La régression linéaire

Sommaire

I. Introduction Générale

Dans les cours qui vont suivre on va parler de modèles mathématiques destinés à prédire un évènement réel à partir d'observations photographiées suite à l'avènement d'un évènement exactement ou quasi similaire à celui à prédire.

Explications :

- On ne peut prédire la réalisation d'un évènement que si il a déjà été observé auparavant (plus il a été observé de fois,mieux c'est)
- Les observations sont photographiés : elles sont figées selon toutes leurs dimensions (temps, espaces, ...).

Donc de ces deux points, on devine que la prédiction est fortement dépendante des observations faites, fortement dépendantes de leurs qualités, de la diversifications des points de vues sur chaque observation, ... Et on devine aussi que la prédiction est sujette à erreur, car elle est la tentative de donner une valeur fixe à une variable en mouvement à partir d'observations figées.

Il est donc nécessaire à chacun qui décide de s'adonner à la prédiction de respecter les axes suivants :

- La qualité des observations
- La multiplication des observations.
- L'estimation de l'erreur qui peut être commise.

II. La prédiction

Revenons-en aux mathématiques...

Dans l'introduction, le mot prédiction à été utilisé plusieurs fois, mais quel en est le sens mathématique?

Prenons un exemple simple pour illustrer son sens :
Nous cherchons à prédire une variable Y à partir d'une variable X. Nous disposons donc d'une suite de n observations $(x_i, y_i)_{i\in[1,n]}$. Prédire la variable Y à partir de X revient à trouver une fonction f:

$$Y \approx f(X)$$

Ainsi, une fois la fonction f trouvée, pour une valeur x_j, on peut estimer la valeur de y_j correspondante, sans pour autant attendre l'observation de l'évènement x_j.

Comme énoncé dans l'introduction, une prédiction est sujette à erreurs (d'où le signe $\approx$. Donc pour avoir une mesure de la qualité de la prédiction (la qualité du choix de la fonction f), nous nous proposons de calculer la grandeur suivante pour chaque fonction f:

$$\arg\min_{f\in F} \sum_{i=1}^{n} L(y_i - f(x_i))$$

Où n est le nombre d'observations, F l'ensemble des fonctions auquel appartient f et L une fonction de coût (exemples : $x \to |x|$, $x \to x^2$,...).

III. La régression linéaire simple

Pour commencer ce cours, nous allons étudier le cas le plus simple. Nous disposons donc d'une variable à expliquer Y à partir d'une unique variable X et nous supposons que notre fonction de prédiction f est affine.

Nous avons donc n observations (x_i, y_i) et nous supposons qu'il existe a,b vérifiant :

$$y_i = f(x_i) + \epsilon_i = ax_i + b + \epsilon_i \; \forall i \in \{1, .., n\}$$

Les ϵ_i sont les erreurs qui décrivent le fait que les n points ne sont jamais alignés sur la droite d'équation y=ax+b.

Afin de conserver la logique et la pertinence de ce modèle, nous faisons les 3 hypothèses suivantes :

$$(\mathcal{H}) \begin{cases} E(\epsilon_i) = 0 \; \forall \; i \in \{1, .., n\} \\ Var(\epsilon_i) = \sigma^2 \; \forall \; i \in \{1, .., n\} \\ Cov(\epsilon_i, \epsilon_j) = 0 \; si \; i \neq j \end{cases}$$

1. Méthode des moindres carrés

Nous cherchons donc à trouver la fonction affine f qui minimise la quantité :

$$\arg\min_{f \in F} \sum_{i=1}^{n} L(y_i - f(x_i))$$

Pour celà nous choisirons la fonction de coût $L: x \to x^2$.

On parle alors de méthode d'estimation par moindres carrés (MCO).

$$S(a, b) = \sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2 = \sum_{i=1}^{n} (y_i - b - ax_i)^2$$

On cherche donc les coefficients $\hat{a}$ et $\hat{b}$ qui minimisent S(a,b).

2. Les équations régiassant $\hat{a}$ et $\hat{b}$

On montre sans grande difficulté que la fonction S(a,b) est strictement convexe ce qui nous permet d'affirmer l'existance d'un minimum en un unique point $(\hat{a}, \hat{b})$, qu'on peut déterminer en annulant les dérivées partielles de S :

$$\begin{cases} \dfrac{\partial S}{\partial a} = -2\sum_{i=1}^{n}(y_i - \hat{b} - \hat{a}x_i) = 0 \\ \dfrac{\partial S}{\partial b} = -2\sum_{i=1}^{n} x_i(y_i - \hat{b} - \hat{a}x_i) = 0 \end{cases}$$

- La première équation donne :

$$\hat{b}n + \hat{a}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

Donc,

$$\hat{b} = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{\hat{a}}{n}\sum_{i=1}^{n} x_i$$

D'où le résultat,

$$\hat{b} = \bar{y} - \hat{a}\bar{x} \qquad (1.1)$$

Avec $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ et $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ (La moyenne arithmétique).

