Sélection d'*a priori* et géométrie de l'information

Hichem Snoussi et Ali Mohammad-Djafari
Laboratoire des Signaux et Systèmes (L2S)
Supélec, Plateau de Moulon,
91192 Gif-sur-Yvette Cedex, France
snoussi@lss.supelec.fr

Résumé

Dans cette contribution, nous étudions le problème de la sélection de distribution *a priori* dans le contexte de la théorie bayésienne. La litérature sur le sujet est abondante et le problème est loin d'être définitivement résolu [1]. Nous revisitons cette problématique avec les outils de la géométrie différentielle pour proposer une construction de l'*a priori* dans le cadre de la théorie bayésienne de décision. Les résultats sont illustrés avec un exemple de classification.

1 Introduction

Une modélisation physique peut être représentée par une machine d'apprentissage (learning machine) liant les entrées $\boldsymbol{x}$ aux sorties $\boldsymbol{y}$ (voir figure 1). La complexité du méchanisme physique gérant la relation entrées sorties ou le manque d'informations rendent difficiles la prédiction des sorties sachant les entrées (problème direct) ou l'estimation des entrées sachant les sorties (problème inverse). Dans le cas où un modèle paramétrique direct $p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})$ est supposé connu, on peut utiliser le maximum de vraisemblance pour estimer le paramètre inconnu $\boldsymbol{\theta}$ ou plus généralement utiliser l'approche bayésienne pour incorporer une information *a priori* $p(\boldsymbol{\theta})$ et former la distribution jointe $p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta})$ contenant toute l'information disponible sur le système considéré. On peut alors effectuer une inférence optimale sur une grandeur particulière qui intéresse l'utilisateur.

Nous supposons dans ce papier que nous disposons de données d'apprentissage $\boldsymbol{z} = (\boldsymbol{x}_{1..N}, \boldsymbol{y}_{1..N})$ et d'une information physique sur la transformation entrée sortie qui consiste en un modèle paramétrique $\mathcal{Q} = \{P(\boldsymbol{z} \mid \boldsymbol{\theta})\}$ de distributions de probabilité. L'objectif de l'apprentissage statistique est de construire une application τ qui associe un jeu de données $\boldsymbol{z} \in \mathcal{Z}$ à une densité prédictive $p \in \mathcal{Q}$:

$$\begin{array}{rccc} \tau: & \mathcal{Z} & \longrightarrow & \mathcal{Q} \\ & \boldsymbol{z} & \mapsto & q = \tau(\boldsymbol{z}) \end{array}$$

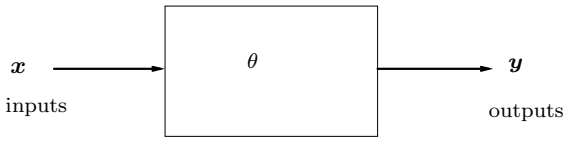

Figure 1. Machine d'apprentissage statistique

L'apprentissage statistique bayésien aboutit à une solution qui dépend de la distribution *a priori* de la densité inconnue p. Dans le cas paramétrique, ceci est équivalent à la distribution *a priori* $\Pi(\boldsymbol{\theta})$ du paramètre $\boldsymbol{\theta}$. Trouver une expression générale pour $\Pi(\boldsymbol{\theta})$ qui reflète la relation entre le modèle restreint et le plus petit ensemble d'ignorance le contenant est la contribution principale de ce papier. Nous montrons que l'expression de l'*a priori* construit (que nous notons δ-*a priori*)

dépend de la géométrie choisie (choix subjectif) sur l'ensemble des mesures de probabilités. L'*a priori* entropique [Rodriguez 1991, [2]] et l'*a priori* conjugué des familles exponentielles sont des cas spécifiques des δ-*a priori* pour $\delta = 1$ et $\delta = 0$ respectivement.

La section I est un rappel des concepts de la théorie bayésienne de prédiction dans le cadre de la géométrie différentielle. Dans la section II, nous développons un critère de sélection d'*a priori* ainsi que la forme explicite de la solution de ce critère. Dans la section III, nous étudions le cas particulier des familles δ-plates. Nous illustrons dans la section IV les résultats dans un problème de classification non supervisée.

