Apprentissage Bayésien Semi-Supervisé par Modélisation Générative

Elouan ARGOUARC'H[1,2] François DESBOUVRIES[2] Eric BARAT[1] Eiji KAWASAKI[1] Thomas DAUTREMER[1]
[1]Université Paris Saclay, CEA List, F-91120 Palaiseau, France
[2]Samovar, Telecom SudParis, Institut Polytechnique de Paris, F-91120 Palaiseau, France

Résumé – Dans cet article, nous comparons les modélisations générative et discriminante sous le prisme de la quantification d'incertitude épistémique, et confrontons ensuite les deux approches à la problématique de l'apprentissage semi-supervisé. Nous expliquons que l'approche générative permet de prendre en compte des données non-labelisées dans la modélisation, ce qui est structurellement impossible avec une approche discriminante. Nous proposons enfin un algorithme d'échantillonnage permettant l'apprentissage semi-supervisé et la quantification de l'incertitude épistémique d'un modèle génératif.

Abstract – In this article, we compare generative and discriminative modeling under the prism of epistemic uncertainty quantification. Next we confront the two approaches to the problem of semi-supervised learning. We show that the generative approach enables to take into account unlabeled data in the modeling, which is structurally impossible under a discriminative approach. Finally, we propose a sampling algorithm that enables semi-supervised learning and the quantification of the epistemic uncertainty of a generative model.

1 Introduction

Les problématiques usuelles de classification et de regression en apprentissage statistique peuvent être reformulées comme la modélisation d'une dépendance entre deux variables aléatoires à l'aide d'une loi de probabilité conditionnelle. Cette formulation probabiliste présente l'avantage de ne pas occulter l'éventuel (ou incontestable) caractère stochastique du processus d'intérêt ; caractérisant ainsi l'incertitude aléatoire. On distingue habituellement deux méthodes de modélisation[7] : l'approche générative et l'approche discriminante. La première s'attache à modéliser uniquement ce qui est inconnu, tandis que la seconde modélise directement ce que l'on cherche à prédire.

En outre, lorsque l'on infère un modèle à partir d'observations, il est nécessaire de caractériser une autre source d'incertitude qui entache la prédiction [2] : l'incertitude épistémique, liée au nombre fini d'observations. Dans ce papier nous considérons une modélisation Bayésienne, qui permet, par le biais de quantités prédictives calculées à partir de l'ensemble des modèles probables, de quantifier cette source d'incertitude.

Enfin, on parle d'apprentissage supervisé lorsque les données utilisées pour inférer le modèle sont complètes, c'est à dire des réalisations jointes des deux variables dont on cherche à modéliser la dépendance. Mais les contextes dans lesquels seulement une variable est observée sont de plus en plus fréquents en apprentissage statistique : on parle d'apprentissage semi-supervisé [8][10].

Dans ce papier nous comparons les approches générative et discriminante sous le prisme de la quantification de l'incertitude épistémique, et concluons en faveur de l'approche générative, qui est davantage compatible avec la philosophie Bayésienne (cf. §3). Par ailleurs, nous confrontons ces deux approches à la problématique de l'apprentissage semi-supervisé [1], et arrivons au constat que l'approche discriminante ne permet pas, dans la méthodologie Bayésienne, l'apprentissage semi-supervisé (cf. §4). Enfin nous proposons un algorithme qui intègre effectivement des données incomplètes dans l'inférence d'un modèle génératif, permettant ainsi l'apprentissage semi-supervisé et la quantification de son incertitude épistémique (cf. §5). Nous illustrons enfin ces mécanismes sur un problème de régression linéaire (cf. §6).

