Introduction à la Vraisemblance et Régression Logistique Bernoulli

Chaîne YouTube: **Coder en Dur**

November 23, 2024

Plan

Qu'est-ce que la vraisemblance ?

Définition: La vraisemblance quantifie la probabilité que des données observées soient issues d'un modèle donné avec des paramètres spécifiques.

Notation: Soit :

- $x_1, x_2, \ldots, x_N$ un échantillon observé,
- $p(x \mid \theta)$ la probabilité donnée par un modèle paramétré par θ.

La **fonction de vraisemblance** est définie par :

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} p(x_i \mid \theta)$$

En pratique, on travaille souvent avec le **log de la vraisemblance** :

$$\ell(\theta) = \log \mathcal{L}(\theta) = \sum_{i=1}^{N} \log p(x_i \mid \theta)$$

Principe du Maximum de Vraisemblance

Objectif: Trouver le paramètre $\hat{\theta}$ qui maximise la vraisemblance :

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, \mathcal{L}(\theta)$$

Équivalent en log :

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, \ell(\theta)$$

Cette méthode est utile pour :

- Estimer les paramètres dans les modèles statistiques,
- Établir une correspondance entre les données et le modèle probabiliste.

Modèle de Régression Logistique

La régression logistique est utilisée pour modéliser la probabilité qu'une observation x_i appartienne à une classe $y_i \in \{0, 1\}$.

Formule:

$$P(y_i = 1 \mid x_i, \theta) = \sigma(\theta^\top x_i) = \frac{1}{1 + e^{-\theta^\top x_i}}$$

où $\sigma(z)$ est la fonction sigmoïde.

Pour $y_i = 0$, la probabilité est :

$$P(y_i = 0 \mid x_i, \theta) = 1 - \sigma(\theta^\top x_i)$$

Vraisemblance pour la Régression Logistique

Fonction de vraisemblance: Pour un échantillon $\{(x_i, y_i)\}_{i=1}^N$, la vraisemblance est donnée par :

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} P(y_i \mid x_i, \theta)$$

En remplaçant $P(y_i \mid x_i, \theta)$ par les expressions sigmoïdes :

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} \sigma(\theta^\top x_i)^{y_i} \cdot (1 - \sigma(\theta^\top x_i))^{1-y_i}$$

Log-vraisemblance:

$$\ell(\theta) = \sum_{i=1}^{N} \left[y_i \log \sigma(\theta^\top x_i) + (1 - y_i) \log(1 - \sigma(\theta^\top x_i)) \right]$$

Maximisation de la Log-vraisemblance

Pour maximiser la log-vraisemblance, nous minimisons la fonction de perte $J(\theta)$ en utilisant la descente de gradient.

Gradient de $J(\theta)$ **:**

$$\frac{\partial J(\theta)}{\partial \theta} = -\frac{1}{N}\sum_{i=1}^{N}\Big(y_i - \sigma(\theta^\top x_i)\Big)x_i$$

Preuve :

1. La fonction sigmoïde est donnée par :

$$\sigma(z) = \frac{1}{1+e^{-z}}, \quad \frac{d\sigma(z)}{dz} = \sigma(z)(1-\sigma(z))$$

2. La perte $J(\theta)$ est :

$$J(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\Big[y_i \log \sigma(\theta^\top x_i) + (1-y_i)\log(1-\sigma(\theta^\top x_i))\Big]$$

Dérivation du Gradient

Étape 1 : Calcul de $\frac{\partial J(\theta)}{\partial \theta}$

$$J(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log \sigma(\theta^\top x_i) + (1 - y_i) \log(1 - \sigma(\theta^\top x_i)) \right]$$

En dérivant chaque terme par rapport à θ :

$$\frac{\partial J(\theta)}{\partial \theta} = -\frac{1}{N} \sum_{i=1}^{N} \left[\frac{y_i}{\sigma(\theta^\top x_i)} \cdot \frac{\partial \sigma(\theta^\top x_i)}{\partial \theta} + \frac{1 - y_i}{1 - \sigma(\theta^\top x_i)} \cdot \frac{\partial (1 - \sigma(\theta^\top x_i))}{\partial \theta} \right]$$

Étape 2 : Remplacer $\frac{\partial \sigma(\theta^\top x_i)}{\partial \theta}$

$$\frac{\partial \sigma(\theta^\top x_i)}{\partial \theta} = \sigma(\theta^\top x_i)(1 - \sigma(\theta^\top x_i)) x_i$$

Dérivation du Gradient

Étape 3 : Simplification En remplaçant $\frac{\partial \sigma}{\partial \theta}$ dans l'expression de $J(\theta)$:

$$\frac{\partial J(\theta)}{\partial \theta} = -\frac{1}{N} \sum_{i=1}^{N} \Big[y_i (1 - \sigma(\theta^\top x_i)) - (1 - y_i) \sigma(\theta^\top x_i) \Big] x_i$$

Étape 4 : Regrouper les termes

$$\frac{\partial J(\theta)}{\partial \theta} = -\frac{1}{N} \sum_{i=1}^{N} \Big(y_i - \sigma(\theta^\top x_i) \Big) x_i$$

Ce qui donne l'expression finale du gradient.

Optimisation par descente de gradient

Pour trouver $\hat{\theta}$, on maximise la log-vraisemblance en utilisant la descente de gradient.

Gradient:

$$\frac{\partial \ell(\theta)}{\partial \theta} = \sum_{i=1}^{N} \left(y_i - \sigma(\theta^\top x_i) \right) x_i$$

Algorithme:

- Initialiser θ,
- Répéter jusqu'à convergence :

$$\theta \leftarrow \theta + \eta \frac{\partial \ell(\theta)}{\partial \theta}$$

où η est le taux d'apprentissage.

Conclusion

- La vraisemblance est une méthode puissante pour estimer les paramètres de modèles probabilistes.
- En régression logistique, elle permet de modéliser la relation entre variables explicatives et une variable cible binaire.
- L'optimisation se fait efficacement avec des algorithmes comme la descente de gradient.

Prochaines étapes : Implémentation pratique avec des exemples en Python.

Merci de suivre Coder en Dur !