2 Apprentissage et géométrie différentielle

2.1 Masse et géométrie

L'apprentissage statistique consiste à construire une application τ qui à chaque jeu de données d'apprentissage $\boldsymbol{z}$ associe une densité prédictive $q = \tau(\boldsymbol{z}) \in \mathcal{Q} \subset \mathcal{P} = \{p \mid \int p = 1\}$. Le sous-ensemble $\mathcal{Q}$ est en général un modèle paramétrique $\mathcal{Q} = \{P(\boldsymbol{z} \,|\, \boldsymbol{\theta})\}$. Ainsi, l'espace d'arrivée de la fonction τ est un sous espace de distributions qu'on doit munir, au moins dans ce travail, d'une masse (champ scalaire) et d'une géométrie. La masse est définie par une distribution *a priori* $\Pi(p)$ sur l'espace $\mathcal{P}$ avant la collecte des données $\boldsymbol{z}$ et elle est modifiée selon la règle de Bayes après l'observation des $\boldsymbol{z}$ pour former la distribution *a posteriori* (voir figure 2) :

$$P(p \mid \boldsymbol{z}) \propto P(\boldsymbol{z} \mid p)\, \Pi(p)$$

où $P(\boldsymbol{z} \mid p)$ est $p(\boldsymbol{z})$ la vraisemblance de la distribution p selon laquelle les données $\boldsymbol{z}$ sont générées.

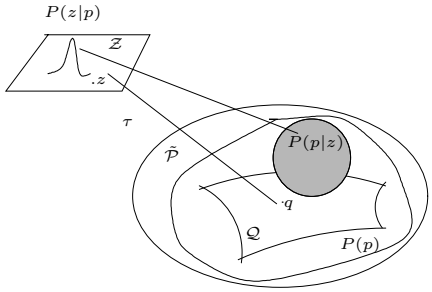

Figure 2. La distribution *a posteriori* est proportionelle au produit de la masse *a priori* et de la vraisemblance .

La géométrie peut être définie par la δ-divergence D_δ :

$$D_\delta(p, q) = \frac{\int p}{1-\delta} + \frac{\int q}{\delta} - \frac{\int p^\delta q^{1-\delta}}{\delta(1-\delta)}$$

qui est une mesure de divergence invariante par rapport au choix de la mesure dominante de l'espace des données $\mathcal{Z}$ et par rapport à la paramétrisation du modèle paramétrique $\mathcal{Q}$. Il est démontré dans [Amari 1985, [3]] que, dans la variété différentielle paramétrique $\mathcal{Q}$, la δ-divergence induit une structure duale $(g, \nabla^\delta, \nabla^{1-\delta})$. g est la métrique de Fisher, ∇^δ est la δ-connection affine avec les symboles de Christoffel $\Gamma^\delta_{ij,k}$ et $\nabla^* = \nabla^{1-\delta}$ est sa connection duale :

$$\begin{cases} g_{ij} &= E_{\boldsymbol{\theta}}\left[\partial_i l(\boldsymbol{\theta})\, \partial_j l(\boldsymbol{\theta})\right] \\ \Gamma^\delta_{ij,k} &= E_{\boldsymbol{\theta}}\left[(\partial_i \partial_j l(\boldsymbol{\theta}) + \delta \partial_i l(\boldsymbol{\theta}) \partial_j l(\boldsymbol{\theta}))\, \partial_k l(\boldsymbol{\theta})\right] \end{cases}$$

La variété paramétrique $\mathcal{Q}$ est δ-plate si et seulement si il existe une paramétrisation $[\theta_i]$ tels que les symboles de Christoffel sont nuls : $\Gamma^\delta_{ij,k}(\boldsymbol{\theta}) = 0$. Les coordonnées $[\theta_i]$ vérifiant cette propriété

sont appelés les coordonnées affines. Un autre système de coordonnées $[\theta_i^{'}]$ ayant les coefficients de la connection nuls est nécesssairement relié au système $[\theta_i]$ par une transformation affine (Il exite une matrice $(n \times n)$ $\boldsymbol{A}$ et un vecteur $\boldsymbol{b}$ tel que $\boldsymbol{\theta}^{'} = \boldsymbol{A\theta} + \boldsymbol{b}$).