2 Apprentissage supervisé

Considérons (X, Y) un couple de variables aléatoires liées par un *Data Generating Process* (DGP) $P_{Y|X}$ qui décrit la loi de Y sachant X. Une prédiction consiste à se prononcer sur la valeur d'un x_0, inconnu, qui a généré une valeur observée y_0 via le DGP :

$$y_0 \sim P_{Y|X}(Y|X = x_0). \tag{1}$$

En inférence Bayésienne, conditionnellement à y_0, l'information sur la valeur d'intérêt x_0 et l'incertitude *aléatoire* (i.e. liée au caractère stochastique du DGP) sont entièrement décrites par la loi *a posteriori* :

$$p_{X|Y}(x_0|y_0) = \frac{p_{Y|X}(y_0|x_0)\pi_X(x_0)}{\int_{\Omega_X} p_{Y|X}(y_0|x)\pi_X(x)\mathrm{d}x}, \tag{2}$$

$p_{Y|X}(y_0|x_0)$ est la densité conditionnelle associée au DGP, souvent nommée par abus de langage *vraisemblance*, et $\pi_X(x_0)$ est la densité associée à la loi qui encode notre *a priori* (avant d'observer y_0) sur les valeurs probables de x_0.

Dans de nombreuses situations la densité associée au DGP est inaccessible - soit car le DGP (et par extension sa densité) est simplement inconnu ; - soit car le DGP n'est accessible que via son processus de simulation stochastique et la densité implicite est intractable [6]. Dans ce cas, l'expression (2) ne peut pas être calculée, pas même un à facteur multiplicatif près. Afin de pallier cette difficulté, il est commun d'utiliser une approche variationnelle afin d'approximer la postérieure d'intérêt à l'aide d'une famille de lois conditionnelles $\mathcal{F}_\Theta = \{P_\theta; \theta \in \Theta\}$ paramétrée par θ. On infère au sein de cette famille à partir de couples $\mathcal{D} = \{(x_i, y_i)|y_i \sim P_{Y|X}(Y|X = x_i)\}_{i=1}^{|\mathcal{D}|}$. Alors, en considérant θ comme une variable aléatoire,

(2) est approximée par la *postérieure prédictive* Bayésienne [4] :

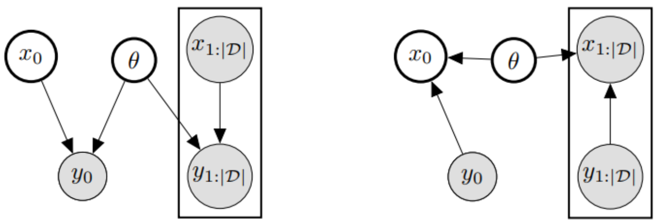
$$p(x_0|y_0,\mathcal{D}) = \int_\Theta p(x_0,\theta|y_0,\mathcal{D})\mathrm{d}\theta; \tag{3}$$

$$\text{avec } p(x_0,\theta|y_0,\mathcal{D}) = p(x_0|y_0,\cancel{\mathcal{D}},\theta,)p(\theta|y_0,\mathcal{D}), \tag{4}$$

ce qui permet de prendre en compte l'incertitude *épistémique* (i.e. liée à un nombre de données fini dans $\mathcal{D}$). Cette formulation inclut les problématiques standard d'apprentissage statistique : on parle de *régression* (resp. *classification*) lorsque X est une variable *continue* (resp. *catégorielle*).

3 Génératif versus Discriminant

L'approximation de la loi a posteriori peut se faire de deux manières duales : -soit indirectement, en modélisant la vraisemblance, i.e. la densité conditionnelle associée au DGP; -soit directement en modélisation la densité a posteriori. Nous étudions cette dualité en distinguant l'approche *Generative* de l'approche *Discriminante*.



