Exercices théoriques

- Dériver les équations normales de la régression linéaire au moyen de l'algèbre linéaire

On cherche $\vec{x}$ qui solutionne $\boldsymbol{A}\vec{x} = \vec{b}$. Comme l'équation n'admet pas de solution pour $\vec{x}$, on cherche plutôt $\vec{\hat{x}}$ qui solutionne $\boldsymbol{A}\vec{\hat{x}} = \vec{p}$, avec l'objectif de minimiser l'erreur $\vec{e} = \vec{p} - \vec{b}$. Puisque $\vec{p} \subset C(\boldsymbol{A})$ et que $\vec{e}$ est minimal quand $\vec{e}$ est perpendiculaire à $C(\boldsymbol{A})$, on conclut, par le théorème fondamental de l'algèbre linéaire, que $\vec{e} \subset N(\boldsymbol{A}^{\mathrm{T}})$:

$$\begin{aligned} \boldsymbol{A}^{\mathrm{T}}\vec{e} &= \vec{0} \\ \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\vec{\hat{x}} - \vec{b}) &= \vec{0} \\ \vec{\hat{x}} &= (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\vec{b} \end{aligned}$$

- Démontrer que $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ est inversible ssi les colonnes de $\boldsymbol{A}$ sont indépendantes

Si $\boldsymbol{A}\vec{x} = \vec{0}$, alors $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\vec{x} = \vec{0}$; si $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\vec{x} = \vec{0}$ alors $\vec{x}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\vec{x} = 0$, $\|\boldsymbol{A}\vec{x}\| = 0$, $\vec{x} = \vec{0}$: $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ et $\boldsymbol{A}$ ont le même espace nul. Si les colonnes de $\boldsymbol{A}$ sont indépendantes, $N(\boldsymbol{A})$ ne contient que le vecteur nul, $N(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})$ ne contient que le vecteur nul, les colonnes de $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ sont indépendantes et donc $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ est inversible.

- Démontrer l'inégalité de Gibbs au moyen de l'inégalité de Jensen

$$\begin{aligned} \text{Inégalité de Gibbs : } & H(X) \leqslant \log|\mathcal{A}_X| \text{ où } |\mathcal{A}_X| \text{ est le cardinal de } \mathcal{A}_X \\ \text{ou similairement } & -H(X) \geqslant -\log|\mathcal{A}_X| \\ -H(X) &= \sum_{x \subset \mathcal{A}_X} p(x)(-1)\log\frac{1}{p(x)} \\ &= E\left[-\log\frac{1}{p(X)}\right] \\ -\log() & \text{ est une fonction convexe} \\ \text{par l'inégalité de Jensen, } & E[f(X)] \geqslant f(E[X]) : \\ -H(X) &\geqslant -\log(E\left[\frac{1}{p(X)}\right]) = -\log|\mathcal{A}_X| \end{aligned}$$

- Écrire les trois premiers termes de l'expansion en série de Taylor de $f(\vec{x})$ où $\vec{x} \in \mathbb{R}^n$

$$f(\vec{x}) = f(\vec{x}_0) + (\vec{x} - \vec{x}_0)^{\mathrm{T}}\,\nabla f(\vec{x})\bigg|_{\vec{x}_0} + \frac{1}{2}(\vec{x} - \vec{x}_0)^{\mathrm{T}}\,H_f(\vec{x})\bigg|_{\vec{x}_0}(\vec{x} - \vec{x}_0) + \ldots$$

$$\begin{aligned} \nabla f(\vec{x})\bigg|_{\vec{x}_0} &: \text{ gradient de } f(\vec{x}) \text{ évalué à } \vec{x}_0 \\ H_f(\vec{x})\bigg|_{\vec{x}_0} &: \text{ matrice hessienne de } f(\vec{x}) \text{ évaluée à } \vec{x}_0 \end{aligned}$$