Toutes les notions introduites ci-dessus peuvent être étendues au cas non paramétrique en remplaçant les dérivées partielles avec les dérivées de Fréchet. Immergeant le modèle $\mathcal{Q}$ dans l'espace des mesures positives $\tilde{\mathcal{P}}$ [Zhu et al. 1995, [4, 5]] non seulement l'espace des mesures de probabilités $\mathcal{P}$, beaucoup de résultats sont dérivés d'une manière plus simple pour la raison principale que $\tilde{\mathcal{P}}$ est δ-plat et δ-convexe $\forall$ δ dans $[0\,,\,1]$ tandis que $\mathcal{P}$ est δ-plat uniquement pour $\delta = \{0, 1\}$ et δ-convexe pour $\delta = 1$. Pour simplifier les notations, nous utilisons les δ-coordonnées $\overset{\delta}{l}$ du point $p \in \tilde{\mathcal{P}}$ définies comme :

$$\overset{\delta}{l}(p) = p^{\delta}/\delta$$

Une courbe liant 2 points a et b est une fonction $\gamma : [0\,,\,1] \longrightarrow \tilde{\mathcal{P}}$, tel que $\gamma(0) = a$ et $\gamma(1) = b$. Une courbe est dite δ-géodésique dans la δ-géométrie si c'est une ligne droite dans les δ-coordonnées.

2.2 Apprentissage bayésien

L'erreur commise par une fonction d'apprentissage τ dans une δ-géométrie fixée peut être quantifiée par la δ-divergence $D_\delta(p, \tau(\boldsymbol{z}))$ entre la vraie probabilité inconnue p et la décision $\tau(\boldsymbol{z})$. Cette divergence est d'abord moyennée par rapport à tous les jeux de données possibles $\boldsymbol{z}$ puis moyennée par rapport à la vraie distribution inconnue p donnant l'erreur de généralisation $E(\tau)$:

$$E_\delta(\tau) = \int_p P(p) \int_{\boldsymbol{z}} P(\boldsymbol{z} \,|\, p) D_\delta(p, \tau(\boldsymbol{z}))$$

La décision optimale τ_δ est donc le minimizeur de l'erreur de généralisation :

$$\tau_\delta = \arg\min_{\tau} \{E_\delta(\tau)\}$$

La cohérence de l'apprentissage bayésien est montré dans [Zhu et al. 1995, [4, 5]] signifiant que l'estimateur optimal τ_δ comme fonction de $\boldsymbol{z}$ peut être calculé en tout point et donc une expression générale de τ_δ n'est pas nécessaire :

$$\hat{p}(\boldsymbol{z}) = \tau_\delta(\boldsymbol{z}) = \arg\min_{q} \int_p P(p \,|\, \boldsymbol{z}) D_\delta(p, q) \tag{1}$$

La solution de (1) est direct par calcul variationel et donne :

$$\hat{p}^\delta = \int p^\delta P(p \,|\, z)$$

Cette solution n'est autre que le centre de gravité de l'ensemble $\tilde{\mathcal{P}}$ muni de la masse $P(p \,|\, \boldsymbol{z})$, la distribution *a posteriori* de p et de la δ-géométrie induite par la δ-divergence D_δ. Nous notons ici l'analogie avec la mécanique statique et l'importance de la géométrie choisie sur l'espace des distributions. L'espace étendue des mesures positives finies $\tilde{\mathcal{P}}$ est δ-convexe et donc indépendemment de la distribution *a posteriori* $P(p \,|\, \boldsymbol{z})$ la solution $\hat{p}$ appartient à $\tilde{\mathcal{P}}$ $\forall$ $\delta \in [0\,,\,1]$.