(a) Génératif (b) Discriminant

FIGURE 1 : Modèles graphiques

3.1 Modélisation générative

L'approche générative se base sur la modélisation de la densité associée au DGP et l'on déduit la postérieure correspondante :

$$p(x_0|y_0,\theta) = \frac{\pi_X(x_0)p_\theta(y_0|x_0)}{p(y_0|\theta)}; \tag{5}$$

En général, cette expression ne peut être calculée qu'à une constante multiplicative près car le numérateur $p(y_0|\theta) = \int_{\Omega_X} p_\theta(y_0|x)\pi_X(x)\mathrm{d}x$ est intractable. On s'intéresse à la loi jointe (4) à l'intérieur de l'intégrale dans (3) : dans cette expression, aucun des deux facteurs ne peut être calculé individuellement, mais la loi jointe peut néanmoins être exprimée à une constante multiplicative près. En effet, en remplacant le premier facteur par son expression donnée par (5), et en remarquant que le second facteur peut s'écrire $p(\theta|y_0,\mathcal{D}) \propto p(\theta|\mathcal{D})p(y_0|\theta)$, on obtient, par simplification du facteur intractable $p(y_0|\theta)$:

$$p(x_0,\theta|y_0,\mathcal{D}) \propto \pi_X(x_0)p_\theta(x_0|y_0)p(\theta|D) \tag{6}$$

$$\text{où } p(\theta|D) \propto \pi_\Theta(\theta)\prod_{i=1}^{|\mathcal{D}|} p_\theta(y_i|x_i). \tag{7}$$

Finalement, en intégrant θ dans (6), la postérieure (3) s'écrit à une constante multiplicative près :

$$p(x_0|y_0,\mathcal{D}) \propto \pi_X(x_0)\int_\Theta p_\theta(y_0|x_0)p(\theta|\mathcal{D})\mathrm{d}\theta, \tag{8}$$

ce qui correspond au prior $\pi_X(x_0)$ multiplié par la *vraisemblance prédictive* $p(y_0|x_0,\mathcal{D})$. Ainsi, ces expressions nous permettent de déduire deux façons d'obtenir des échantillons distribués suivant (3) dans le cas d'une modélisation générative : - soit par échantillonnage de couples à partir de la densité jointe connue à une constante près (6); - soit, si possible, par calcul de la vraisemblance marginale et échantillonnage de la densité (8) connue à une constante près. Un échantillonnage en deux étapes correspondrait à $\theta \sim p(\theta|y_0,\mathcal{D})$ (que l'on ne sait malheureusement pas calculer) puis $x_0 \sim p(x_0|y_0,\theta)$.

3.2 Modélisation discriminante

L'approche discriminante se base sur la modélisation de la loi a posteriori :

$$p(x_0|y_0,\theta) = p_\theta(x_0|y_0), \tag{9}$$

(p_θ dénote ce qui est effectivement prédit par la modèle θ - comparer (5) et (9)). On s'intéresse à la loi jointe (4) à l'intérieur de l'intégrale dans (3) : le second facteur $p(\theta|y_0,\mathcal{D})$ se réduit à $p(\theta|\mathcal{D})$ (nous développons ce point dans la section 3.3.3); et finalement

$$p(x_0,\theta|y_0,\mathcal{D}) = p_\theta(x_0|y_0)p(\theta|\mathcal{D}) \tag{10}$$

$$\text{où } p(\theta|\mathcal{D}) \propto \pi_\Theta(\theta)\prod_{i=1}^{|\mathcal{D}|} p_\theta(x_i|y_i). \tag{11}$$

Nous déduisons finalement l'expression de la postérieure prédictive (3) en intégrant θ dans (10) :

$$p(x_0|y_0,\mathcal{D}) = \int_\Theta p_\theta(x_0|y_0)p(\theta|\mathcal{D})\mathrm{d}\theta. \tag{12}$$

Ainsi, ces expressions nous permettent de déduire trois façons d'obtenir des échantillons distribués suivant (3) dans le cas d'une modélisation discriminante : - soit par échantillonnage joint à partir de la densité (10);- soit, s'il est possible de la calculer, par échantillonnage à partir de la densité postérieure prédictive (12); - soit, motivé par (12), par échantillonnage en deux étapes $\theta \sim p(\theta|\mathcal{D})$ avec (11) puis $x_0 \sim p_\theta(x_0|y_0)$ avec (9).