- Exprimer $\Delta^2 = (\vec{x}-\vec{\mu})^{\mathrm{T}}\Sigma^{-1}(\vec{x}-\vec{\mu})$ en termes des D valeurs propres de Σ

$$\begin{aligned}
\text{Soit } \vec{u}_i \text{ tel que } \Sigma\, \vec{u}_i &= \lambda_i \vec{u}_i \\
\Sigma &= \sum_{i=1}^{D} \lambda_i \vec{u}_i \vec{u}_i^{\mathrm{T}} \\
\Sigma^{-1} &= \sum_{i=1}^{D} \frac{1}{\lambda_i} \vec{u}_i \vec{u}_i^{\mathrm{T}} \\
\text{Notons } y_i &= \vec{u}_i^{\mathrm{T}}(\vec{x}-\vec{\mu}) \\
&= (\vec{x}-\vec{\mu})^{\mathrm{T}}\vec{u}_i \\
\Delta^2 &= (\vec{x}-\vec{\mu})^{\mathrm{T}}\Sigma^{-1}(\vec{x}-\vec{\mu}) \\
&= (\vec{x}-\vec{\mu})^{\mathrm{T}} \left(\sum_{i=1}^{D} \frac{1}{\lambda_i} \vec{u}_i \vec{u}_i^{\mathrm{T}} \right) (\vec{x}-\vec{\mu}) \\
&= \sum_{i=1}^{D} \frac{1}{\lambda_i} (\vec{x}-\vec{\mu})^{\mathrm{T}} \vec{u}_i \vec{u}_i^{\mathrm{T}} (\vec{x}-\vec{\mu}) \\
&= \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i}
\end{aligned}$$

- Pour le modèle de la régression linéaire, énoncer les équations pour : la distribution de la variable cible ; les distributions *a priori* et *a posteriori* du vecteur des pondérations ; la distribution prédictive de la variable cible dans le cas où $\boldsymbol{S}_0 = \alpha \boldsymbol{I}$ et $\vec{m}_0 = 0$

$$\begin{aligned}
p(t\,|\vec{x},\vec{w}) &= \mathcal{N}(t\,|y(\vec{x},\vec{w}),\beta^{-1}) \\
p(\vec{w}) &= \mathcal{N}(\vec{w}\,|\vec{m}_0,\boldsymbol{S}_0) \\
p(\vec{w}\,|\vec{\mathsf{t}},\mathbf{X}) &= \mathcal{N}(\vec{w}\,|\vec{m}_N,\boldsymbol{S}_N) \\
\text{où } \boldsymbol{S}_N^{-1} &= \boldsymbol{S}_0^{-1} + \beta \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} \\
\vec{m}_N &= \boldsymbol{S}_N(\boldsymbol{S}_0^{-1}\vec{m}_0 + \beta\boldsymbol{\Phi}^{\mathrm{T}}\vec{\mathsf{t}}) \\
p(t\,|\vec{x},\vec{\mathsf{t}},\mathbf{X}) &= \mathcal{N}(t\,|y(\vec{x},\vec{m}_N),\sigma_N^2(\vec{x})) \\
\text{où } \sigma_N^2(\vec{x}) &= \frac{1}{\beta} + \vec{\phi}(\vec{x})^{\mathrm{T}}\boldsymbol{S}_N\vec{\phi}(\vec{x})
\end{aligned}$$

- Établir la correspondance entre le paramètre de régularisation dans la fonction d'erreur des moindres carrés et le paramètre de précision de la distribution a priori du vecteur des pondérations

$$
\begin{aligned}
p(\vec{w}\,|\vec{\mathsf{t}}) &\propto \prod_{n=1}^{N} p(t_n\,|\vec{x}_n, \vec{w})\; p(\vec{w}) \\
&= \prod_{n=1}^{N} \mathcal{N}(t_n\,|y_n, \beta^{-1})\; \mathcal{N}(\vec{w}\,|0, \alpha^{-1}) \text{ où } y_n = y(\vec{x}_n, \vec{w}) \\
&= k \prod_{n=1}^{N} \exp(-\frac{\beta}{2}(t_n - y_n)^2)\; \exp(-\frac{\alpha}{2}\vec{w}^{\mathrm{T}}\vec{w}) \\
E(\vec{w}) = -\ln p(\vec{w}\,|\vec{\mathsf{t}}) &= \frac{\beta}{2}\sum_{n=1}^{N}(t_n - y_n)^2 + \frac{\alpha}{2}\vec{w}^{\mathrm{T}}\vec{w} + K \\
E(\vec{w}) \text{ est miminum quand } &\frac{\beta}{2}\sum_{n=1}^{N}(t_n - y_n)^2 + \frac{\alpha}{2}\vec{w}^{\mathrm{T}}\vec{w} \text{ est minimum} \\
\text{quand } &\frac{1}{2}\sum_{n=1}^{N}(t_n - y_n)^2 + \frac{1}{2}\frac{\alpha}{\beta}\vec{w}^{\mathrm{T}}\vec{w} \text{ est minimum}
\end{aligned}
$$

