IFT1169 – TRAVAIL PRATIQUE #2 – 1er novembre 2022

« Science des données »
2e partie
Régression linéaire

Mohamed Lokbani

Équipes : le travail est à faire en monôme (une seule personne).
Remise : une seule remise est à effectuer par voie électronique le mardi 22 novembre 2022, 23h59 au plus tard, sans possibilités de prorogation.
Conseils : n'attendez pas le dernier jour avant la remise pour engager votre travail. Vous n'aurez pas le temps nécessaire pour le réaliser.

But : ce TP a pour but de vous faire pratiquer les STL.

Ce travail est la deuxième partie d'un projet de session à réaliser en 3 étapes. Vu que les étapes se suivent, il est donc important de réaliser chacune d'elle avant de pouvoir passer à la suivante.

Énoncé : dans la première partie, vous avez travaillé sur la lecture des données. Vous les avez préservées dans une perspective orientée objet. Finalement, vous avez calculé quelques statistiques descriptives. Dans cette seconde partie, vous allez améliorer l'organisation des données en utilisant pour cela des conteneurs de la STL et vous allez chercher à calculer la relation entre deux variables.

Régression linéaire simple : elle cherche à établir une relation linéaire entre une variable de sortie par rapport à une autre variable.

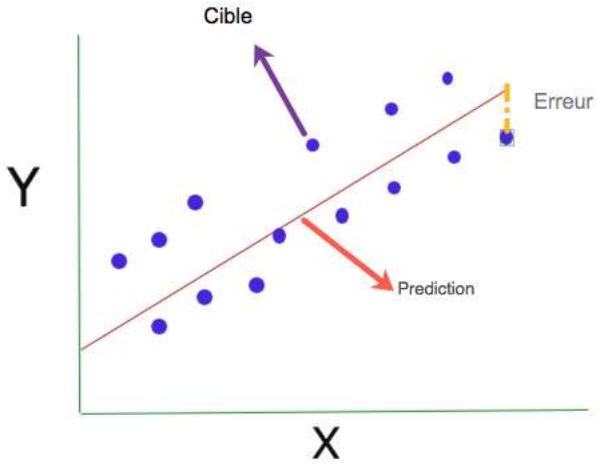
$\hat{y}(x) = a * x + b$

a représente la pente, b l'intersection et $\hat{y}$ est notre prédiction.

y représente la variable dépendante (elle dépend de x).

x est la variable indépendante.

On essaye donc de trouver la meilleure droite en fonction de l'ensemble des points x_i.



Calcul des paramètres

$$a = \frac{\sum_{i=1}^{N} x_i * y_i - \bar{y} * \sum_{i=1}^{N} x_i}{\sum_{i=1}^{N} x_i^2 - \bar{x} \sum_{i=1}^{N} x_i}$$

$$b = \frac{\bar{y} \sum_{i=1}^{N} x_i^2 - \bar{x} \sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2 - \bar{x} \sum_{i=1}^{N} x_i}$$

où

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

et

$$\bar{y} = \frac{\sum_{i=1}^{N} y_i}{N}$$

Qualité de prédiction : la qualité de prédiction mesure à quel point l'équation de régression est adaptée pour décrire la distribution des points. On calcule pour cela le coefficient de détermination linéaire de Pearson. Il est représenté par le symbole « R^2 » ou « r^2 ».

Si R^2 est proche de la valeur 0, la droite de régression colle à 0% avec l'ensemble des points donnés! Le modèle mathématique n'explique pas la distribution des points. Le nuage de points est complètement dispersé autour de la droite de régression.

Si R^2 est proche de la valeur 0.5, la moitié de la variation observée dans le modèle calculé peut être expliquée par les points.

Si R^2 est proche de la valeur 1, la régression détermine 100% de la distribution des points. Le nuage de points est resserré autour de la droite de régression.

En pratique, la valeur de R^2 est rarement égale à 1. Elle se situe généralement entre 0.85 et 1.

Elle est calculée comme suit :

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y_i})^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2}$$

Pour la suite de ce travail, on s'intéresse uniquement à l'effet de la masse corporelle (« bmi ») sur le diabète (« Y »). De ce fait,les valeurs des paramètres de cet exemple sont :

a: 10.23
b: -117.77
R^2 : 0.34

Travail demandé

- Lire le fichier de données (TP#01)
- Organiser les données dans une approche orientée objet (TP#01)
- Maximiser l'utilisation de la STL (conteneurs et algorithmes).
- Calculer les paramètres « a » et « b » de la droite de régression linéaire entre « bmi » et « Y ».
- Calculer le coefficient de détermination R^2.
- Réaliser une prédiction. On calcule $\hat{y}$ à partir d'une valeur fournie en entrée.

Affichage: nous allons ajouter au menu l'affichage des paramètres de la droite de régression, le coefficient de détermination et la possibilité d'effectuer une prédiction.

Vous allez afficher le menu suivant :

Sélectionner un nombre de la liste suivante :

[1] Description de l'ensemble de données
[2] L'entête des données
[3] Queue des données
[4] Propriétés des données
[5] L'ensemble des données
[6] Coefficients de la droite de régression
[7] Coefficient de détermination
[8] Prédiction
[9] Quitter le programme

Chacun des affichages demandés est fourni sur le site web du travail pratique. Il vous est demandé d'obtenir exactement le même affichage. Vous allez utiliser pour cela les propriétés de l'affichage nouvellement introduites dans le standard C++20. Ces propriétés ont été décrites dans la séance de démonstration #02.