2.3 Restriction paramétrique

Dans les situations pratiques, on restreint l'espace des décisions à sous ensemble paramétrique $\mathcal{Q} \in \tilde{\mathcal{P}}$ qu'on suppose différentiable. La variété $\mathcal{Q}$ est donc paramétrée par un système de coordonnées $[\theta_i]_{i=1}^n$ avec n la dimension de la variété. Ce modèle n'est pas cependant déconnecté d'éventuelles manipulations non paramétriques et comme nous allons montrer la décision *a priori* ou la décision *a posteriori* peuvent être localisées à l'extérieur de ce modèle $\mathcal{Q}$.

La densité prédictive optimale est le minimiseur de l'erreur de généralisation :

$$\hat{q}(\boldsymbol{z}) = \tau_\delta(\boldsymbol{z}) = \arg\min_{q \in \mathcal{Q}} \int_{p \in \mathcal{Q}} P(p \mid \boldsymbol{z}) D_\delta(p, q) = \arg\min_{q \in \mathcal{Q}} \int_{\boldsymbol{\theta}} P(\boldsymbol{\theta} \mid \boldsymbol{z}) D_\delta(p_\theta, q) d\boldsymbol{\theta} \tag{2}$$

La solution est la δ-projection du barycentre $\hat{p}$ de ($\mathcal{Q}$, $P(\boldsymbol{\theta} \mid \boldsymbol{z})$, D_δ) sur le modèle $\mathcal{Q}$ (voir figure 4).

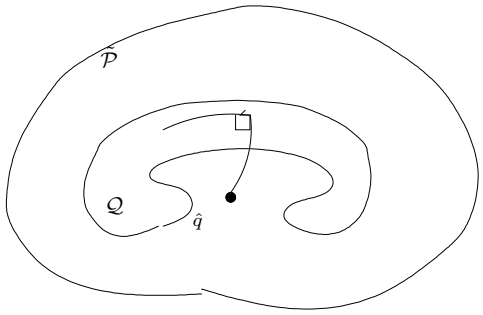

Figure 4. Solution : projection du barycentre
sur le modèle paramétrique

La manipulation de ces concepts est très abstraite dans le cas général. Cependant, dans le cas des familles δ-plates la solution a une forme explicite en fonction des paramétres de la variété différentielle $\mathcal{Q}$.

3 Sélection d'*a priori*

Le problème de la sélection d'*a priori* peut se résumer comme suit :

`Comment construire un` *`a priori`* $P(p)$ `en respectant la règle suivante : Exploiter les informations disponibles` *`a priori`* `sans rajouter des informations subjectives.`

On note que ce critère est une sorte de compromis entre un comportement souhaité (comportement de référence) et la contrainte d'uniformité. Nous insistons ici sur le fait que la sélection doit être effectuée avant la collecte des données $\boldsymbol{z}$, sinon la cohérence de la règle de Bayes n'est plus vérifiée.

Dans un cadre de décision, le comportement de référence peut être décrit comme suit : Avant la collecte des données d'apprentissage $\boldsymbol{z}$, fournir une distribution de référence p_0 comme décision avec un degré de confiance γ_e. La distribution de référence peut être fournie par un expert ou par notre expérience antérieure. On a alors le problème inverse du problème statistique d'apprentissage. Avant, la distribution *a posteriori* (masse) est fixée et on doit trouver la distribution optimale (barycentre) tandis que maintenant la décision optimale est fixée et on doit chercher la masse *a priori* optimale avec la contrainte d'uniformité. Pour conserver les notions usuelles de dérivation et d'intégration, nous supposons que notre objectif est de trouver l'*a priori* sur le modèle paramétrique $\mathcal{Q} = \{q_\theta \mid \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^n\}$.