Le paramètre de régularisation λ est donc équivalent à $\frac{\alpha}{\beta}$

- Au moyen d'une notation sans omission, formuler la triple intégrale de la distribution *prédictive* obtenue dans le cadre d'un traitement bayésien complet (portant sur $\vec{w}$, α, β) du modèle de régression linéaire

$$p(t\,|\vec{x}, \vec{\mathsf{t}}, \mathbf{X}) = \iiint p(t\,|\vec{x}, \vec{w}, \beta)\; p(\vec{w}\,|\vec{\mathsf{t}}, \mathbf{X}, \alpha)\; p(\alpha, \beta\,|\vec{\mathsf{t}}, \mathbf{X})\; d\vec{w}\; d\alpha\; d\beta$$

- Énoncer la relation qui lie une sortie y_k au vecteur d'entrée $\vec{x}$ (à D composantes) dans un réseau neuronal à une couche cachée composée de M unités

$$y_k = f(a_k) = f\left(\sum_{j=0}^{M} w_{kj}^{(2)} \sigma\left(\sum_{i=0}^{D} w_{ji}^{(1)} x_i\right)\right)$$

- Dans un réseau neuronal, pour les modèle de régression, de classification binaire et de classification multinomiale, énoncer : le modèle pour la variable cible ; la formulation de y_k d'une unité de sortie en fonction de son activation a_k ; la fonction d'erreur

$$\begin{aligned}
p(t\,|\vec{x},\vec{w}) &= \mathcal{N}(t\,|y(\vec{x},\vec{w}),\beta^{-1}) \\
y(\vec{x},\vec{w}) &= f(a) = a \\
E(\vec{w}) &= \frac{\beta}{2}\sum_{n=1}^{N}\{y(\vec{x}_n,\vec{w})-t_n\}^2 \\
p(t\,|\vec{x},\vec{w}) &= y(\vec{x},\vec{w})^t\{1-y(\vec{x},\vec{w})\}^{1-t} \\
y(\vec{x},\vec{w}) &= f(a) = \sigma(a) = \frac{1}{1+exp(-a)} \\
E(\vec{w}) &= -\sum_{n=1}^{N} t_n ln\{y(\vec{x}_n,\vec{w})\} + (1-t_n)ln\{1-y(\vec{x}_n,\vec{w})\} \\
p(\vec{t}|\vec{x},\vec{w}) &= \prod_{k=1}^{K}\{y_k(\vec{x},\vec{w})\}^{t_k} \\
y_k(\vec{x},\vec{w}) &= f(a_k) = \frac{exp(a_k)}{\sum_j exp(a_j)} \\
E(\vec{w}) &= -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\, ln\{y_k(\vec{x}_n,\vec{w})\}
\end{aligned}$$

où t_{nk} est la k^e composante de la cible $\vec{t}$ de la n^e observation

- Pour un réseau neuronal à K sorties, avec une couche cachée composée de M unités, dériver les équations de $\nabla_{\vec{w}} E_n$ calculé par rétropropagation

$$\begin{aligned}
\frac{\partial E_n}{\partial w_{ji}^{(l)}} = \frac{\partial E_n}{\partial a_j^{(l)}}\frac{\partial a_j^{(l)}}{\partial w_{ji}^{(l)}} &= \frac{\partial E_n}{\partial a_j^{(l)}} z_i^{(l-1)} \text{ où } z_i^{(1)} = h(a_i^{(1)}), z_i^{(0)} = x_i \\
&= \delta_j^{(l)} z_i^{(l-1)} \text{ où } \delta_j^{(l)} \equiv \frac{\partial E_n}{\partial a_j^{(l)}}
\end{aligned}$$

$\delta_k^{(2)}$ se calcule directement à partir de la fonction d'erreur.