Aide : compléter le programme avec une option d'aide.

tp2.exe --help ou tp2.exe -h

Ces deux options font la même chose. Elles affichent une courte description sur la manière d'utiliser votre programme.

Fichiers à remettre :

Vous devez regrouper l'ensemble de ces fichiers dans « TP2_A22.zip » et le remettre par l'entremise du système Studium, dans le dépôt « tp02 ».

Le fichier « TP2_A22.zip » va contenir :

L'ensemble des programmes (*.h/*.cpp) relatifs à ce travail pratique.

Le makefile permettant de compiler simplement ce travail pratique.

« TP2_A22 _Rapport.pdf » : le rapport décrivant le travail effectué. Vous allez trouver sur le site web du cours, dans la section « Foire aux questions », les détails qui doivent figurer dans un rapport.

« TP2_A22 _feuille2route.pdf » : une auto-évaluation du travail réalisé.

La correction va se faire sur un poste de la DESI en utilisant le compilateur « g++ » (12.x). Vérifiez que le code que vous avez produit (devant normalement fonctionner correctement chez vous) fonctionne aussi bien sur les ordinateurs de la DESI.

Barème : ce TP2 est noté sur **15 points**. Les points sont répartis comme suit :

Affichage : 3 points (affichage en sortie, aide, gestion de l'erreur)
Utilisation de la STL : 5 points
Calcul des paramètres (a, b et R^2) : 4 points
Rapport : 1.5
Avis global : 1.5

En plus du précédent barème, vous risquez de perdre des points dans les cas suivants ....

- La non-remise électronique (volontaire ou par erreur) est sanctionnée par la note 0.
- La non-remise du fichier « TP1_A22_feuille2route.pdf » : -1.
- Un programme qui ne compile pas : 0.
- Un programme qui compile, mais ne réalise pas les choses prévues dans la spécification : 0.
- Les avertissements (warnings) non corrigés : cela dépend de la quantité! À partir de -0.25 et plus.
- Aberration dans le codage : même si tous les chemins mènent à Rome, faites l'effort nécessaire pour éviter de prendre le plus long!

Des questions à propos de ce TP? Une seule adresse : dift1169@iro.umontreal.ca

Pour faciliter le traitement de votre requête, inclure dans le sujet de votre courriel, au moins la chaîne [IFT1169] et une référence au tp02.

Mise à jour
1-11-2022, diffusion de l'énoncé.

Annexe

En apprentissage automatique, la régression est utilisée quand on veut estimer une valeur de sortie à partir d'un ensemble de valeurs en entrées. On calcule pour cela une fonction mathématique. Dans l'exemple de la régression linéaire simple avec une seule caractéristique, la fonction mathématique est définie ainsi :

$$\hat{y}(x) = a * x + b$$

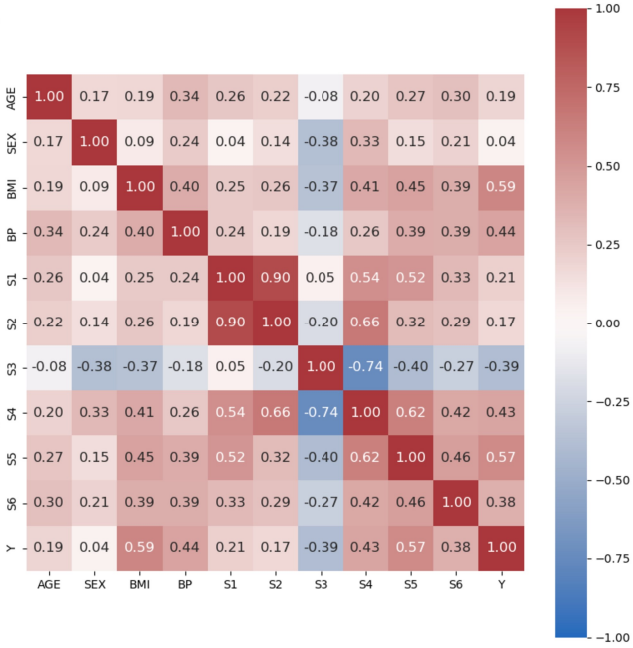
Le jeu de données contient 10 caractéristiques. Doit-on les prendre toutes en considération dans le calcul de la fonction de régression? On peut réaliser une étude préalable pour déterminer les caractéristiques les plus influentes. Une de ces études consiste à chercher s'il y a une corrélation entre les données.

Corrélation des données

Le coefficient de corrélation mesure la force de la relation entre deux variables. Sa valeur se situe dans l'intervalle [-1,+1].

> Si le coefficient est proche de 1, nous sommes en présence d'une forte corrélation positive.
> Si le coefficient est proche de -1, nous sommes en présence d'une forte corrélation négative.
> Si le coefficient est proche de 0 en valeur absolue, il y a une faible corrélation.

Nous avons calculé pour vous ce coefficient pour l'ensemble des données. Le résultat est présenté sur la figure.



Le choix de la caractéristique « bmi » n'est pas le fruit du hasard. On constate qu'il y a une corrélation moyenne entre « bmi » et la cible « Y ». La valeur est de « 0.59 ».

En prenant en considération juste le « bmi », la valeur du coefficient de détermination R^2 est égale à « 0.34 ». La valeur est faible. Ça s'explique par le fait que nous n'avons utilisé qu'une seule caractéristique.

En pratique, on calcule une régression plus complexe. Ce calcul tient compte des caractéristiques qui ont une influence statistique.