3.3 Discussion

Tout d'abord, nous notons que, dans l'approche générative, l'équation (5) n'admet pas, en général (sauf sans le cas favorable où le modèle et le prior sont conjugués), une forme explicite et est d'autre part connue à une constante près. Une étape inférentielle (par MCMC [5] ou Inférence Variationnelle [9]) est donc nécessaire. Nous discutons dans cette section les spécificités, pratiques et conceptuelles, de chacune des approches; et nous remarquons que l'approche générative est peut être davantage compatible avec la philosophie Bayésienne dans la mesure où elle permet de spécifier une loi a priori (§3.3.1) qui est effectivement mise à jour avec une ou plusieurs observations (§3.3.2).

3.3.1 Spécification explicite du prior

La modélisation générative présente le premier avantage de pouvoir spécifier un prior π_X sur la valeur d'intérêt de x_0. L'approche discriminante quant à elle ne requiert pas de spécifier un prior car on modélise directement la postérieure;

dans ce cas, la loi marginale P_X est définie implicitement par le modèle : $p(x_0) = \int_\Theta \int_{\Omega_Y} p_\theta(x_0|y)p_Y(y)\pi_\Theta(\theta)\mathrm{d}y\mathrm{d}\theta$ où $p_Y(y) = \int_{\Omega_X} p_{X|Y}(x|y)\pi_X(x)\mathrm{d}x$. Cependant ce prior implicite n'intervient pas dans le calcul de la loi jointe (10) ; cette approche ne permet pas d'utiliser un a priori, potentiellement informatif, dans l'inférence de x_0.

3.3.2 Multiplicité des observations

Si l'on observe $y_{0,1}, ..., y_{0,N_0} \overset{iid}{\sim} P_{Y|X}(Y|X = x_0)$, alors d'une part, l'approche générative permet de réécrire l'éq. (5) :

$$p(x_0|y_{0,1}, ..., y_{0,N_0}, \theta) = \frac{\pi_X(x_0)\prod_{n=1}^{N_0} p_\theta(y_{0,n}|x_0)}{p(y_{0,1}, ..., y_{0,N_0}|\theta)}; \quad (13)$$

Cette équation peut être calculée au dénominateur près et donc toutes les équations de la section 3.1 s'écrivent de manière similaire. D'autre part, avec une approche discriminante, (9) devient :

$$p(x_0|y_{0,1}, ..., y_{0,N_0}, \theta) = \frac{\prod_{i=1}^{N_0} p_Y(y_{0,i})p_\theta(x_0|y_{0,i})}{p(x_0|\theta)^{N_0-1}p_Y(y_{0,1}, ..., y_{0,N_0})}; \quad (14)$$

mais le terme $p(x_0|\theta) = \int_{\Omega_Y} p_\theta(x_0|y_0)p_Y(y)\mathrm{d}y$ n'admet pas de forme explicite en général et cette expression ne peut pas être calculée, pas même à une constante près lorsque $N_0 > 1$. Ainsi, seule l'approche générative permet de prédire la valeur de x_0 à partir d'une multiplicité d'observations.

3.3.3 $p(\theta|y_0, \mathcal{D})$ ou $p(\theta|\mathcal{D})$?

Dans les sections 3.1 et 3.1 nous avons discuté de l'échantillonnage de la postérieure prédictive (3) qui est la loi marginale de x_0 de la loi jointe (4). Nous mettons à présent en évidence une différence notable entre les deux approches au regard de l'autre loi marginale :

$$p(\theta|y_0, \mathcal{D}) = p(\theta|\mathcal{D})\frac{p(y_0|\theta)}{p(y_0)} \quad (15)$$

Avec une modélisation discriminante $p(y_0|\theta) = p(y_0)$ et cette loi se réduit à $p(\theta|\mathcal{D})$: l'observation y_0 n'apporte aucune information sur le modèle θ. De manière moins formelle mais plus intuitive, cela reviendrait à dire que la loi a posteriori modélisée par θ doit produire un x_0 inconnu pour y_0 ; ce que l'on sait déjà car $P_\theta(.|Y = y_0)$ est une distribution de probabilité.