Le critère est construit comme une somme pondérée de l'erreur de généralisation et de la divergence entre l'*a priori* recherché et l'*a priori* de Jeffreys (racine carré du déterminant de la matrice d'information de Fisher [6]) représentant l'uniformité. On note qu'on manipule deux espaces différents : l'espace $\tilde{\mathcal{P}}$ des mesures finies et l'espace des distributions *a priori*. Ces deux espaces sont distincts donc on peut choisir deux géométries différentes. Par exemple, si on choisit la δ-géométrie pour l'espace $\tilde{\mathcal{P}}$ et la 1-géométrie pour l'espace des *a priori*, nous obtenons le critère suivant à minimiser :

$$J(\Pi) = \gamma_e \int \Pi(\boldsymbol{\theta}) D_\delta(p_\theta, p_0) d\,\boldsymbol{\theta} + \gamma_u \int \Pi(\boldsymbol{\theta}) \log \Pi(\boldsymbol{\theta}) / \sqrt{g(\boldsymbol{\theta})} d\,\boldsymbol{\theta} \tag{3}$$

où γ_e est le degré de confiance sur la distribution de référence p_0 et γ_u est le degré d'uniformité. Considérés indépendemment, ces deux coefficients ne sont pas significatifs. Cependant, leur rapport joue un rôle important dans la suite. Le critère (3) peut être ré-écrit comme suit :

$$\begin{cases} J(\Pi) = \gamma_e E(\tau_0) + \gamma_u \int \Pi(\boldsymbol{\theta}) \log \Pi(\boldsymbol{\theta}) / \sqrt{g(\boldsymbol{\theta})} d\,\boldsymbol{\theta} \\ \\ \frac{\partial \tau_0}{\partial \boldsymbol{z}} = 0 \end{cases}$$

où $E(\tau_0)$ est l'erreur de généralisation pour la fonction de décision constante τ_0. Par un calcul variationel, on obtient la solution de la minimisation de la fonction (3) :

$$\Pi(\boldsymbol{\theta}) \propto e^{-\frac{\gamma_e}{\gamma_u} D_\delta(p_\theta, p_0)} \sqrt{g(\boldsymbol{\theta})} \tag{4}$$

On note que si $\delta = 1$ alors le critère (3) est la divergence de kullback-Leibler entre les distributions conjointes des données et des paramètres comme dans [Rodriguez 1991, [2]] et si $\delta = 0$ on obtient l'équivalent de l'*a priori* conjugué pour les familles exponentielles (voir exemple dans la section IV). Quand la valeur du rapport γ_e/γ_u tend vers 0, on obtient l'*a priori* de Jeffreys et quand ce rapport tend vers ∞ on obtient la fonction dirac concentrée sur p_0.

4 Familles δ-plates

Nous étudions le cas particulier des familles δ-plates. $\mathcal{Q}$ est une variété différentielle δ-plate si et seulement si il existe un système de coordonnées $[\theta_i]$ tels que les coefficients de la connection $\Gamma_\delta(\boldsymbol{\theta})$ sont identiquement nuls. $[\theta_i]$ est ainsi un système de coordonnée affine. Il est montré que la δ-platitude est équivalente à la $(1-\delta)$-platitude. Il existe alors un système de coordonnées duale $[\eta_i]$ tels que $\Gamma_{1-\delta}(\boldsymbol{\eta}) = 0$. Une des propriétés intéressantes de la platitude est que la δ-divergence D_δ possède une forme simple en fonction de $\boldsymbol{\theta}$ et $\boldsymbol{\eta}$:

$$D_\delta(p, q) = \psi(p) + \phi(q) - \theta_i(p)\eta_i(q)$$

où ψ et ϕ sont les potentiels duales de la transformation de Legendre :

$$\begin{cases} \frac{\partial \eta_j}{\partial \theta_i} = g_{ij} \quad \frac{\partial \theta_i}{\partial \eta_j} = g_{ij}^{-1} \\ \\ \partial_i \psi = \eta_i \quad \partial_i \phi = \theta_i \end{cases}$$

Par exemple, la famille exponentielle est 1-plate avec les paramètres canoniques comme coordonnées affines. $\tilde{\mathcal{P}} = \{p, \int p < \infty\}$ est δ-plat quelque soit $\delta \in [0\,,\,1]$.

