Parallel Computing for Data Science
Lab x001 : Warm-up

Jairo Cugliari

S1 2016–2017

1 Calcul flotant

Donnez le résultat que vous attendez de commandes R qui suivent
- `is.integer(2)`
- `if(sqrt(2) * sqrt(2) != 2) print("what ?!")`
- `if(0.1 + 0.2 == 0.3) print("result is ok")`
- `if(0.1 + 0.2 != 0.3) print("no way !!!!")`

Ne continuez pas à travailler sur `R` avant de vous assurer que vous comprenez ce qui se passe dans cet exercice.

2 Optimisation

Exercice extrait du cours de A. Phillipe.

1. Construire une fonction qui calcule les valeurs de la fonction f définie par
$$f(x)=\sin(x)^2+\sqrt{|x-3|}.$$
2. Tracer la courbe représentative de la fonction f sur le domaine $[-6,4]$.
3. Donner une valeur approché de l'intégrale de la fonction f sur $[-6,4]$.
4. Donner une valeur approché du minimum de f sur $[-6,4]$. En quel point le minimum est il atteint ? (Astuce : regarder la fonction `optimise`)
5. Même question pour le maximum.

3 Problème

Nous voulons évaluer quelques procédures d'optimisation sur un tâche de datamining : résumer les n données univariées $y_1,y_2,...,y_n$ dans une seule valeur $\hat{y}$.

Nous appelons s à une candidate de $\hat{y}$, la meilleure valeur possible.

Ensuite, nous définissons une famille de fonctions de perte indexées par le paramètre p que nous supposons fini[1]
$$\text{loss}_p(s,y_1,y_2,...,y_n)=\left(\sum_{i=1}^{n}(s-y_i)^p\right)^{1/p}.$$

La famille contient la distance euclidienne ($p=2$) et la distance Manhantan ($p=1$) comme de cas particuliers.

1. Écrire la fonction `simuData(n)` qui simule un ensemble de données de taille n.
2. Écrire la fonction `perte(s, y, p)` qui calcule la distance de Minkowski de paramètre p entre la valeur s et le vecteur de données y.
3. Pour chaque valeur $p=1,2,5,1/2$, obtenir la valeur $\hat{y}=\text{argmin}_s\text{perte}(s,y,p)$ par optimisation numérique à l'aide de la fonction `optimize`. Ainsi, la valeur $\hat{y}$ est la valeur qui rend la plus petite perte de représentation des données y par une statistique s.
4. Représenter de manière graphique la fonction de perte ainsi que la valeur optimale.
5. Obtenir la solution du problème de manière analytique pour les valeurs de $p=1,2$.
6. Rajouter au graphiques correspondantes les valeurs obtenues.
7. Mesurer l'erreur de calcul.

[1] Il est possible de définir $\text{loss}_\infty(s,y_1,y_2,...,y_n)$ avec la distance du suprême mais nous ne traiterons pas ce cas ici.