UNIVERSITÉ DE SHERBROOKE

IFT 712 - Devoir 2

Entropie et classification linéaire

Ait ichou Yoann

Cai Yunfan

Gaye ElHadji Habib

Faculté des sciences,
Département Informatique

4 novembre 2022

Table des matières

1 Minimisation de fonction d'énergie

Soit la mesure d'entropie suivante dans le cas d'une régression Ridge :

$$E(\vec{w}) = \sum_{n=1}^{N} (t_n - \vec{w}^T \phi(x_n))^2 + \lambda \vec{w}^T \vec{w}$$

Démontrons que l'optimisation suivante :

$$\vec{w} = argmin_{\vec{w}}(E(\vec{w}))$$

A pour solution :

$$\vec{w} = (\Phi^T \Phi + \lambda I)^{-1} \Phi^T t$$

Calculons d'abord le gradient de la fonction d'énergie $E(\vec{w})$:

$$\begin{aligned}
\nabla E(\vec{w}) &= -2 \sum_{n=1}^{N} (t_n - \vec{w}^T \phi(x_n)) \phi(x_n)^T + 2\lambda \vec{w}^T && (1) \\
&= -2 \sum_{n=1}^{N} (t_n \phi(x_n)^T) + \vec{w}^T \sum_{n=1}^{N} \phi(x_n) \phi(x_n)^T + 2\lambda \vec{w}^T && (2) \\
&= -2 t^T \Phi + 2 \vec{w}^T \Phi^T \Phi + 2\lambda \vec{w}^T && (3)
\end{aligned}$$

Forçons maintenant le gradient à 0 pour obtenir la solution de notre optimisation :

$$\nabla E(\vec{w}) = 0$$

$$-t^T \Phi + \vec{w}^T \Phi^T \Phi + \lambda \vec{w}^T = 0$$

Par transposition, nous obtenons :

$$-\Phi^T t + \Phi^T \Phi \vec{w} + \lambda \vec{w} = 0$$

Puis en isolant $\vec{w}$:

$$(\Phi^T \Phi + \lambda I) \vec{w} = \Phi^T t$$

$$\vec{w} = (\Phi^T \Phi + \lambda I)^{-1} \Phi^T t$$

2 Entropie croisée et régression logistique : gradient

Démontrons que l'entropie croisée étudiée ici, qui s'écrit :

$$E(\vec{w}) = -\sum_{n=1}^{N} t_n ln(y_w(x_n)) + (1-t_n) ln(1-y_w(x_n))$$

Est équivalente à la formulation ci après :

$$\nabla E(\vec{w}) = \sum_{n=1}^{N} \phi(x_n)(y_n - t_n)$$

dans le cas de la fonction de probabilité conditionnelle suivante :

$$p(C_1|\vec{\phi}(\vec{x})) = \sigma(\vec{w}^T\vec{\phi}(\vec{x})) = \frac{1}{1+e^{-\vec{w}^T\vec{\phi}(\vec{x})}}$$

Rappelons que la dérivée de la fonction sigmoïde est équivalente à :

$$\frac{d\sigma}{dx}(w^T\vec{\phi}(\vec{x})) = \vec{\phi}(\vec{x})\sigma(w^T\vec{\phi}(\vec{x})) \times (1-\sigma(w^T\vec{\phi}(\vec{x})))$$

d

Démonstration :

$$\begin{aligned}
\frac{d\sigma(x)}{dx} &= \frac{e^{-x}}{(1+e^{-x})^2} && (4)\\
&= \frac{1}{1+e^{-x}} \times \frac{e^{-x}}{1+e^{-x}} && (5)\\
&= \sigma(x) \times \frac{1+e^{-x}-1}{1+e^{-x}} && (6)\\
&= \sigma(x) \times [\frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}}] && (7)\\
&= \sigma(x) \times (1-\sigma(x)) && (8)
\end{aligned}$$

De plus :

$$\frac{df(g(x))}{dx} = \frac{dg(x)}{dx} \times \frac{df(g(x))}{dg(x)}$$

Nous obtenons donc :

$$\frac{d\sigma(w^T\vec{\phi}(\vec{x}))}{dw} = \frac{w^T\vec{\phi}(\vec{x})}{dw}\sigma(w^T\vec{\phi}(\vec{x})) \times (1-\sigma(w^T\vec{\phi}(\vec{x}))) = \vec{\phi}(\vec{x})^T\sigma(w^T\vec{\phi}(\vec{x})) \times (1-\sigma(w^T\vec{\phi}(\vec{x})))$$

