TD 6: Régression logistique: Ridge, Lasso, Bayésienne - Corrigé

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. On peut ajouter une colonne constante égale à 1 à notre vecteur de variables explicatives. Soit $\tilde{\mathbf{x}} = (1, x_1, x_2, \ldots, x_6) \in \mathbb{R}^7$ et $\tilde{\beta} = (\alpha, \beta_1, \beta_2, \ldots, \beta_6) \in \mathbb{R}^7$. Alors on a :

   $$\alpha + \langle \beta, \mathbf{x} \rangle = \langle \tilde{\beta}, \tilde{\mathbf{x}} \rangle$$

   Et donc,

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

   Dans toute la suite on considère par abus de notation que ce changement de variable est fait, ainsi $\beta, \mathbf{x} \in \mathbb{R}^7$ et:

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

2. On a

   $$\frac{\mathbb{P}(y = 1|\mathbf{X} = \mathbf{x})}{\mathbb{P}(y = 0|\mathbf{X} = \mathbf{x})} = \frac{\text{sigmoid}(\langle \beta, \mathbf{x} \rangle)}{1 - \text{sigmoid}(\langle \beta, \mathbf{x} \rangle)} = \frac{\frac{1}{1+e^{-\langle \beta, \mathbf{x} \rangle}}}{1 - \frac{1}{1+e^{-\langle \beta, \mathbf{x} \rangle}}} = \frac{1}{1 + e^{-\langle \beta, \mathbf{x} \rangle}} \cdot \frac{1 + e^{-\langle \beta, \mathbf{x} \rangle}}{e^{-\langle \beta, \mathbf{x} \rangle}} = e^{\langle \beta, \mathbf{x} \rangle}$$

   Si on augmente la valeur de la variable x_j d'une unité, en gardant toutes les autres variables constantes, alors les cotes (odds) sont multipliées par e^{β_j}. Donc :

   - Si $\beta_j > 0$, alors une augmentation de x_j augmente la probabilité de $y = 1$
   - Si $\beta_j < 0$, alors une augmentation de x_j diminue la probabilité de $y = 1$
   - Si $\beta_j = 0$, alors la variable x_j n'a pas d'influence sur la probabilité de $y = 1$

3. La loi de $y|\mathbf{X}$ est une loi de Bernoulli de paramètre $p = \text{sigmoid}(\langle \beta, \mathbf{x} \rangle)$, c'est-à-dire :

   $$y|\mathbf{X} \sim \text{Bernoulli}(\text{sigmoid}(\langle \beta, \mathbf{x} \rangle))$$

   Car $\mathbb{P}(y = 1|\mathbf{X} = \mathbf{x}) = \text{sigmoid}(\langle \beta, \mathbf{x} \rangle) \in [0, 1]$.

4. Même si la loi jointe $(\mathbf{X}, y)$ est inconnue, on a un modèle paramétrique pour la loi conditionnelle $y|\mathbf{X}$, ce qui nous permet de construire la vraisemblance avec Bayes:

   $$\ell(\beta) = \log \prod_{i=1}^{n} P(y_i, \mathbf{X}_i; \beta) = \log \prod_{i=1}^{n} P(y_i|\mathbf{X}_i; \beta) P(\mathbf{X}_i) = \sum_{i=1}^{n} \log P(y_i|\mathbf{X}_i; \beta) + \sum_{i=1}^{n} \log P(\mathbf{X}_i)$$

Ainsi, comme $\log P(\mathbf{X}_i)$ ne dépend pas de β, maximiser le log de la loi jointe est équivalent à maximiser:

$$\sum_{i=1}^{n} \log P(y_i|\mathbf{X}_i; \beta)$$

5. Développons l'expression de la question précédente:

$$\begin{aligned}
&\sum_{i=1}^{n} \log \left[(\text{sigmoid}(\langle \beta, \mathbf{x}_i \rangle))^{y_i} \cdot (1 - \text{sigmoid}(\langle \beta, \mathbf{x}_i \rangle))^{1-y_i}\right] \\
&= \sum_{i=1}^{n} \left[y_i \log(\text{sigmoid}(\langle \beta, \mathbf{x}_i \rangle)) + (1 - y_i) \log(1 - \text{sigmoid}(\langle \beta, \mathbf{x}_i \rangle))\right] \\
&= \sum_{i=1}^{n} \left[y_i \log\left(\frac{1}{1 + e^{-\langle \beta, \mathbf{x}_i \rangle}}\right) + (1 - y_i) \log\left(\frac{e^{-\langle \beta, \mathbf{x}_i \rangle}}{1 + e^{-\langle \beta, \mathbf{x}_i \rangle}}\right)\right] \\
&= \sum_{i=1}^{n} \left[y_i \log\left(\frac{1}{1 + e^{-\langle \beta, \mathbf{x}_i \rangle}}\right) + (1 - y_i) \log\left(\frac{1}{1 + e^{\langle \beta, \mathbf{x}_i \rangle}}\right)\right] \\
&= \sum_{i=1}^{n} \left[-y_i \log(1 + e^{-\langle \beta, \mathbf{x}_i \rangle}) - (1 - y_i) \log(1 + e^{\langle \beta, \mathbf{x}_i \rangle})\right] \\
&= -\sum_{i=1}^{n} \left[y_i \log(1 + e^{-\langle \beta, \mathbf{x}_i \rangle}) + (1 - y_i) \log(1 + e^{\langle \beta, \mathbf{x}_i \rangle})\right]
\end{aligned}$$

Multiplions par -1 pour transformer en problème de minimisation et divisons par n pour obtenir la moyenne empirique :

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \left[y_i \log(1 + e^{-\langle \beta, \mathbf{x}_i \rangle}) + (1 - y_i) \log(1 + e^{\langle \beta, \mathbf{x}_i \rangle})\right]$$

On retrouve le problème d'optimisation vu en cours avec la perte cross-entropy.

6. Avec $\lambda = 0$, l'ordre de grandeur des coefficients β correspondant aux variables continues (Months et MonthlyCharges) sera très différent de celui des coefficients des variables binaires.

   En effet, comme les variables continues ont un ordre de grandeur beaucoup plus grand (Months peut être de l'ordre de dizaines et MonthlyCharges de l'ordre de milliers), les coefficients β correspondants seront naturellement beaucoup plus petits pour que le produit $\beta_j \cdot x_j$ reste du même ordre de grandeur que les autres variables.

   Avec $\lambda > 0$, la pénalité ℓ_2 a tendance à réduire l'amplitude de tous les coefficients, mais cet effet est plus prononcé sur les coefficients de grande valeur. Ainsi, les coefficients des variables binaires (qui pourraient être naturellement plus grands) seront davantage pénalisés que les coefficients des variables continues (qui sont déjà petits).

   Cette inégalité dans la pénalisation est problématique car elle peut conduire à une sélection biaisée des variables. C'est pourquoi la standardisation des variables continues est nécessaire avant

d'appliquer la régression logistique pénalisée, pour que tous les coefficients soient comparables et pénalisés de façon équitable.

Enfin, l'optimisation se fait par descente de gradient sur l'espace $\mathbb{R}^7$, si le meilleur β a des coordonnées avec des ordres de grandeur très différents, il sera situé dans une région avec des courbures très différentes (valeur absolue du gradient qui dépend fortement de la direction) ce qui risque d'introduire une instabilité lors de l'optimisation.

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. Soit $\pi = \mathcal{N}(0, \gamma \mathrm{Id})$ la densité a priori sur β. D'après le théorème de Bayes, la densité a posteriori est proportionnelle au produit de la vraisemblance et de la densité a priori :

$$p(\beta|\text{données}) \propto p(\text{données}|\beta) \cdot p(\beta)$$

La log-vraisemblance est :

$$\log p(\text{données}|\beta) = -\sum_{i=1}^{n} \left[y_i \log(1 + e^{-\langle \beta, \mathbf{x}_i \rangle}) + (1 - y_i) \log(1 + e^{\langle \beta, \mathbf{x}_i \rangle}) \right]$$

La log-densité a priori est :

$$\log p(\beta) = \log \left(\frac{1}{(2\pi\gamma)^{d/2}} \exp\left(-\frac{1}{2\gamma} \|\beta\|_2^2 \right) \right) = -\frac{d}{2} \log(2\pi\gamma) - \frac{1}{2\gamma} \|\beta\|_2^2$$