Dans ce cas, l'expression de la distribution δ-*a priori* Π_δ (4) est la suivante :

$$\Pi(\boldsymbol{\theta}) \propto e^{-\frac{\gamma_e}{\gamma_u}\left(\psi(\boldsymbol{\theta}) - \theta_i\, \eta_i^0\right)} \sqrt{g(\theta)}$$

où $[\theta_i^0]$ et $[\eta_i^0]$ sont les coordonnées duales affines de p_0.

On a donc une expression explicite de l'*a priori*.

Dans le cas euclidien, correspondant à une connection auto-duale : $\nabla = \nabla^*$ ou d'une manière équivalente à l'égalité des coordonnées affines $[\theta_i] = [\eta_i]$, la distribution δ-*a priori* est gaussienne de moyenne $\boldsymbol{\theta}_0$ et de précision $2\,\frac{\gamma_e}{\gamma_u}$:

$$\Pi(\theta) \propto e^{-\frac{\gamma_e}{\gamma_u}\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2}$$

5 Exemples

Dans cette section, nous donnons l'expression des δ-*a priori* dans un problème de classification. Supposons que les T données observées $\boldsymbol{x}_i$, $i = 1..T$ ($\boldsymbol{x}_i \in \mathbb{R}^n$) sont générées selon un mélange de gaussiennes multivariables [7] (un mélange de distributions appartenant à des familles 0-plates) :

$$p(\boldsymbol{x}_i) = \sum_{k=1}^{K} w_k \, \mathcal{N}(\boldsymbol{x}_i \,;\, \boldsymbol{m}_k, \boldsymbol{R}_k) \tag{5}$$

où w_k, $\boldsymbol{m}_k$ et $\boldsymbol{R}_k$ sont le poids, la moyenne et la covariance de la classe k. Ceci peut être interprété par un problème à données incomplètes où les données manquantes sont les étiquettes $(c_i)_{i=1..T}$ des différentes classes. Le mélange est donc une marginalisation par rapport à c_i :

$$p(\boldsymbol{x}_i) = \sum_{c_i} p(c_i) \, \mathcal{N}(\boldsymbol{x}_i \,|\, c_i, \boldsymbol{\theta})$$

où $\boldsymbol{\theta}$ est l'ensemble des moyennes et des covariances. L'ensemble des paramètres inconnus est alors $\boldsymbol{\eta} = (\boldsymbol{\theta}, \boldsymbol{w})$. L'expression de l'*a priori* est dérivée pour les vraisemblances complétées (le calcul des divergences et de la matrice de Fisher est basé sur $p(\boldsymbol{x}, c \,|\, \boldsymbol{\eta})$). Les détails du calcul ainsi qu'un autre exemple en séparation de sources peuvent être consultés dans [Snoussi 2002 [8]]. Nous donnons ici directement les expressions des deux *a priori* Π_0 et Π_1 :

- $\delta = 0$:

$$\begin{aligned} \Pi_0(\boldsymbol{\eta}_i) &\propto \exp\left[-\tfrac{\gamma_e}{\gamma_u}\left(w_i^0 D_0(\mathcal{N}_i : \mathcal{N}_i^0) + w_i^0 \log \tfrac{w_i^0}{w_i}\right)\right] \sqrt{|g_i(\boldsymbol{\eta}_i)|} \\ &\propto \mathcal{N}\left(\boldsymbol{m}_i \,;\, \boldsymbol{m}_0, \tfrac{\boldsymbol{R}_i}{\alpha\, w_i^0}\right) \mathcal{W}_n\left(\boldsymbol{R}_i^{-1} \,;\, \nu_0, \boldsymbol{R}_0^{-1}\right) w_i^{\beta_0} \end{aligned} \tag{6}$$

avec,

$$\alpha = \tfrac{\gamma_e}{\gamma_u}, \quad \nu_0 = \alpha\, w_i^0, \quad \beta_0 = \alpha w_i^0 + \tfrac{n^2+n-1}{2}$$

$\mathcal{W}_n$ est la distribution Wishart d'une matrice $n \times n$:

$$\mathcal{W}_n(\boldsymbol{R} \,;\, \nu, \boldsymbol{\Sigma}) \propto |\boldsymbol{R}|^{\frac{\nu-(n+1)}{2}} \exp\left[-\frac{\nu}{2}\mathrm{Tr}\left(\boldsymbol{R}\boldsymbol{\Sigma}^{-1}\right)\right]$$

Le 0-*a priori* est donc Normal Inverse Wishart pour la moyenne et la covariance $(\boldsymbol{m}_i, \boldsymbol{R}_i)$ et Dirichlet pour les poids w_i et possède ainsi une forme équivalente de l'*a priori* **conjugué**.

