Régression logistique

La régression logistique est utilisée pour la classification et non pour la régression. Cependant, elle est considérée comme une méthode de régression, car elle permet d'estimer la probabilité qu'une observation appartienne à une classe donnée. Il existe trois types de régression logistique :

- **Régression logistique binaire**: ici, le but de la classification est d'identifier si un échantillon appartient à une classe ou non.
- **Régression logistique multinomiale**: ici, le but de la classification est d'identifier à quelle classe appartient-t-il un échantillon parmi plusieurs classes.
- **Régression logistique ordinale**: ici, le but de la classification est de chercher la classe d'un échantillon parmi des classes ordonnées. Un exemple de classes: non satisfait, satisfait, très satisfait.

Régression logistique binaire

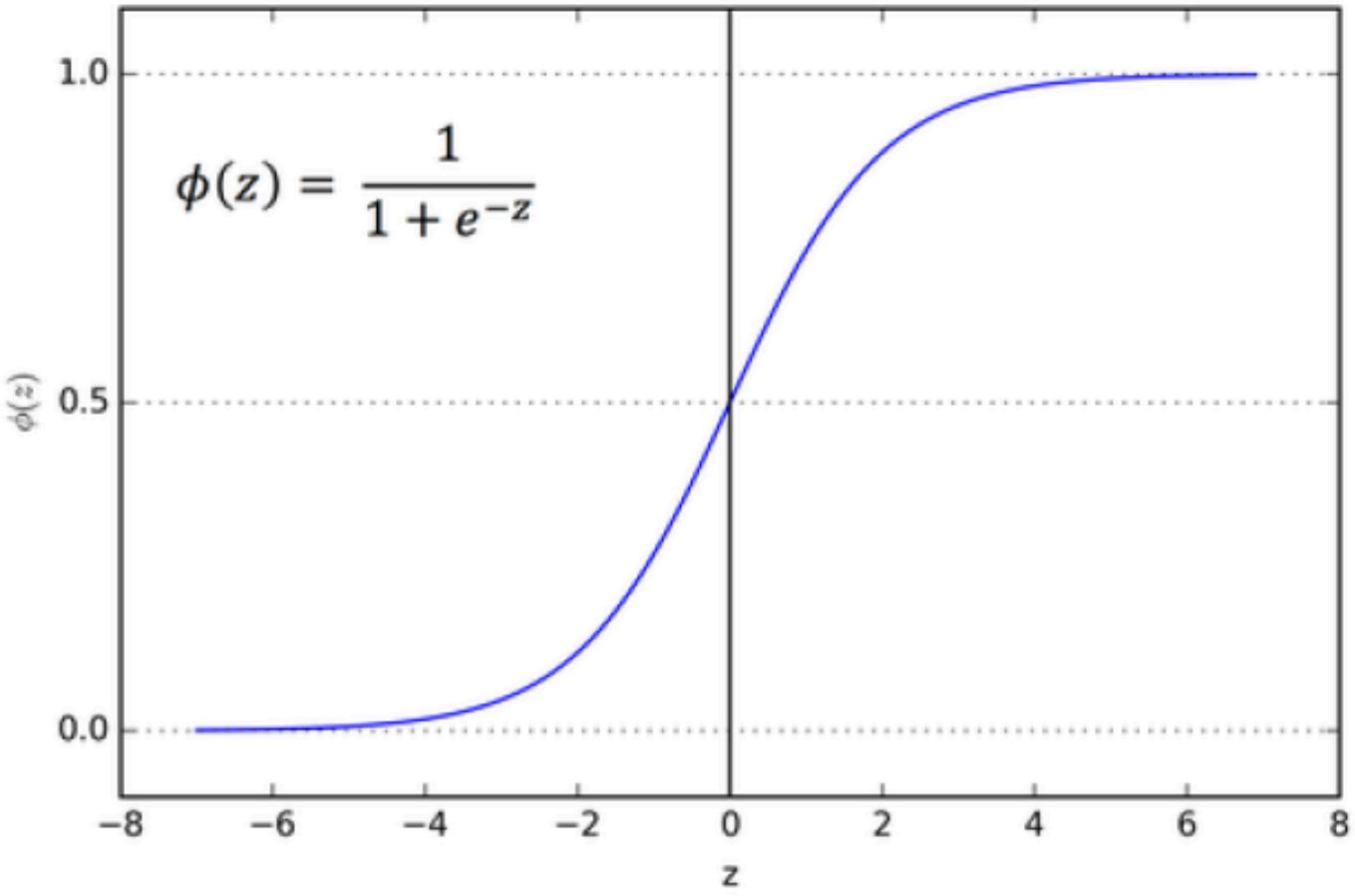
La régression logistique (**binaire**)est une méthode statistique et un algorithme de machine learning utilisé pour prédire des résultats binaires. Contrairement à la régression linéaire, qui prédit des valeurs continues, la régression logistique est conçue pour modéliser la probabilité qu'un événement se produise, généralement en réponse à des variables indépendantes.

Dans ce type de problème, nous disposons d'un dataset contenant une variable cible y qui peut prendre seulement deux valeurs, par exemple 0 ou 1 :

- Si y=0, cela signifie que l'email n'est pas un spam.
- Si y=1, cela indique que l'email est un spam.