Avec une modélisation générative en revanche, $p(y_0|\theta)$ s'écrit sous la forme $p(y_0|\theta) = \int_{\Omega_X} p_\theta(y_0|x_0)\pi_X(x)\mathrm{d}x$. Ce qui signifie que les modèles probables sont les modèles qui, avec grande probabilité, produisent y_0 pour une valeur inconnue distribuée selon le prior π_X (il s'agit en réalité du x_0 d'intérêt). La loi $p(\theta|y_0, \mathcal{D})$ dépend donc effectivement de l'observation y_0.

4 Apprentissage semi-supervisé

Dans cette section, nous empruntons la terminologie spécifique de la problématique de classification : nous désignons par *label* la variable cachée X et par *observation* la variable observée Y produite par le DGP. Comme mentionné précédemment, le problème inverse et sa formulation Bayésienne correspondent à l'apprentissage *supervisé* : on cherche à inférer les modèles θ qui rendent probables les données *labélisées* de $\mathcal{D}$. Cependant, dans de nombreux contextes d'apprentissage statistique, on dispose, en plus de $\mathcal{D}$, de données *non-labelisées* $\mathcal{Y}$. Ces données se composent uniquement d'observations $\widetilde{y}_j$ dont on sait qu'elles sont produite par le DGP, mais pour un label inconnu $\widetilde{x}_j$ (d'intérêt ou non) :

$$\mathcal{Y} = \{\widetilde{y}_j | \exists \widetilde{x}_j \in \Omega_X, \widetilde{y}_j \sim P_{Y|X}(Y|X = \widetilde{x}_j)\}_{j=1}^{|\mathcal{Y}|}. \quad (16)$$

Lorsque les observations de $\mathcal{Y}$ et celles de $\mathcal{D}$ couvrent des régions différentes de Ω_Y et/ou lorsque $\mathcal{Y}$ contient un grand nombre d'observations, alors $\mathcal{Y}$ apporte une information non-négligeable relativement à $\mathcal{D}$. L'apprentissage semi-supervisé s'attache donc à inférer un modèle à partir de $\mathcal{D}$ et $\mathcal{Y}$.

La postérieure prédictive devient :

$$p(x_0|y_0, \mathcal{D}, \mathcal{Y}) = \int_\Theta p(x_0|y_0, \theta)p(\theta|y_0, \mathcal{D}, \mathcal{Y})\mathrm{d}\theta, \quad (17)$$

que l'on souhaite calculer, ou, à défaut, échantillonner. Cette formulation générale du problème inverse semi-supervisé se réduit au cas supervisé lorsque $\mathcal{Y} = \varnothing$.

Nous notons tout d'abord que la modélisation discriminante de permet pas l'apprentissage semi-supervisé dans la méthodologie présentée. En effet, de la même manière que, sans connaître x_0, y_0 n'apporte pas d'information sur θ (cf. §3.3.3), sans connaître le label $\widetilde{x}_j$, l'observation $\widetilde{y}_j$ n'apporte pas d'information sur le modèle. Finalement $p(\theta|y_0, \mathcal{D}, \mathcal{Y}) = p(\theta|\mathcal{D})$, (17) se réduit à (12), et toutes les autres équations de la section 3.2 restent inchangées. Ainsi, outre les raisons invoquées dans la section 3.3, cet argument favorise l'utilisation d'une modélisation générative dans le cas où l'on dispose également d'observations que l'on suppose informatives.

