LAMA-WeST Lab
Artificial Intelligence
Natural language processing
Semantic web

Corrélation de Spearman

INF8460 - Traitement automatique de la langue naturelle

Edouard Albert-Roulhac
Polytechnique Montréal

16 septembre 2024

Introduction à la corrélation de Spearman

La corrélation de Spearman est une mesure non paramétrique de la dépendance monotone entre deux variables. Contrairement à la corrélation de Pearson, qui mesure la relation linéaire, Spearman se base sur les rangs des données plutôt que sur les valeurs brutes.

- Utilisée pour évaluer des relations monotones
- Basée sur les rangs des données
- Plus robuste aux valeurs aberrantes

Exemple visuel

Considérons un jeu de données où la relation entre x et y est monotone mais non linéaire : $y = \exp(x)$

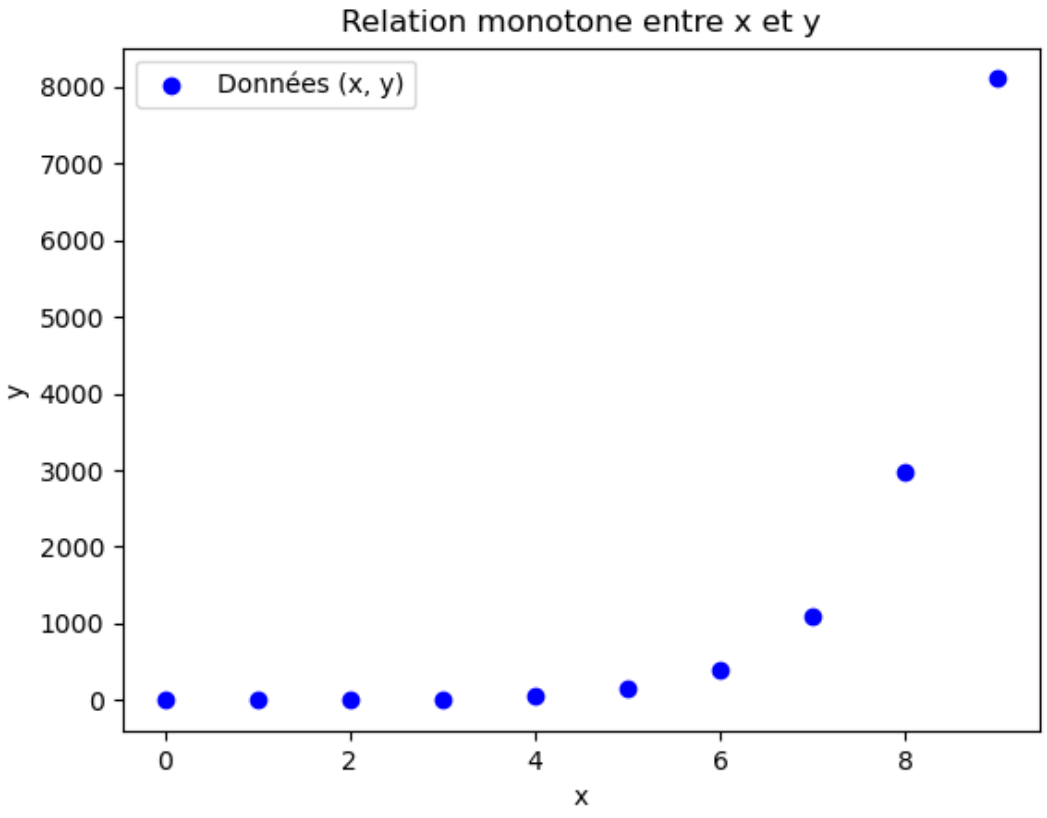

Figure – Relation monotone exponentielle entre x et y

La corrélation de Pearson, mesure de corrélation usuelle que vous connaissez déjà, évalue la relation linéaire entre deux variables. Elle est calculée par la formule suivante :

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Dans notre exemple, elle vaut :

$$\text{Corrélation de Pearson} = 0.72$$

- La valeur n'est pas 1, car la relation n'est pas linéaire.
- La corrélation de Pearson est sensible aux valeurs extrêmes et aux non-linéarités.

Tableau des rangs pour la corrélation de Spearman

Pour calculer la corrélation de Spearman, nous transformons les données en rangs. Les colonnes **Rangs** correspondent aux rangs des valeurs dans une liste comme dans le tableau :

x	**y**	**Rang(x)**	**Rang(y)**	**Diff Rangs**
0	1.00	1	1	0
1	2.72	2	2	0
2	7.39	3	3	0
3	20.09	4	4	0
4	54.60	5	5	0
5	148.41	6	6	0
6	403.43	7	7	0
7	1096.63	8	8	0
8	2980.96	9	9	0
9	8103.08	10	10	0

La corrélation de Spearman entre deux listes x et y de taille n est calculée à partir des rangs des données avec la formule suivante :

$$\rho(x, y) = 1 - \frac{6 \sum_{i=1}^{n} (\text{Rang(x)}_i - \text{Rang(y)}_i)^2}{n(n^2 - 1)}$$

- Dans l'exemple, puisque les rangs de x et y sont parfaitement alignés, la corrélation de Spearman est 1.
- Spearman est parfait pour détecter des relations monotones, même non linéaires.
- La corrélation de Spearman varie entre -1 (pour une relation strictement décroissante) et 1 (pour une relation strictement croissante).
- 0 indique qu'il n'y a pas de relation monotone entre x et y

Exemple avec Simlex (1)

Dans le cadre du TP2, le jeu de validation est Simlex. Voici un petit extrait de Simlex ainsi qu'un exemple de similarité cosinus dans la colonne **Pred**.
La colonne **Diff Rangs** est la différence en valeur absolue de rangs des valeurs dans les colonnes **Simlex** et **Pred**.

mot_1	**mot_2**	**Simlex**	**Pred**	**Rang Sim-lex**	**Rang Pred**	**Diff Rangs**
weird	odd	9.77	0.75	5	4	1
bad	immoral	7.77	0.64	3	2	1
sad	funny	0	0.56	1	1	0
wonderful	great	9.54	0.87	4	5	1
guilty	ashamed	7.31	0.67	2	3	1

Exemple avec Simlex (2)

Corrélation de Spearman : 0.8
La valeur de la corrélation est haute même si les valeurs et distributions sont différentes et si la relation n'est pas linéaire, car l'ordre des prédictions est presque conservé.