Nous obtenons la formule suivante pour le gradient de l'entropie croisée :

$$\nabla E(\vec{w}) = -\sum_{n=1}^{N}[t_n\vec{\phi}(\vec{x_n})^T \frac{y_w(\vec{\phi}(\vec{x_n}))(1-y_w(\vec{\phi}(\vec{x_n})))}{y_w(\vec{\phi}(\vec{x_n}))}$$
$$\text{- (1-t}_n\text{)}\vec{\phi}(\vec{x_n})^T \frac{y_w(\vec{\phi}(\vec{x_n}))(1-y_w(\vec{\phi}(\vec{x_n})))}{1-y_w(\vec{\phi}(\vec{x_n}))}]$$

Posons :

$$y_n = y_w(\vec{\phi}(\vec{x_n}))$$

Nous obtenons donc :

$$\begin{aligned}
\nabla E(\vec{w}) &= -\sum_{n=1}^{N}[t_n\vec{\phi}(\vec{x_n})^T \frac{y_n(1-y_n)}{y_n} - (1-t_n)\frac{\vec{\phi}(\vec{x_n})^T y_n(1-y_n)}{1-y_n}] && (9)\\
&= -\sum_{n=1}^{N}[t_n\vec{\phi}(\vec{x_n})^T(1-y_n) - (1-t_n)\vec{\phi}(\vec{x_n})^T y_n] && (10)\\
&= \sum_{n=1}^{N} = \vec{\phi}(\vec{x_n})^T(-t_n + y_n t_n + y_n - y_n t_n) && (11)\\
&= \sum_{n=1}^{N} \vec{\phi}(\vec{x_n})^T(y_n - t_n) && (12)
\end{aligned}$$

Ce qui est équivalent à :

$$\nabla E(\vec{w}) = \sum_{n=1}^{N}(y_n - t_n)\vec{\phi}(\vec{x_n})$$

3 Application du Lagrangien

La fonction entropie s'exprime comme suit :

$$H = -\sum_i P_i log_2 P_i$$

Notre système est soumis aux deux contraintes suivantes :

$$\text{C} : \begin{cases} P_1 + P_2 + P_3 = 1 \\ P_1 = 2P_2 \end{cases}$$

Nous cherchons à maximiser l'entropie tout en appliquant les contraintes précédentes à notre problème d'optimisation pour restreindre notre espace de solutions à celles qui nous intéressent. Pour ce faire nous devons combiner notre formule de l'entropie à nos contraintes, en appliquant le principe de Lagrange.

Soit λ_1 et $\lambda_2 \in R$. Le lagrangien L s'écrit :

$$\begin{aligned} L &= H - \lambda_1(P_1 + P_2 + P_3 - 1) - \lambda_2(P_1 - 2P_2) && (13) \\ &= -\sum_i P_i log_2 P_i - \lambda_1(P_1 + P_2 + P_3 - 1) - \lambda_2(P_1 - 2P_2) && (14) \end{aligned}$$

Etudions maintenant les dérivées partielles du lagrangien L en P_1, P_2 et P_3 , puis forçons le à 0 pour résoudre notre problème d'optimisation :

- $\frac{\partial L}{\partial P_1} = -log_2 P_1 - 1 - \lambda_1 - \lambda_2 = 0$
- $\frac{\partial L}{\partial P_2} = -log_2 P_2 - 1 - \lambda_1 + 2\lambda_2 = 0$
- $\frac{\partial L}{\partial P_3} = -log_2 P_3 - 1 - \lambda_1 = 0$
- $\frac{\partial L}{\partial \lambda_1} = -(P_1 + P_2 + P_3 - 1) = 0$
- $\frac{\partial L}{\partial \lambda_2} = -(P_1 - 2P_2) = 0$

Nous pouvons déduire les P_i satisfaisant ces contraintes en fonction des λ_1 et λ_2 . :

$$-log_2 P_i + k = 0$$

$$\Leftrightarrow log_2 P_i = k$$

$$\Leftrightarrow P_i = 2^k$$

Ce qui nous donne finalement :
$P_1 = 2^{-1-\lambda_1-\lambda_2}$
$P_2 = 2^{-1-\lambda_1+2\lambda_2}$
$P_3 = 2^{-1-\lambda_1}$