Nous explicitons à présent l'apprentissage semi-supervisé à l'aide d'une approche générative. L'équation (6) devient :

$$p(x_0, \theta|y_0, \mathcal{D}, \mathcal{Y}) \propto \pi_X(x_0)p_\theta(y_0|x_0)p(\theta|D)p(\mathcal{Y}|\theta) \quad (18)$$

$$\text{où } p(\mathcal{Y}|\theta) = \prod_{j=1}^{|\mathcal{Y}|} \int_{\Omega_X} p_\theta(\widetilde{y}_j|\widetilde{x}_j)\pi_{Xj}(\widetilde{x}_j)\mathrm{d}\widetilde{x}_j. \quad (19)$$

Dans l'équation (19), le terme fait intervenir un a priori sur le label $\widetilde{x}_j$ associé à l'observation $\widetilde{y}_j$ que l'on note π_{X_j}, qui n'est pas nécessairement commun à toutes les observations $\widetilde{y}_j$, ni le même prior utilisé pour inférer x_0 dans (2) ou (5). En général, ce terme n'est défini que sous forme intégrale et n'est pas calculable analytiquement, rendant (18) intractable, pas même à une constante près, ce qui soulève la question pratique de l'échantillonnage de cette postérieure prédictive. Nous proposons dans la section suivante un algorithme pour échantillonner cette postérieure.

5 Echantillonnage de Gibbs pour la postérieure prédictive

Nous présentons dans cette section l'Algorithme 1 permettant d'échantillonner la postérieure prédictive (17). Cet algorithme se base sur le principe de l'échantillonnage de Gibbs [3], qui est un algorithme MCMC dont la transition Markovienne consiste qui consiste à tirer sucessivement les conditionnelles dans la loi jointe $p(x_0, \widetilde{x}_1, ..., \widetilde{x}_{|\mathcal{Y}|}, \theta|y_0, \mathcal{D}, \mathcal{Y})$.

Algorithme 1 : Echantillonnage de Gibbs de la postérieure prédictive $p(x_0|y_0, \mathcal{D}, \mathcal{Y})$

Data : $y_0, \mathcal{D}, \mathcal{Y}$, nombre d'étapes T
Result : échantillon $x_0^{(T)}$
$\theta^{(0)} \sim p(\theta|\mathcal{D})$
for *t=1 to T* **do**
  $x_0^{(t)} \sim p(x_0|y_0, \theta^{(t-1)})$ avec l'équation (5)
  for $j = 1$ *to* $|\mathcal{Y}|$ **do**
    $\widetilde{x}_j^{(t)} \sim p(\widetilde{x}_j|\widetilde{y}_j, \theta^{(t-1)})$
  end
  $\mathcal{D}_+^{(t)} = \mathcal{D} \cup \{x_0^{(t)}, y_0\} \cup \{(\widetilde{x}_j^{(t)}, \widetilde{y}_j\}_{j=1}^{|\mathcal{Y}|}$
  $\theta^{(t)} \sim p(\theta|\mathcal{D}_+^{(t)})$ avec l'équation (7)
end

6 Expérience illustrative

Nous illustrons dans cette section l'effet de l'apprentissage semi-supervisé dans le cas d'une modélisation générative. Considérons $p_{Y|X}(y|x) = \mathcal{N}(y; \alpha_1 x + \alpha_0, \sigma^2_{Y|X})$ avec $\alpha_1 \neq 0$, et le prior $\pi_X(x) = \mathcal{N}(x; \mu_X, \sigma^2_X)$. Nous supposons α_1, α_0 et $\sigma^2_{Y|X}$ inconnus, et on cherche à les inférer à partir de $\mathcal{D}$ et $\mathcal{Y}$ (voir figure 2).