- $\delta = 1$:

$$\begin{aligned} \Pi_1(\boldsymbol{\eta}_i) &\propto \exp\left[-\tfrac{\gamma_e}{\gamma_u}\left(w_i D_1(\mathcal{N}_i : \mathcal{N}_i^0) + w_i \log \tfrac{w_i}{w_i^0}\right)\right] \sqrt{|g_i(\boldsymbol{\eta}_i)|} \\ &\propto \mathcal{N}\left(\boldsymbol{m}_i \,;\, \boldsymbol{m}_0, \tfrac{\boldsymbol{R}_i}{\alpha\, w_i}\right) \mathcal{W}_n\left(\boldsymbol{R}_i \,;\, \alpha w_i - 1, \tfrac{\alpha w_i - 1}{\alpha w_i}\boldsymbol{R}_0\right) \\ &\quad w_i^{\frac{n^2+n-1}{2} - (1+\frac{n}{2})\alpha w_i} \, (w_i^0)^{\alpha w_i} \, \Gamma_n(\tfrac{\alpha w_i - 1}{2}) \end{aligned} \tag{7}$$

avec Γ_n est la fonction Gamma généralisée de dimension n ([6] page 427) :

$$\Gamma_n(b) = \left[\Gamma(\frac{1}{2})\right]^{\frac{1}{2}n(n-1)} \prod_{i=1}^{n} \Gamma(b + \frac{i-n}{2}), \;\; b > \frac{n-1}{2}$$

6 Conclusion

Dans ce paper, nous avons mis en évidence l'importance de munir l'espace des distributions avec une géométrie et une mesure de divergence. Une géométrie différente donne un réseau d'apprentissage différent. La construction d'une distribution *a priori* est établie dans le cadre de la théorie de prédiction bayésienne en minimisant un critère représentant un compromis entre une distribution de référence et la contrainte d'uniformité (ignorance).

Références

[1] R. E. Kass and L. Wasserman, "Formal rules for selecting prior distributions : A review and annotated bibliography", Technical report no. 583, Department of Statistics, Carnegie Mellon University, 1994.

[2] C. Rodríguez, "Entropic priors", *Tech. rep. Electronic form* `http : omega.albany.edu :8008/entpriors.ps`, (1991).

[3] S. Amari, *Differential-Geometrical Methods in Statistics*, Volume 28 of Springer Lecture Notes in Statistics, Springer-Verlag, New York, 1985.

[4] H. Zhu and R. Rohwer, "Bayesian invariant measurements of generalisation", in *Neural Proc. Lett.*, 1995, vol. 2 (6), pp. 28–31.

[5] H. Zhu and R. Rohwer, "Bayesian invariant measurements of generalisation for continuous distributions", Technical report, NCRG/4352, ftp ://cs.aston.ac.uk/neural/zhuh/continuous.ps.z, Aston University, 1995.

[6] G. E. P. Box and G. C. Tiao, *Bayesian inference in statistical analysis*, Addison-Wesley publishing, 1972.

[7] H. Snoussi and A. Mohammad-Djafari, "Penalized maximum likelihood for multivariate gaussian mixture", in *Bayesian Inference and Maximum Entropy Methods*, R. L. Fry, Ed. MaxEnt Workshops, August 2002, pp. 36–46, Amer. Inst. Physics.

[8] H. Snoussi and A. Mohammad-Djafari, "Information Geometry and Prior Selection.", in *Bayesian Inference and Maximum Entropy Methods*. MaxEnt Workshops, August 2002, to appear in Amer. Inst. Physics.