre la distance de Manhatan ou la distance euclidienne.

Distance de Manhatan	Distance Euclidienne
$\sum_{i=0}^{n} \|v_i - w_i\|$ 	$\sum_{i=0}^{n} (v_i - w_i)^2$ 

Ces deux distances seront dans notre cas les même puisque l'on travaille dans l'ensemble $\{\{0,0\},\{0,1\},\{1,0\},\{1,1\}\}$. Prenons donc la distance de Marshal (notée D pour la suite).

Soit 3 étudiants ayant fait les choix représentés par les vecteurs suivants:

- $E_1 = [1,0,0]$
- $E_2 = [1,1,0]$
- $E_3 = [1,1,1]$

On rappelle que cela signifie que :

- l'étudiant E_1 est intéressé par le critère 1.
- l'étudiant E_2 est intéressé par le critère 1 et 2.
- l'étudiant E_3 est intéressé par le critère 1, 2 et 3.

On s'intéresse alors a la distance entre les élèves E_1 et E_2 et la distance entre les élèves E_1 et E_3.
On obtient: $D(E_1, E_2) = 1$ et $D(E_1, E_3) = 1$
On a alors $D(E_1, E_2) < D(E_1, E_3)$ ce qui est logique puisque E_1 et E_2 ont 2 critères en commun alors que E_1 et E_3 en ont 1 seul.

Cette approche est donc raisonnable, mais pas idéale. Nous allons expliquer pourquoi dans le prochain paragraphe.

Problèmes posés par la solution classique

Bien qu'efficasse, les distances précédaments proposées ne sont pas idéales.

Reprenons notre exemple avec:

- $E_1 = [1,0,0]$
- $E_2 = [1,1,0]$
- $E_3 = [1,1,1]$

Et cette fois calculons la distance entre E_1 et E_2 et entre E_2 et E_3.

On obtient: $D(E_1, E_2) = 1$ et $D(E_2, E_3) = 1$. Donc $D(E_1, E_2) = D(E_2, E_3)$.

Cela peut sembler logique puisque chaque couple possède seulement un critère de différence.
Cependant si on pense au sens de ces critères, ce qui importe n'est pas tant que il en ai un maximum en commun que le fait qu'ils en ai un maximum **selectionné** en commun. C'est à dire un maximum de critère que les deux étudiants apprécis. Aussi, deux personnes s'apprecient pour les choses qu'elles partagent et pas pour l'infinité de non points commun qu'elles ont.

Ainsi il nous faut prendre en compte cette dissymétrie qui n'est pas représentée par les distances classiques *par definition*.
En effet on rappelle que pour D une distance on aura $D(a, a) = 0$.

Dans le paragraphe suivant nous tacherons donc de construire une distance entre deux personnes qui ne sera pas une distance suivant la définition classique mais qui répondra mieux à notre problème.

Solution au problème de symétrie

Nous aurions pu limité le problème a notre exemple et contruire une fonction discrete (puisque nous travaillons dans l'ensemble $\{\{0, 0\}, \{0, 1\}, \{1, 0\}, \{1, 1\}\}$).
Cependant dans un soucis de généralité nous proposons une foncion continue.

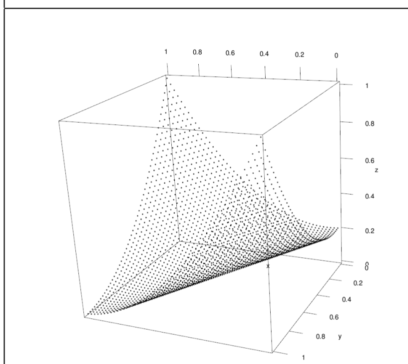
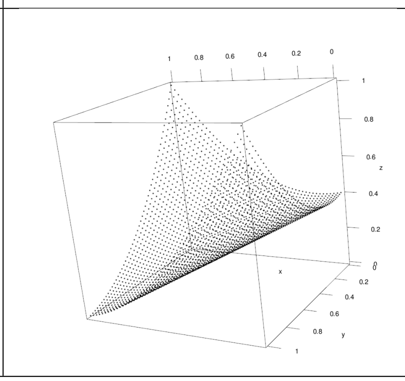
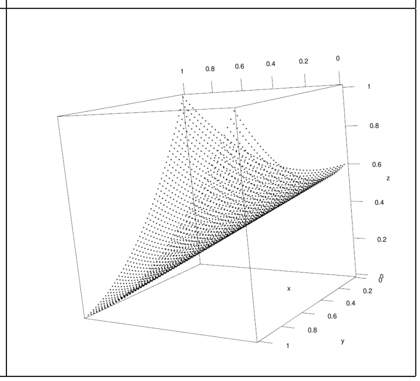
Les propriétés de cette fonction (notée D_P) seront:

- Continuité
- Dérivabilité
- $D_P(1, 1) < D_P(0, 0) < D_P(1, 0)$ et $D_P(1, 0) = D_P(0, 1)$

On définit un paramètre P qui sera la pénalité appliquée au point $(0, 0)$.
On définit alors:

$$\begin{array}{rccl} D_P : & \mathbb{R}^2 & \longrightarrow & \mathbb{R} \\ & (x, y) & \longmapsto & P * \left[1 - 0.5(x + y) + \frac{1}{P - 0.5} * (x - y)^2\right] \end{array}$$

Différents graphiques représentants D_P en faisant varier P.

P=0.2	P=0.4	P=0.6
		

Les propriétés de continuité et dérivabilité sont vérifiées de manière évidente.

De plus on a:

- $D_P(0,0) = P$
- $D_P(1,1) = 0$
- $D_P(1,0) = D_P(0,1) = 1$

Il suffit donc de prendre $P \in]0,1[$ et on a bien $D_P(1,1) < D_P(0,0) < D_P(1,0)$ et $D_P(1,0) = D_P(0,1)$.

Dans la suite on prendra donc $P \in]0,1[$.

Mise en oeuvre de la solution sur notre problème

Reprenons nos trois étudiants:

- $E_1 = [1,0,0]$
- $E_2 = [1,1,0]$
- $E_3 = [1,1,1]$

Appliquons maintenant la nouvelle fonction D_P au problème:

- $D_P(E_1, E_2) = 1 + P$
- $D_P(E_2, E_3) = 1$

Ainsi on a bien $D_P(E_2, E_3) < D_P(E_1, E_2)$.

La fonction D_P répond bien a notre problème.