- La deuxième équation donne :

$$\hat{b}\sum_{i=1}^{n} x_i + \hat{a}\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i\, y_i$$

En remplacant $\hat{b}$ par son expression obtenue (1.1) on obtient:

$$\hat{a} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (1.2)$$

Cette relation (1.2) suppose que le dénominateur $\sum_{i=1}^{n}(x_i - \bar{x})^2$ est non nul. Or, $\sum_{i=1}^{n}(x_i - \bar{x})^2 = 0$ implique que $x_i - \bar{x} = 0 \; \forall \; i \; \Leftrightarrow x_i = \bar{x} \; \forall i$. Ce cas est naturellement écarté car il n'est pas intéressant (tout les x_i sont égaux donc le vecteur X n'ifluencerai pas les coefficient du vecteur Y)

Remarques :

- La relation $\hat{a} = \bar{y} - \hat{b}\bar{x} \; \left(\bar{y} = \hat{b}\bar{x} + \hat{a}\right)$ implique que la droite des MCO passe par le centre de gravité (x_i, y_i) du nuage des points.

- L'estimateur $\hat{a}$ peut aussi s'écrire :

$$\hat{a} = a + \frac{\sum_{i=1}^{n}(x_i - \bar{x})\epsilon_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

  Cette décomposition de $\hat{a}$ n'est pas interessante pour le calcul car elle fait intervenir la valeur de a (la valeur théorique de la pente de la droite affine y = f(x) que ne nous pouvons qu'estimer sa valeur à travers notre échantillon) et les valeurs de ϵ_i qui nous sont inconnues. Mais elle à un intérêt thérique qui sert à démontrer des propriétés importantes des estimateurs $\left(\hat{a}, \hat{b}\right)$ (leurs biais et leur variance), propriété qui justifie la convergence des résultats obtenue par la pratique (à travers les observations) vers le résultat théorique lorsque la quantité et la qualité des observations s'améliore.

3. Les propriétés statistiques des estimateurs

Sous les hypothèses régissant ϵ_i (les hypothèses $\mathcal{H}$), les estimateurs $\hat{a}$, $\hat{b}$ présentes les proprietés suivantes :

a. Les estimateurs sont sans biais

$$E[\hat{a}] = a \;\; et \; E\left[\hat{b}\right] = b$$

(la démonstration est décrite dans le document (1) en référence à la fin du cours)

Interprétation du résultat :

Si on considère n ensembles d'échantillons E_i, lorsqu'on calcule les $\hat{a}_1, \hat{a}_2, \dots, \hat{a}_n$ les estimateurs à partir des E_i respectifs, ces estimateurs ont comme espérance a. (de même pour $\hat{b}$)

b. Les estimateurs $\hat{a}, \hat{b}$ sont ceux avec une variance minimale

$$\forall \; \tilde{a} \; estimateur \; de \; a, Var(\tilde{a}) \geq Var(\hat{a})$$

$$\forall \; \tilde{b} \; estimateur \; de \; a, Var\left(\tilde{b}\right) \geq Var\left(\hat{b}\right)$$

(la démonstration est décrite dans le document (1) en référence à la fin du cours)

Interprétation du résultat :

L'estimateur $\hat{a}$ $\left(respectivement \; \hat{b}\right)$ sont les estimateurs des coefficients de f qui varient le moins autour de la valeur réelle théorique de a (respectivement de b).

4. La notion de résidu

En ayant déterminé les valeurs de nos estimateurs $\hat{a}$ de la pente de f et $\hat{b}$ de son ordonnée à l'origine, nous avons, pour chaque valeur x_i de l'échantillon deux points possibles :

Le point (x_i, y_i) : point observé et le point $(x_i, \hat{y}_i)$ point qui se trouve sur la droite affine définie par f. Les résidus sont définit par :

$$\hat{\mathcal{E}}_i = y_i - \hat{y}_i$$

Or, $\hat{y}_i = \hat{a}x_i + \hat{b}$:

$$\hat{\mathcal{E}}_i = y_i - \hat{y}_i = y_i - \hat{a}x_i - \hat{b}$$

D'après la relation (1.1) on a :

$$\hat{b} = \bar{y} - \hat{a}\bar{x}$$

D'où l'expression du résidu i :

$$\hat{\mathcal{E}}_i = y_i - \bar{y} - \hat{a}(x_i - \bar{x})$$

Par construction on a donc :

$$\sum_i \hat{\mathcal{E}}_i = \sum_i (y_i - \bar{y}) - \hat{a} \sum_i (x_i - \bar{x}) = 0$$

