Statistique Bayésienne

Fondaments de théorie de la decision

Anna Simoni[2]

[2]CREST - Ensae and CNRS

Outline

L'inférence comme un problème decisionnel. I

- L'objectif de la plupart des études inférentielles est de fournir une décision au statisticien.
- Il est donc necessaire de definir un critère d'évaluation des procédures de décision qui prenne en compte les conséquences de chaque décision et dépende des paramètres du modèle, c'est-à-dire du vrai état du monde (ou de la nature).
- Importance du choix du critère car les conséquences de cette décision ne sont pas négligeables.
- Ce critère est appelé coût.

L'inférence comme un problème decisionnel. II

- Le problème decisionnel statistique est constitué d'un modèle $f(x|\theta)$ et d'un espace d'actions (ou *espace de décision*) $\mathcal{D}$.
- Une fonction de decision est une fonction $\delta(x) : \mathcal{X} \to \mathcal{D}$.

Définition

Une fonction de coût est une fonction $L : \Theta \times \mathcal{D} \to [0, +\infty]$.

La fonction de coût évalue l'érreur $L(\theta, \delta)$ associée à la décision δ quand le paramètre prend la valeur θ.

La fonction d'utilité.

- La fonction d'utilité $U(\theta, \delta)$ peut être vue comme une mesure de proximité entre l'estimation proposée δ et la vraie valeur $h(\theta)$.

- Combien nous coûte le fait de ne pas connaître θ avec certitude ? Si nous connaissions θ avec certitude, nous prendrions une décision

$$\delta^*(\theta) = \arg \sup_{\delta \in \mathcal{D}} U(\theta, \delta)$$

et la conséquence certaine serait $\sup_{\delta \in \mathcal{D}} U(\theta, \delta)$.

- Une fois que la fonction d'utilitée U a été construite, nous construisons la fonction de coût correspondante : $L(\theta, \delta) = -U(\theta, \delta)$ (et donc $U(\theta, \delta) \leq 0$) ou plus generalement

$$L(\theta, \delta) = (\sup_{\delta \in \mathcal{D}} U(\theta, \delta)) - U(\theta, \delta).$$

- Sauf pour les cas les plus triviaux, il est généralement impossible de minimiser (en δ) uniformément (en θ) la fonction de coût $L(\theta, \delta)$ quand θ est inconnu.

La fonction d'utilité.

- La fonction d'utilité $U(\theta, \delta)$ peut être vue comme une mesure de proximité entre l'estimation proposée δ et la vraie valeur $h(\theta)$.

- Combien nous coûte le fait de ne pas connaître θ avec certitude ? Si nous connaissions θ avec certitude, nous prendrions une décision

$$\delta^*(\theta) = \arg \sup_{\delta \in \mathcal{D}} U(\theta, \delta)$$

et la conséquence certaine serait $\sup_{\delta \in \mathcal{D}} U(\theta, \delta)$.

- Une fois que la fonction d'utilitée U a été construite, nous construisons la fonction de coût correspondante : $L(\theta, \delta) = -U(\theta, \delta)$ (et donc $U(\theta, \delta) \leq 0$) ou plus generalement

$$L(\theta, \delta) = (\sup_{\delta \in \mathcal{D}} U(\theta, \delta)) - U(\theta, \delta).$$

- Sauf pour les cas les plus triviaux, il est généralement impossible de minimiser (en δ) uniformément (en θ) la fonction de coût $L(\theta, \delta)$ quand θ est inconnu.

La fonction d'utilité.

- La fonction d'utilité $U(\theta, \delta)$ peut être vue comme une mesure de proximité entre l'estimation proposée δ et la vraie valeur $h(\theta)$.

- Combien nous coûte le fait de ne pas connaître θ avec certitude ? Si nous connaissions θ avec certitude, nous prendrions une décision

$$\delta^*(\theta) = \arg \sup_{\delta \in \mathcal{D}} U(\theta, \delta)$$

  et la conséquence certaine serait $\sup_{\delta \in \mathcal{D}} U(\theta, \delta)$.

- Une fois que la fonction d'utilitée U a été construite, nous construisons la fonction de coût correspondante : $L(\theta, \delta) = -U(\theta, \delta)$ (et donc $U(\theta, \delta) \leq 0$) ou plus generalement

$$L(\theta, \delta) = (\sup_{\delta \in \mathcal{D}} U(\theta, \delta)) - U(\theta, \delta).$$

- Sauf pour les cas les plus triviaux, il est généralement impossible de minimiser (en δ) uniformément (en θ) la fonction de coût $L(\theta, \delta)$ quand θ est inconnu.

La fonction d'utilité.

