TD 6: Régression logistique bayésienne

Objectifs:

1. (Re)-découvrir la régression logistique pénalisée
2. Comprendre l'équivalence entre approche fréquentiste et l'estimateur bayésien MAP
3. Comprendre les causes de l'underfitting/overfitting avec un modèle bayésien
4. Découvrir l'utilité des modèles hiérarchiques dans la sélection de modèle

1. Problématique, baseline et modélisation Dans les entreprises qui vendent un service (opérateur téléphonique par exemple), fidéliser les clients est un problème majeur. Le *churn rate* correspond au pourcentage de clients qui décident d'annuler leur abonnement (pour changer d'opérateur par exemple). Si l'entreprise réussit à anticiper si un tel client a grande une probabilité de *churn*, elle peut le cibler avec des services en plus, des gestes commerciaux etc. Nous avons la base de données d'un opérateur téléphonique avec les variables:

- `Dependents` : Variable binaire indiquant si le client a des personnes à charge (0/1)
- `TechSupport` : Variable binaire indiquant si le client a souscrit au support technique (0/1)
- `Contract` : Variable binaire indiquant le type de contrat (0: mensuel, 1: long terme)
- `InternetService` : Variable binaire indiquant le type de service internet (0: fibre, 1: DSL)
- `Months` : Variable numérique continue indiquant la durée d'abonnement en mois
- `MonthlyCharges` : Variable numérique continue indiquant le montant mensuel facturé en dollars
- `Churn` : Variable binaire cible indiquant si le client a résilié son contrat (0: non, 1: oui)

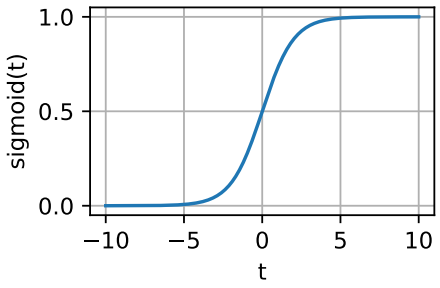
On note la variable Churn par y et les autres variables par un vecteur aléatoire $\mathbf{X} \in \mathbb{R}^6$. On souhaite utiliser $\mathbf{X}$ pour prédire y. Soit x un vecteur d'observation d'un client ($x \in \mathbb{R}^6$). La régression logistique (fréquentiste) consiste à modéliser la probabilité conditionnelle: $\mathbb{P}(y = y|\mathbf{X} = \mathbf{x})$ en utilisant un modèle linéaire simple avec un paramètre inconnue β. On considère alors une combinaison linéaire $\alpha + \beta_1 x_1 + \dots \beta_6 x_6$ que l'on peut noter par le produit scalaire $\alpha + \langle \beta, \mathbf{x} \rangle$. Or cette combinaison linéaire est réelle, pour qu'elle modélise une probabilité il faut la transformer en [0, 1]. Pour cela on utilise la fonction *sigmoid* (aussi appelée *logistic*): $\sigma : t \mapsto \frac{1}{1+e^{-t}}$:

1. Régression logistique classique La fonction sigmoid transforme donc la droite rélle en [0, 1] d'une façon "lisse" en passant par 0.5 en 0. La régression logistique correspond au modèle:

$$\mathbb{P}(y = 1|\mathbf{X} = \mathbf{x}) = \text{sigmoid}(\alpha + \langle \beta, \mathbf{x} \rangle)$$

Ainsi, plus le score de la combinaison linéaire sera grand, plus la probabilité s'approchera de 1.

1. Avec quel changement de variables peut-on redéfinir le même modèle sous la forme $\mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}) = \text{sigmoid}(\langle \beta, \mathbf{x} \rangle)$ avec $\beta \in \mathbb{R}^7$?



2. Montrez que sous ce modèle, les cotes de probabilités (*odds*) vérifient:

$$\frac{\mathbb{P}(y=1|\mathbf{X}=\mathbf{x})}{\mathbb{P}(y=0|\mathbf{X}=\mathbf{x})} = \exp(\langle \beta, \mathbf{x} \rangle)$$

   En déduire une interprétation des coefficients de régression β.

3. Quelle est la loi de $y|\mathbf{X}$?

4. La régression logistique fréquentiste consiste à trouver le meilleur paramètre β en maximisant la log-vraisemblance. On suppose qu'on observe $(\mathbf{X}_1, y_1), \ldots, (\mathbf{X}_n, y_n)$ i.i.d suivant une loi jointe inconnue mais telle que la loi conditionnelle $y|\mathbf{X}$ est celle de la question précédente. Peut-on quand même écrire le problème d'optimisation du maximum de vraisemblance en fonction des données ?

