Projet : Classification du cancer du sein

Objectifs

- Comprendre le principe de classification en utilisant la régression logistique
- Mettre en œuvre une recherche de minimum sans contrainte par la méthode du gradient par lot
- Mettre en œuvre une recherche de minimum sans contrainte par la méthode du gradient stochastique
- Analyser le comportement de ces méthodes

Objectif de la régression logistique et définition du modèle (Fonction logistique)

L'objectif de la régression logistique est de prédire la probabilité d'une variable de sortie binaire en fonction des caractéristiques d'entrée fournies. Par exemple, prédire la présence d'une maladie en fonction de certaines caractéristiques du patient.

La régression logistique utilise la fonction logistique pour transformer la combinaison linéaire des caractéristiques d'entrée en une valeur de probabilité comprise entre 0 et 1. La formule de la fonction logistique est la suivante :

$$\sigma(X.\theta) = \frac{1}{(1+e^{(-X.\theta)})}$$

avec θ(w0, w1, w2, ..., wn) sont les paramètres de régression (à prédire), et X(x1, x2, ..., xn) sont les caractéristiques d'entrée.

n est le nombre de caractéristiques (nombre de colonnes de X)

A partir de cette fonction, il est possible de définir une frontière de décision entre les données étiquetées 1 (cancer du sein) et les données étiquetées 0 (pas de cancer). le seuil est défini à 0.5 comme ceci :

$$\begin{cases} y = 0 \; si \; \sigma(X.\theta) \leq 0.5 \\ y = 1 \; si \; \sigma(X.\theta) \geq 0.5 \end{cases}$$

Fonction de coût

la fonction de coût pour la régression logistique est définie par :

$$J(\theta) = -\frac{1}{N}\sum_{i=1}^{N}(y^{i}_{true} log(y^{i}_{pred}) + (1 - y^{i}_{true}) log(1 - y^{i}_{pred}))$$

où : N est le nombre d'échantillons

y^{i}_{true} est l'étiquette réelle de l'échantillon i

$y^{i}_{pred} = \sigma(X.\theta)$ est la prédiction du modèle (fonction logistique) pour l'échantillon i

Classification du cancer du sein en utilisant la régression logistique

L'objectif du projet est de développer un modèle prédictif capable de distinguer de manière précise et fiable les tumeurs bénignes des tumeurs malignes. En utilisant les données prétraitées et normalisées fournies, vous devrez optimiser les paramètres θ du modèle afin de minimiser la fonction de coût.

Problème simplifié

Phase 1 :

Le modèle est définie par la fonction ci-dessous, dans laquelle il y'a 2 paramètres : θ(w0, w1).

$$\sigma(X.\theta) = \frac{1}{(1+e^{(-X.\theta)})}$$

avec X une matrice de 455 lignes (nombre d'exemples) et 2 colonnes (caractéristiques)
Les données à fitter sont un ensemble de points $\{(x_i,y_i)\}_{i=1,N}$. Le but du problème est de trouver les paramètres optimaux θ tels que le modèle sépare efficacement les deux classes (tumeurs malignes et bénignes) dans le nuage de points. Pour cela, il faut minimiser la fonction de coût ci-dessous.

$$J(\theta) = -\frac{1}{N}\sum_{i=1}^{N}(y^{i}_{true}\,log(y^{i}_{pred}) + (1 - y^{i}_{true})log(1 - y^{i}_{pred}))$$

Un fichier comportant les coordonnées de 455 points est fourni (phase1_2D), et il faut déterminer les coefficients θ(w0, w1) qui minimisent le critère de distance $J(\theta)$.

Deux méthodes seront implémentées et testées : une descente de gradient par lot, puis la descente de gradient stochastique. Les algorithmes de ces méthodes sont donnés ci-dessous.

Phase 2 :

Le modèle est définie par la fonction $\sigma(X.\theta)$, dans laquelle il y'a 31 paramètres : θ(w0, w1, w2, …, w31), avec X une matrice de 455 lignes (nombre d'exemples) et 30 colonnes (caractéristiques).