Donc, la log-densité a posteriori est (à une constante additive près) :

$$\log p(\beta|\text{données}) \propto -\sum_{i=1}^{n} \left[y_i \log(1 + e^{-\langle \beta, \mathbf{x}_i \rangle}) + (1 - y_i) \log(1 + e^{\langle \beta, \mathbf{x}_i \rangle}) \right] - \frac{1}{2\gamma} \|\beta\|_2^2$$

8. L'estimateur MAP (Maximum A Posteriori) correspond au mode de la distribution a posteriori, c'est-à-dire au maximum de la log-densité a posteriori :

$$\arg\max_{\beta} \left\{ -\sum_{i=1}^{n} \left[y_i \log(1 + e^{-\langle \beta, \mathbf{x}_i \rangle}) + (1 - y_i) \log(1 + e^{\langle \beta, \mathbf{x}_i \rangle}) \right] - \frac{1}{2\gamma} \|\beta\|_2^2 \right\}$$

En multipliant par -1 et en divisant par n, on obtient un problème de minimisation :

$$\hat{\beta}_{MAP} = \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[y_i \log(1 + e^{-\langle \beta, \mathbf{x}_i \rangle}) + (1 - y_i) \log(1 + e^{\langle \beta, \mathbf{x}_i \rangle}) \right] + \frac{1}{2n\gamma} \|\beta\|_2^2 \right\}$$

En comparant avec le problème Ridge, on voit que les deux problèmes sont équivalents en posant $\lambda = \frac{1}{n\gamma}$.

9. La relation entre λ et γ est $\lambda = \frac{1}{n\gamma}$.

Effet des paramètres :

- γ est la variance de la distribution a priori gaussienne. Plus γ est grand, plus la distribution a priori est diffuse, ce qui signifie que nous avons moins d'informations a priori sur les valeurs de β. À l'inverse, plus γ est petit, plus la distribution a priori est concentrée autour de zéro, ce qui signifie que nous avons une forte croyance a priori que les valeurs de β sont proches de zéro.
- λ est le paramètre de régularisation dans la régression logistique pénalisée. Plus λ est grand, plus la pénalité sur les grands coefficients est forte, ce qui conduit à un modèle plus simple (underfitting / sous-apprentissage). À l'inverse, plus λ est petit, moins la pénalité est forte, ce qui peut conduire à un surapprentissage / overfitting si les données sont bruitées.

La relation $\lambda = \frac{1}{n\gamma}$ montre que ces paramètres sont inversement proportionnels :

- Un petit γ (a priori concentré autour de zéro) correspond à un grand λ (forte régularisation).
- Un grand γ (a priori diffus) correspond à un petit λ (faible régularisation).

10. Au lieu d'obtenir une seule estimation ponctuelle comme dans l'approche classique, l'approche bayésienne fournit une distribution complète pour les paramètres β et donc pour les probabilités prédites. Cela permet de quantifier l'incertitude de nos prédictions avec des intervalles de crédibilité pour les probabilités prédites, ce qui donne une mesure de la confiance dans nos prédictions.

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:

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

11. Pour obtenir un MAP équivalent au problème Lasso (1), la loi a priori sur β doit être une distribution de Laplace. En effet, la densité de la loi de Laplace est proportionnelle à $e^{-\tau|\beta|}$, et le logarithme de cette densité est proportionnel à $-\tau|\beta|$.

    Ainsi, si on choisit $\pi(\beta) \propto \exp(-\tau\|\beta\|_1)$ avec $\tau = \lambda \cdot n$, où $\|\beta\|_1 = \sum_{j=1}^{d+1} |\beta_j|$ est la norme L_1 de β, alors le problème MAP correspondant sera équivalent au problème Lasso.