5. Montrez que ce problème d'optimisation peut s'écrire sous la forme classique ML avec la fonction de perte *cross-entropy*:

$$\min_{\beta \in \mathbb{R}^7} \quad \frac{1}{n} \sum_{i=1}^{n} y_i \log(1 + e^{-\beta^\top \mathbf{x}_i}) + (1 - y_i) \log(1 + e^{\beta^\top \mathbf{x}_i})$$

6. Pour éviter le sur-apprentissage, on ajoute une pénalité de type ℓ_2:

$$\min_{\beta \in \mathbb{R}^7} \quad \frac{1}{n} \sum_{i=1}^{n} y_i \log(1 + e^{-\beta^\top \mathbf{x}_i}) + (1 - y_i) \log(1 + e^{\beta^\top \mathbf{x}_i}) + \frac{\lambda}{2} \|\beta\|_2^2 \tag{1}$$

   avec $\lambda > 0$ un hyperparamètre. Les variables continues n'ont pas le même ordre de grandeur (dizaines / milliers). Quel est l'impact sur l'ordre de grandeur des β correspondant avec $\lambda = 0$? et $\lambda > 0$?

   On suppose que chaque variable continue est standardisée: centrée et divisée par son écart-type. Cette étape est nécessaire pour les modèles linéaires pour (1) avoir des coefficients β comparables, (2) pénalisés de façon égale, (3) optimisés avec une descente de gradient stable.

2. Régression logistique bayésienne (Ridge) On suppose à présent que les β ne sont plus des paramètres mais des variables aléatoires suivant une loi a priori π donnée.

7. Déterminer la formule de la densité de la loi a posteriori si $\pi = \mathcal{N}(0, \gamma \mathrm{Id})$.

8. Établir le problème d'optimisation de l'estimateur MAP (Maximum a posteriori) et montrez qu'il est équivalent au problème (1).

9. Interprétez la relation entre λ et γ. Quel est l'effet de chacun de ces paramètres ?

10. L'approche bayésienne consiste à simuler plusieurs échantillons $\beta_1, \ldots, \beta_m$ suivant la loi a posteriori $\beta|\mathbf{x}_i$ puis de calculer pour le même $\mathbf{x}_i$ plusieurs probabilités $\mathbb{P}(y_i = 1)$ notées par $p_1, \ldots, p_m$ avec la sigmoid. Quel est l'intérêt de l'approche bayésienne dans le cadre de la régression logistique comparé à l'approche classique ?

3. Régression logistique sparse On suppose à présent que nous sommes en grande dimension avec $d > n$. Il est raisonnable donc de supposer que seuls quelques variables sont utiles pour la prédiction. On considère alors la pénalité de type Lasso qui donne un β "sparse" c-à-d avec beaucoup de 0:

$$\min_{\beta \in \mathbb{R}^{d+1}} \quad \frac{1}{n} \sum_{i=1}^{n} y_i \log(1 + e^{-\beta^\top \mathbf{x}_i}) + (1 - y_i) \log(1 + e^{\beta^\top \mathbf{x}_i}) + \lambda \|\beta\|_1 \tag{2}$$

11. Pour quelle loi a priori sur β aurait-on un Maximum-a-Posteriori (MAP) équivalent au problème (2) ?

12. Quelle difficulté principale risque de se poser pour simuler une chaîne MCMC suivant la loi a posteriori avec l'algorithme HMC ou NUTS ?

13. Soit $\tau > 0$, $Z \sim \mathcal{E}xp(\tau)$ et $\mathbf{A}|Z \sim \mathcal{N}(0, 2Z)$. Déterminez la loi de $\mathbf{A}$.

14. En déduire un modèle bayésien hiérarchique pour simuler la loi a posteriori avec l'apriori de la question 11.

4. Séléction de modèle Dans cette section, on reprend le modèle avec pénalité "Ridge". Dans le cadre fréquentiste, on choisit l'hyperparamètre λ en utilisant la validation croisée. Ce paramètre est celui qui minimise l'erreur de prédiction sur des données *test*. Ainsi ce paramètre est "data-driven" (inféré par les données).

Dans le cadre bayésien, il est également data-driven mais en suivant une procédure différente: on considère que ce paramètre est une variable aléatoire suivant une hyper-prior avec une variance assez grande comme la loi Cauchy tronquée sur les réels positifs. Ainsi, "on laisse" le modèle simuler toute une distribution sur ce paramètre de régularisation. On obtient un modèle hiérarchique:

$$\beta|\gamma \sim \mathcal{N}(0, \gamma \mathrm{Id}) \qquad \gamma \sim \mathrm{HalfCauchy}(0, 1)$$

15. À quoi correspond l'intégrale $\int f(\beta|\mathrm{data}, \gamma) f(\gamma) \mathrm{d}\gamma$?

16. En déduire une interprétation de la sélection de modèle bayésienne.