Les données à fitter sont un ensemble de points $\{(x_i,y_i)\}_{i=1,N}$. Le but du problème est de trouver les paramètres optimaux θ tels que le modèle sépare efficacement les deux classes (tumeurs malignes et bénignes) dans le nuage de points. Pour cela, il faut minimiser la fonction de coût $J(\theta)$.

Un fichier comportant les coordonnées de 455 points est fourni (phase2_ND), et il faut déterminer les coefficients θ(w0, w1, w2, …, w31) qui minimisent le critère de distance $J(\theta)$.

Deux méthodes seront implémentées et testées : une descente de gradient par lot, puis la descente de gradient stochastique.

Méthodes du gradient

Le problème traité est de minimiser une certaine fonction $J(X)$, avec $X = (x_1, x_2, ..., x_{30})$. On suppose également que l'on connaît l'expression analytique du gradient.

Principe des méthodes de gradient :

De manière générale, les méthodes de descente sont des méthodes itératives qui génèrent une suite de points X_n tels que : $\forall n \geq 0 : J(X_n) < J(X_{n-1})$. Pour cela, une technique simple consiste à faire diminuer la valeur de J en se déplaçant d'une certaine quantité dans la direction de plus grande pente, donnée par le gradient au point considéré et normale à l'isovaleur passant par ce point.

Le schéma général de la méthode du gradient est donc le suivant :

$$X_n = X_{n-1} - \alpha . \nabla J(X_{n-1}) \qquad \text{avec} \quad \alpha > 0 \text{ choisi pour avoir : } J(X_n) < J(X_{n-1})$$

Selon la valeur choisie de α choisie, on se déplace plus ou moins loin dans la direction opposée au gradient.

Au voisinage de X_{n-1}, la fonction J est décroissante dans la direction opposée au gradient. En choisissant une valeur de α « très petite », on est sûr de faire diminuer J, ce qui garantit la convergence de l'algorithme, mais le déplacement à chaque itération, $\|X_n - X_{n-1}\|$, est petit et la convergence risque d'être longue. Il y a donc un compromis à trouver afin d'utiliser avec une valeur de α qui assure la convergence, sans trop la ralentir.

Schéma général de l'algorithme à implémenter :

Dans le cadre de ce projet, seule la méthode du gradient par lot sera implémentée.

- Notations :
  - X_{n-1} et X_n : point initial et point final de l'itération n.
  - n et n_{max} : compteur d'itérations et nombre maximal d'itérations autorisé
  - X_0 : point de départ de l'algorithme
  - α : pas de recherche

a. Algorithme de descente de gradient par lot

- Choix des paramètres de l'algorithme : X_0, α, n_{max}
- Initialisation : $X_{n-1} \leftarrow X_0$,, $n \leftarrow 0$
- Tant que t $n < n_{max}$:
  - $X_n \leftarrow X_{n-1} - \alpha . \nabla J(X_{n-1})$
  - $n \leftarrow n + 1$

b. Algorithme de descente de gradient stochastique

- Choix des paramètres de l'algorithme : X_0, α, n_{max}
- Initialisation : $X_{n-1} \leftarrow X_0$,, $n \leftarrow 0$
- Tant que $n < n_{max}$:
  - Choisir une exemple i aléatoirement
  - $X_n \leftarrow X_{n-1} - \alpha . \nabla J_i(X_{n-1})$

- $n \leftarrow n + 1$

Travail demandé

- Implémenter la fonction de coût J(θ) pour évaluer les performances du modèle.
- Mettre en œuvre la descente de gradient stochastique pour mettre à jour les paramètres du modèle et minimiser la fonction de coût.
- Appliquer la descente de gradient par lot pour comparer ses performances avec celles du gradient stochastique.
- Tracer les isovaleurs de J dans le plan (w1,w2), de même que la suite de points intermédiaires (w1(n),w2(n)). **(phase 1)**
- Évaluer le modèle en comparant les résultats prédits et les valeurs réelles du test.
- Analyser les résultats obtenus avec les deux méthodes pour déterminer le meilleur modèle de régression logistique pour la classification du cancer du sein.
- Présenter les conclusions et recommandations, en mettant en évidence les avantages et les inconvénients des méthodes de descente de gradient utilisées.