    Plus précisément, si on choisit $\pi(\beta) = \prod_{j=1}^{d+1} \frac{\tau}{2} \exp(-\tau|\beta_j|)$, c'est-à-dire que chaque coefficient β_j suit indépendamment une loi de Laplace de paramètre τ, alors la log-densité a priori est (à une constante additive près) : $\log \pi(\beta) \propto -\tau\|\beta\|_1$

    En posant $\tau = \lambda \cdot n$, le problème MAP devient :

$$\hat{\beta}_{MAP} = \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[y_i \log(1 + e^{-\langle \beta, \mathbf{x}_i \rangle}) + (1 - y_i) \log(1 + e^{\langle \beta, \mathbf{x}_i \rangle}) \right] + \lambda \|\beta\|_1 \right\}$$

12. Les algorithmes HMC (Hamiltonian Monte Carlo) et NUTS (No-U-Turn Sampler) reposent sur le calcul des gradients de la log-densité pour proposer des mouvements efficaces dans l'espace des paramètres. Or, la distribution de Laplace n'est pas différentiable lorsque le paramètre est exactement égal à zéro, ce qui peut causer des problèmes numériques.

13. Nous allons montrer que si $\tau > 0$, $\sigma^2 \sim \mathcal{E}xp(\frac{\tau^2}{2})$ et $\mathbf{A}|\sigma^2 \sim \mathcal{N}(0, \sigma^2)$, alors $\mathbf{A} \sim \text{Laplace}(\tau)$.
Pour cela, calculons la densité marginale de $\mathbf{A}$ en intégrant sur σ^2 :

$$\begin{aligned} p(\mathbf{A}) &= \int_0^\infty p(\mathbf{A}|\sigma^2)p(\sigma^2)d\sigma^2 \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\mathbf{A}^2}{2\sigma^2}\right) \cdot \frac{\tau^2}{2} \exp\left(-\frac{\tau^2}{2}\sigma^2\right) d\sigma^2 \\ &= \frac{\tau^2}{2\sqrt{2\pi}} \int_0^\infty \frac{1}{\sqrt{\sigma^2}} \exp\left(-\frac{\mathbf{A}^2}{2\sigma^2} - \frac{\tau^2}{2}\sigma^2\right) d\sigma^2 \end{aligned}$$

Posons $\sigma = \sqrt{\sigma^2}$ pour mieux visualiser l'intégrale. Avec $d\sigma^2 = 2\sigma d\sigma$, nous obtenons :

$$\begin{aligned} p(\mathbf{A}) &= \frac{\tau^2}{2\sqrt{2\pi}} \int_0^\infty \frac{1}{\sigma} \exp\left(-\frac{\mathbf{A}^2}{2\sigma^2} - \frac{\tau^2}{2}\sigma^2\right) \cdot 2\sigma d\sigma \\ &= \frac{\tau^2}{\sqrt{2\pi}} \int_0^\infty \exp\left(-\frac{1}{2}\left(\frac{\mathbf{A}^2}{\sigma^2} + \tau^2\sigma^2\right)\right) d\sigma \end{aligned}$$

On complète le carré de l'argument de l'exponentielle pour pouvoir l'intégrer:

$$\frac{1}{2}\left(\frac{\mathbf{A}^2}{\sigma^2} + \tau^2\sigma^2\right) = \frac{1}{2}\left(\frac{|\mathbf{A}|}{\sigma} - \tau\sigma\right)^2 + |\mathbf{A}|\tau$$

$$\begin{aligned} p(\mathbf{A}) &= \frac{\tau^2}{\sqrt{2\pi}} \int_0^\infty \exp\left(-\frac{1}{2}\left(\frac{|\mathbf{A}|}{\sigma} - \tau\sigma\right)^2 - |\mathbf{A}|\tau\right) d\sigma \\ &= \frac{\tau^2}{\sqrt{2\pi}} e^{-|\mathbf{A}|\tau} \int_0^\infty \exp\left(-\frac{1}{2}\left(\frac{|\mathbf{A}|}{\sigma} - \tau\sigma\right)^2\right) d\sigma \end{aligned}$$