Afin de déterminer les valeurs des P_i, nous devons calculer les valeurs de λ_1 et λ_2 en se référant aux contraintes de départ, retrouvées en dérivant notre lagrangien L par λ_1 et λ_2 :

$$\mathrm{C}: \begin{cases} P_1 + P_2 + P_3 = 1 \\ P_1 = 2P_2 \end{cases} \Leftrightarrow \begin{cases} P_1 + P_2 + P_3 = 1 \\ 2^{-1-\lambda_1-\lambda_2} = 2^{-\lambda_1+2\lambda_2} \end{cases} \Leftrightarrow \begin{cases} P_1 + P_2 + P_3 = 1 \\ -1-\lambda_1-\lambda_2 = -\lambda_1+2\lambda_2 \end{cases}$$

$$\Leftrightarrow \begin{cases} 2^{-1-\lambda_1-\lambda_2} + 2^{-1-\lambda_1+2\lambda_2} + 2^{-1-\lambda_1} = 1 \\ \lambda_2 = -1/3 \end{cases} \Leftrightarrow \begin{cases} 2^{-1-\lambda_1+\frac{1}{3}} + 2^{-1-\lambda_1-\frac{2}{3}} + 2^{-1-\lambda_1} = 1 \\ \lambda_2 = -1/3 \end{cases}$$

$$\Leftrightarrow \begin{cases} 2^{-\lambda_1}(2^{-1+\frac{1}{3}} + 2^{-1-\frac{2}{3}} + 2^{-1}) = 1 \\ \lambda_2 = -1/3 \end{cases} \Leftrightarrow \begin{cases} \lambda_1 = log_2(2^{-1+\frac{1}{3}} + 2^{-1-\frac{2}{3}} + 2^{-1}) \\ \lambda_2 = -1/3 \end{cases}$$

Les valeurs des paramètres obtenus précédemment nous permettent de calculer les P_i :

$$\begin{aligned} P_1 &= \frac{2^{\frac{1}{3}}}{2(2^{-\frac{2}{3}} + 2^{-1-\frac{2}{3}} + 2^{-1})} && (15) \\ &= \frac{2^{\frac{1}{3}}}{2^{\frac{1}{3}} + 2^{-\frac{2}{3}} + 1} && (16) \\ &= \frac{1}{1 + 2^{-1} + 2^{-\frac{1}{3}}} && (17) \\ &= \frac{2}{2 + 1 + 2^{\frac{2}{3}}} && (18) \\ &= \frac{2}{3 + 2^{\frac{2}{3}}} && (19) \end{aligned}$$

$$\begin{aligned} P_2 &= \frac{1}{2}P_1 && (20) \\ &= \frac{1}{3 + 2^{\frac{2}{3}}} && (21) \end{aligned}$$

$$
\begin{aligned}
P_3 &= 1 - P_1 - P_2 && (22)\\
&= 1 - 3P_2 && (23)\\
&= \frac{1}{3+2^{\frac{2}{3}}} && (24)\\
&= \frac{3+2^{\frac{2}{3}}}{3+2^{\frac{2}{3}}} - \frac{3}{3+2^{\frac{2}{3}}} && (25)\\
&= \frac{2^{\frac{2}{3}}}{3+2^{\frac{2}{3}}} && (26)
\end{aligned}
$$

Vérifions désormais que l'optimum ainsi trouvé correspond aux paramètres qui maximisent l'entropie. Pour cela, nous allons calculer la Hessienne de notre lagrangien respectivement par rapport aux variables P_i (pour les 3 premières colonnes et lignes) :

$$
H = \begin{bmatrix} \frac{-1}{P_1} & 0 & 0 \\ 0 & \frac{-1}{P_2} & 0 \\ 0 & 0 & \frac{-1}{P_3} \end{bmatrix}
$$

Nous remarquons en calculant la Hessienne que la diagonale de la matrice ne contient que des valeurs propres négatives ou nulles puisque les P_i sont positifs, la Hessienne est donc **définie négative**.
Nous atteignons donc bien un maximum local aux points P_i satisfaisant les contraintes :

$$
\mathrm{C} : \begin{cases} P_1 + P_2 + P_3 = 1 \\ P_1 = 2P_2 \end{cases}
$$

tels que :

$$
P_1 = \frac{2}{3+2^{\frac{2}{3}}} \; ; \; P_2 = \frac{1}{3+2^{\frac{2}{3}}} \; ; \; P_3 = \frac{2^{\frac{2}{3}}}{3+2^{\frac{2}{3}}}
$$