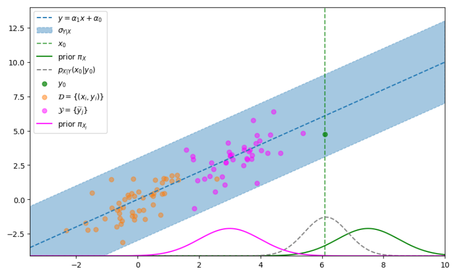


FIGURE 2 : Apprentissage semi-supervisé d'un modèle linéaire avec $\mathcal{D}$ et $\mathcal{Y}$ et postérieure pour y_0

Nous nous intéressons à la postérieure prédictive pour la même valeur observée de y_0, et nous comparons ainsi $p(x_0|y0, \mathcal{D}, \mathcal{Y})$ et $p(x_0|y_0, \mathcal{D})$ sous la forme d'histogrammes construits à partir d'échantillons obtenus grâce à l'algorithme de Gibbs présenté dans la section 5. Nous utilisons comme référence $p_{X|Y}(x_0|y_0)$, c'est-à-dire la vraie loi a posteriori, supposée inconnue. Nous observons sur la figure 3 que la postérieure prédictive $p(x_0|y_0, \mathcal{D}, \mathcal{Y})$ est plus proche de la postérieure $p_{X|Y}(x_0|y_0)$ que ne l'est $p(x_0|y_0, \mathcal{D})$. Cet exemple illustre deux éléments : (i) prendre en compte $\mathcal{Y}$ dans une modélisation générative permet effectivement de réduire l'incertitude épistémique de la prédiction, (ii) l'algorithme de Gibbs permet d'échantillonner la postérieure prédictive correspondante.

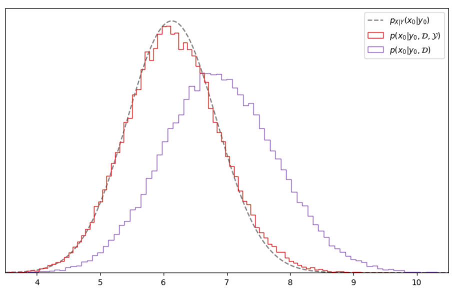


FIGURE 3 : Prédictives comparées à la loi a posteriori

7 Conclusion

Dans cet article nous avons comparé les approches générative et discriminante sous le prisme de la quantification Bayésienne de l'incertitude épistémique. Nous avons montré que l'approche générative est un outil versatile, mais la question de son utilisation pratique se pose car l'inférence du modèle génératif ne peut pas être faite indépendamment de l'observation. Nous avons utilisé ce constat pour expliquer que l'approche générative permet l'apprentissage semi-supervisé, ce qui est structurellement impossible avec l'approche discriminante. Enfin nous avons proposé un algorithme d'échantillonnage de la postérieure prédictive en modélisation générative semi-supervisée, que nous avons illustré sur un exemple de régression linéaire.

Références

[1] A. CHAN, A. ALAA, Z. QIAN et M. VAN DER SCHAAR : Unlabelled data improves bayesian uncertainty calibration under covariate shift. *In International Conference on Machine Learning*, pages 1392–1402. PMLR, 2020.

[2] A. DER KIUREGHIAN et O. DITLEVSEN : Aleatory or epistemic ? does it matter ? *Structural safety*, 31(2):105–112, 2009.

[3] A. E GELFAND : Gibbs sampling. *Journal of the American statistical Association*, 95(452):1300–1304, 2000.

[4] A. GELMAN, J. B. CARLIN, H. S. STERN et D. B. RUBIN : *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

[5] G. O. ROBERTS et J.S. ROSENTHAL : General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20 – 71, 2004.

[6] S. A. SISSON, Y. FAN et M. BEAUMONT : *Handbook of approximate Bayesian computation*. CRC Press, 2018.

[7] I. ULUSOY et C.M. BISHOP : Generative versus discriminative methods for object recognition. *In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 258–265 vol. 2, 2005.

[8] X. YANG, Z. SONG, I. KING et Z. XU : A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[9] C. ZHANG, J. BÜTEPAGE, H. KJELLSTRÖM et S. MANDT : Advances in variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):2008–2026, 2019.

[10] X. J. ZHU : Semi-supervised learning literature survey. 2005.