Effectuons le changement de variable $u = \frac{|\mathbf{A}|}{\sigma} - \tau\sigma$. Quand $\sigma \to 0^+$, $u \to +\infty$, et quand $\sigma \to +\infty$, $u \to -\infty$. $u = \frac{|\mathbf{A}|}{\sigma} - \tau\sigma \Rightarrow \tau\sigma^2 + u\sigma - |\mathbf{A}| = 0$. En gardant la racine positive de ce polynôme, nous obtenons :

$$\sigma = \frac{-u + \sqrt{u^2 + 4\tau|\mathbf{A}|}}{2\tau}$$

On a $d\sigma = \frac{-du + \frac{udu}{\sqrt{u^2+4\tau|\mathbf{A}|}}}{2\tau} = -\frac{1}{2\tau}\left(1 - \frac{u}{\sqrt{u^2+4\tau|\mathbf{A}|}}\right) du$. Ainsi l'intégrale devient:

$$
\begin{aligned}
p(\mathbf{A}) &= \frac{\tau^2}{\sqrt{2\pi}} e^{-|\mathbf{A}|\tau} \int_{+\infty}^{-\infty} -\exp\left(-\frac{u^2}{2}\right) \frac{1}{2\tau}\left(1 - \frac{u}{\sqrt{u^2 + 4\tau|\mathbf{A}|}}\right) du \\
&= \frac{\tau}{2} e^{-|\mathbf{A}|\tau} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{u^2}{2}\right)\left(1 - \frac{u}{\sqrt{u^2 + 4\tau|\mathbf{A}|}}\right) du \\
&= \frac{\tau}{2} e^{-|\mathbf{A}|\tau} \frac{1}{\sqrt{2\pi}} \left[\underbrace{\int_{-\infty}^{+\infty} \exp\left(-\frac{u^2}{2}\right) du}_{\text{intégrale d'une gaussienne}} - \underbrace{\int_{-\infty}^{+\infty} \exp\left(-\frac{u^2}{2}\right) \frac{u}{\sqrt{u^2 + 4\tau|\mathbf{A}|}} du}_{=0 \text{ car intégrale d'une fonction impaire}} \right] \\
&= \frac{\tau}{2} e^{-|\mathbf{A}|\tau}
\end{aligned}
$$

Ce qui correspond bien à la densité d'une loi de Laplace de paramètre τ.

14. La question précédente nous permet de définir un modèle bayésien hiérarchique équivalent au Lasso, avec $\tau = \lambda \cdot n$:

$$(\beta_j|\sigma^2) \sim \mathcal{N}(0, \sigma^2) \qquad \sigma^2 \sim \mathcal{E}xp\left(\frac{\tau^2}{2}\right)$$

Ce modèle hiérarchique a l'avantage d'être constitué uniquement de distributions différentiables (gaussiennes et exponentielles), ce qui le rend compatible avec les méthodes MCMC basées sur le gradient comme HMC ou NUTS.

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. L'intégrale $\int f(\beta|\mathrm{data}, \gamma) f(\gamma) \mathrm{d}\gamma$ correspond à la densité marginale a posteriori de β sachant les données, où on a intégré sur tous les valeurs possibles de l'hyperparamètre γ.

Plus précisément : - $f(\beta|\mathrm{data}, \gamma)$ est la densité a posteriori de β sachant les données et une valeur fixée de γ - $f(\gamma)$ est la densité a priori de γ - L'intégrale représente une moyenne pondérée des densités a posteriori de β pour différentes valeurs de γ, où les poids sont donnés par la densité a priori de γ.

16. Dans l'approche fréquentiste, on sélectionne une seule valeur "optimale" de l'hyperparamètre λ (ou équivalent, γ) et on base toutes nos prédictions sur cette unique valeur. Cela revient à choisir un seul modèle parmi une famille de modèles potentiels.

    En revanche, dans l'approche bayésienne hiérarchique, on ne se limite pas à une seule valeur de γ. Au lieu de cela, on simule et on intègre sur toute la distribution a posteriori de γ. Cela revient à considérer tous les modèles possibles (correspondant à différentes valeurs de γ) et à pondérer leurs prédictions en fonction de leur probabilité a posteriori. La sélection de modèle bayésienne peut être interprétée comme une forme de "model averaging" (moyennage de modèles). Ceci permet de prendre en compte l'incertitude sur la valeur de l'hyperparamètre γ, ce qui peut être important lorsque les données ne contraignent pas fortement cette valeur – si par exemple la variance de l'erreur de validation croisée est très grande.