Si la fonction f dans $y_k = f(a_k)$ est l'inverse de la fonction de liaison canonique de la distribution de $t\,|\vec{x}$:

$$\delta_k^{(2)} = y_k - t_k$$

Quelle que soit la forme de la fonction f : $\delta_j^{(1)} = \frac{\partial E_n}{\partial a_j^{(1)}} = \sum_{k=1}^{K}\frac{\partial E_n}{\partial a_k^{(2)}}\frac{\partial a_k^{(2)}}{\partial a_j^{(1)}}$

$$= \sum_{k=1}^{K}\delta_k^{(2)}\frac{\partial a_k^{(2)}}{\partial a_j^{(1)}}$$

De la somme $a_k^{(2)} = w_{k0}^{(2)}h(a_0^{(1)}) + w_{k1}^{(2)}h(a_1^{(1)}) + w_{k2}^{(2)}h(a_2^{(1)}) + ...$

un seul terme lie $a_k^{(2)}$ et $a_j^{(1)}$: $w_{kj}^{(2)}h(a_j^{(1)})$

$$\frac{\partial a_k^{(2)}}{\partial a_j^{(1)}} = w_{kj}^{(2)}\left.\frac{dh(a)}{da}\right|_{(a_j^{(1)})}$$

$$\delta_j^{(1)} = h'(a_j^{(1)})\sum_{k=1}^{K} w_{kj}^{(2)}\delta_k^{(2)}$$

- Résumer la *représentation noyau* du modèle de régression linéaire régularisé (Ridge) en indiquant : la définition reliant le nouveau vecteur des paramètres $\vec{a}$ et l'ancien vecteur $\vec{w}$; la fonction noyau ; la relation entre $\vec{w}$ et $\vec{a}$ qui produit une valeur minimum pour la fonction d'erreur ; la solution de $\vec{a}$ qui minimise la fonction d'erreur ; la solution de régression $y(\vec{x})$ en termes de la fonction noyau et du vecteur des paramètres $\vec{a}$

$$
\begin{aligned}
\mathbf{\Phi}\ \vec{w} &= \vec{\mathsf{t}} - \lambda \vec{a} \\
k(\vec{x}_n, \vec{x}_m) &= \vec{\phi}(\vec{x}_n)^{\mathrm{T}} \vec{\phi}(\vec{x}_m) \\
\vec{w} &= \mathbf{\Phi}^{\mathrm{T}}\ \vec{a} \\
\vec{a} &= (\boldsymbol{K} + \lambda \boldsymbol{I})^{-1} \vec{\mathsf{t}} \\
\text{où } \boldsymbol{K} &= \mathbf{\Phi}\ \mathbf{\Phi}^{\mathrm{T}} \\
&= K_{nm} = k(\vec{x}_n, \vec{x}_m) \\
y(\vec{x}) &= \vec{a}^{\mathrm{T}} \vec{k}(\vec{x}) \\
\text{où } \vec{k}(\vec{x}) &= \mathbf{\Phi}\ \vec{\phi}(\vec{x}) \\
k_n(\vec{x}) &= k(\vec{x}_n, \vec{x})
\end{aligned}
$$

- Donner une formulation complète du modèle de Nadaraya-Watson

Au moyen d'une densité de probabibilité modélisée par un estimateur de Parzen :

$$p(\vec{x}, t) = \frac{1}{N} \sum_{n=1}^{N} f(\vec{x} - \vec{x}_n, t - t_n)$$

on déduit la fonction de régression $y(\vec{x}) = \mathbb{E}[t\,|\vec{x}]$

$$
\begin{aligned}
y(\vec{x}) &= \sum_{n=1}^{N} k(\vec{x}, \vec{x_n}) t_n \\
\text{où } k(\vec{x}, \vec{x_n}) &= \frac{g(\vec{x} - \vec{x}_n)}{\sum_{m=1}^{N} g(\vec{x} - \vec{x}_m)} \\
g(\vec{x}) &\equiv \int_{-\infty}^{\infty} f(\vec{x}, t)\ dt
\end{aligned}
$$