Processus de la Régression Logistique:

- ▢ **Données et Variables** :
- Dans le contexte de la régression logistique, on a des variables indépendantes (features) qui sont souvent notées x1,x2,...,xn et une variable dépendante qui représente la classe à prédire (par exemple, "Spam" ou "Non Spam").
- ▢ **Combinaison Linéaire** :
- Avant d'appliquer la fonction logistique on calcule une combinaison linéaire des variables indépendantes
  $z=\beta 0+\beta 1x1+\beta 2x2+...+\beta nxn$
- Ici, z est un score qui n'est pas encore interprété en termes de probabilité.
- ▢ **Fonction sigmoide :**
- Pour obtenir une probabilité à partir de zzz, on applique la fonction sigmoïde :$P(Y=1 \mid X)=1/1+e-z1$



- Cette transformation convertit le score z en une probabilité entre 0 et 1.
- À partir de cette fonction, il est possible d'établir une frontière de décision. En général, un seuil de 0,5 est défini comme suit :

$$y=0 \text{ si } \sigma(X \cdot \theta) < 0$$

$$y=1 \text{ si } \sigma(X \cdot \theta) \geq 0$$

☐ **Frontière de Décision**

- La frontière de décision est déterminée par le seuil de 0,5 sur la probabilité. Cela signifie que si la probabilité prédite P(Y=1 | X) est supérieure ou égale à 0,5, on classe l'observation comme "présence" (par exemple, "Non Spam"), et si elle est inférieure à 0,5, on la classe comme "absence" (par exemple, "Spam").
- Dans un espace à deux dimensions (par exemple, x1 et x2), cette frontière est une ligne droite qui sépare les deux classes.

☐ **La Fonction du coût**

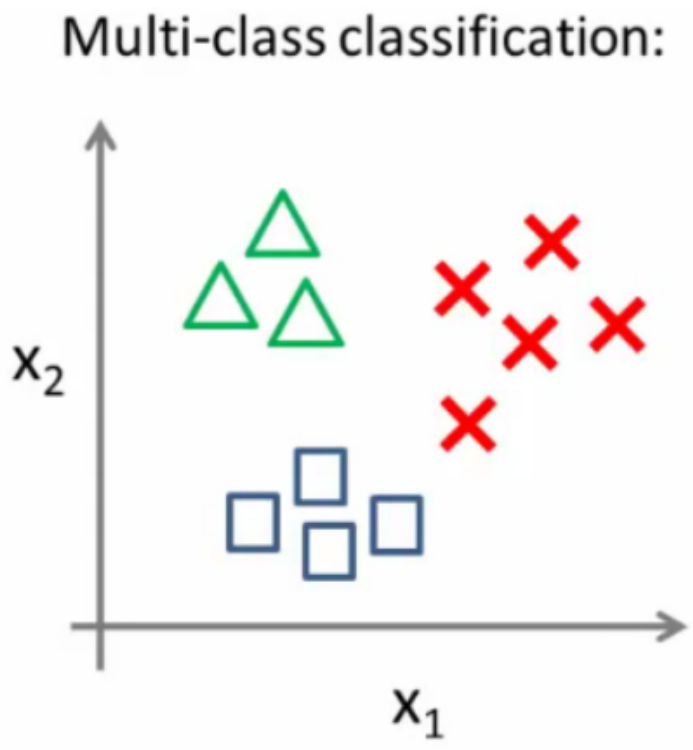
Pour la régression linéaire, la Fonction Coût $J(\theta) = \frac{1}{2m}\Sigma(X.\theta - Y)^2$ donnait une courbe convexe (qui présente un unique minima). C'est ce qui fait que l'algorithme de Gradient Descent fonctionne. En revanche, utiliser cette fonction pour le modèle Logistique ne

donnera pas de courbe convexe (dû à la non-linéarité) et l'algorithme de Gradient Descent se bloquera au premier minima rencontré, sans trouver le minimum global.

Adaptation de la régression logistique à la classification multi-classes:

Comme nous l'avons observé jusqu'à présent, la régression logistique ne permet de classer les observations que de manière binaire (Spam/Non Spam, Malin/Bénin, Noir/Blanc...), ce qui peut s'avérer limitant.

Imaginez maintenant que vous devez classifier une observation dans l'une de trois catégories. Par exemple, il s'agit de classer un article de presse dans l'une des trois catégories suivantes : Sport, High-Tech ou Politique. Dans ce cas, on parle de classification multi-classes, où l'étiquette $Y \in \{0, 1, 2\}$.



Pour la régression logistique multiclasse, on utilise la fonction softmax, qui généralise la sigmoïde à plusieurs classes. La fonction softmax softmax(zi)\text{softmax}(z_i)softmax(zi) pour une classe iii est définie comme **:**

softmax(zi) = e^(zi) / Σ (e^(zj))

où :

- zi est la sortie du modèle pour la classe i,
- K est le nombre total de classes.

La fonction softmax est une extension de la fonction sigmoïde, utilisée dans la régression logistique pour la classification multiclasse. Elle transforme un vecteur de scores en probabilités, dont la somme est égale à 1. Chaque probabilité indique la vraisemblance qu'une observation appartienne à l'une des K classes. En normalisant les scores, la fonction softmax permet de sélectionner la classe la plus probable pour chaque observation, facilitant ainsi le processus de classification.

Application de la fonction Softmax

Pour obtenir des probabilités à partir de ces scores, nous utilisons la fonction softmax. La fonction softmax transforme les logits en probabilités qui totalisent 1 (ou 100%).

Résultat

Après application de la fonction softmax, nous obtenons les probabilités suivantes :

- Probabilité de la classe A : 0.83 (83%)
- Probabilité de la classe B : 0.12 (12%)
- Probabilité de la classe C : 0.05 (5%)

Interprétation

Dans cet exemple, le modèle prédit que l'objet appartient à la classe A avec une probabilité de 83%. Ainsi, en utilisant softmax, nous avons converti les scores du modèle en probabilités qui permettent de classifier l'objet dans la classe ayant la probabilité la plus élevée.