Par contre, la variance des résidus nous est inconnue car les résidus représentent en réalité les éléments qui influencent la variable Y et que le modèle n'a pas pris en compte. Mais nous pouvons construire un estimateur de cette variance :

$$\widehat{\sigma^2} = \sum_{i=1}^{n} \frac{\hat{\mathcal{E}}_i^2}{n-2}$$

Cette estimateur de la variance de $\mathcal{E}_i$ est un estimateur sans biais.

(la démonstration est décrite dans le document (1) en référence à la fin du cours)

Interprétation du résultat :

Estimer la variance des résidus est important car il permet de déterminer la qualité des observations. Des résidus avec une grande variance veut dire que notre modélisation est incomplète et qu'il existe une variable qui influence de manière considérable notre variable à prédire Y et qui n'a pas été prise en compte.

5. La prévision

Comme expliqué dans le début du cours, l'un des buts de la régression est la prédiction. Il s'agit de prédire la valeur y_{n+1} étant donné un valeur x_{n+1} qui n'a pas été observée dans notre échantillon.

En utilisant notre modélisation, on sait que :

$$y_{n+1} = ax_{n+1} + b + \epsilon_{n+1}$$

Avec ϵ_{n+1} l'erreur, et qui suit les hypothèses énoncé dans $(\mathcal{H})$. Or ne connaissant pas la valeur réelle de ϵ_{n+1}, nous allons donc utiliser le modèle ajusté par les estimateurs :

$$\hat{y}_{n+1} = \hat{a}x_{n+1} + \hat{b}$$

C'est au cours de cet ajustement que notre prévision risque l'erreur (importante ou non) car deux éléments sont incertain : ϵ_{n+1} est inconnu et les estimateurs $\hat{a}\ et\ \hat{b}$ sont une approximation des valeurs de a et b.

On cherche donc toujours à évaluer l'importance de l'erreur que nous pouvons commettre, erreur qui peut causer d'important dégâts selon son ampleur.

On obtient donc les deux résultats suivants :

$$\begin{cases} E[\hat{\epsilon}_{n+1}] = 0 \\ Var(\hat{\epsilon}_{n+1}) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) \end{cases}$$

(la démonstration est décrite dans le document (1) en référence à la fin du cours)

Interprétation du résultat :

La moyenne des écarts des prévisions par rapport à la valeur réelle est nulle (Donc plus nous faisons de prédictions plus nous auront de succès, il faut juste faire en sorte que le coût des échecs est couvert par ces succès).

La variance des écarts des prévisions par rapport à la valeur réelle dépend de n et de l'écart $x_{n+1} - \bar{x}$. Cette dépendance est comme suit :

- *Plus n est grand (plus la taille de notre echatillon est grande) plus la variance des résidus est petite*
- *Plus la valeur x_{n+1} s'éloigne de l'espérance $\bar{x}$, plus la variance des résidus est grande.*

Or la variance des résidus représente, en d'autre terme, le risque de s'éloigner de la valeur réelle à prédire. Donc, il s'agit de la minimiser et c'est dans ce sens qu'il faut éviter de faire des prédictions pour des points trop éloignés du nuage (les outliers), car c'est les points sur lequels l'impact des données manquantes est le plus important.

IV. Conclusion

Avant de foncer tête basse et de faire des prédictions à tort et à travers en appliquant des algorithmes disponibles un peu partout sur internet, il faut bien assimiler les différents aspects de l'outil utilisé.

La compréhension de l'outil permet une meilleure maîtrise et précision des prédictions qui en sortiront ainsi qu'une bonne préparation et anticipation face aux potentielles erreurs qui pourront être induites par toute les approximations faites en passant d'une étape à une autre.

De plus, même avec une bonne compréhension de l'outil, un algorithme avec de mauvaises données en entrées donnera de mauvaises prédictions en sortie.

Il faut donc savoir adapter les données de manière à minimiser l'impact des données manquantes ou mal exprimées.

N.B : Dans ce premier chapitre, nous avons visité les points les plus importants de la regression linéaire simple mais ce cours est loin d'être exhaustif.

Références

(1) http://www.lsta.upmc.fr/guyader/files/teaching/Regression.pdf

(2) http://www.math.sciences.univ-nantes.fr/~rochet/enseignement/cours_Reg(sans_preuve).pdf