- La fonction d'utilité $U(\theta, \delta)$ peut être vue comme une mesure de proximité entre l'estimation proposée δ et la vraie valeur $h(\theta)$.

- Combien nous coûte le fait de ne pas connaître θ avec certitude ? Si nous connaissions θ avec certitude, nous prendrions une décision

$$\delta^*(\theta) = \arg \sup_{\delta \in \mathcal{D}} U(\theta, \delta)$$

  et la conséquence certaine serait $\sup_{\delta \in \mathcal{D}} U(\theta, \delta)$.

- Une fois que la fonction d'utilitée U a été construite, nous construisons la fonction de coût correspondante : $L(\theta, \delta) = -U(\theta, \delta)$ (et donc $U(\theta, \delta) \leq 0$) ou plus generalement

$$L(\theta, \delta) = (\sup_{\delta \in \mathcal{D}} U(\theta, \delta)) - U(\theta, \delta).$$

- Sauf pour les cas les plus triviaux, il est généralement impossible de minimiser (en δ) uniformément (en θ) la fonction de coût $L(\theta, \delta)$ quand θ est inconnu.

Outline

Fonction de Risque. I

Pour obtenir un critère de comparaison utilisable à partir d'une fonction de coût dans un contexte aléatoire on peut utiliser la *fonction de risque*.

- Risque fréquentiste :

$$\begin{aligned} R(\theta, \delta) &= \mathbf{E}_\theta[L(\theta, \delta(X))] \\ &= \int_{\mathcal{X}} L(\theta, \delta(x)) f(x|\theta) dx. \end{aligned}$$

- La fonction $\delta(\cdot) : \mathcal{X} \to \mathcal{D}$ est habituellement appelée *estimateur* (tandis que la valeur $\delta(x)$ est appelée *estimation* de θ.

Difficultés liées au risque frequentiste :

- Le critère de risque évalue les procédures selon leurs performances en moyenne et non directement pour une observation x donnée.
- L'analyse fréquentiste du problème de décision suppose tacitement que le même problème sera rencontré de nombreuses fois pour que l'évaluation en fréquence ait un sens. Répétabilité des expériences ?
- Pour une procédure δ, le risque $R(\theta, \delta)$ est une fonction du paramètre θ. L'approche fréquentiste *n'induit donc pas un ordre total* sur l'ensemble des procédures.

Fonction de Risque. I

Pour obtenir un critère de comparaison utilisable à partir d'une fonction de coût dans un contexte aléatoire on peut utiliser la *fonction de risque*.

- Risque fréquentiste :

$$\begin{aligned} R(\theta, \delta) &= \mathbf{E}_\theta[L(\theta, \delta(X))] \\ &= \int_{\mathcal{X}} L(\theta, \delta(x)) f(x|\theta) dx. \end{aligned}$$

- La fonction $\delta(\cdot) : \mathcal{X} \to \mathcal{D}$ est habituellement appelée *estimateur* (tandis que la valeur $\delta(x)$ est appelée *estimation* de θ.

Difficultés liées au risque frequentiste :

- Le critère de risque évalue les procédures selon leurs performances en moyenne et non directement pour une observation x donnée.
- L'analyse fréquentiste du problème de décision suppose tacitement que le même problème sera rencontré de nombreuses fois pour que l'évaluation en fréquence ait un sens. Répétabilité des expériences ?
- Pour une procédure δ, le risque $R(\theta, \delta)$ est une fonction du paramètre θ. L'approche fréquentiste *n'induit donc pas un ordre total* sur l'ensemble des procédures.

Fonction de Risque. II

L'approche bayésienne de la Théorie de la Décision intègre sur l'espace Θ, car θ est inconnu, plutôt que de le faire sur l'espace $\mathcal{X}$, x étant connu.

- Coût moyenne à posteriori :

$$\begin{aligned} \rho(\pi, \delta) &= \mathbf{E}^{\pi}[L(\theta, \delta(x))|x] \\ &= \int_{\Theta} L(\theta, \delta(x))\pi(\theta|x)d\theta. \end{aligned}$$

- Le coût moyen a posteriori est une fonction de x mais cette dépendance n'est pas gênante, contrairement à la dépendance fréquentiste du risque au paramètre puisque x, à la différence de θ, est connu.

Fonction de Risque. III

- Soit π une distribution à priori. On définit le *risque intégré* : (qui est le risque fréquentiste moyenné sur les valeurs de θ selon π)

$$\begin{aligned} r(\pi, \delta) &= \mathbf{E}^{\pi}[L(\theta, \delta(X))] \\ &= \int_{\Theta} \int_{\mathcal{X}} L(\theta, \delta(x)) f(x|\theta) dx \pi(\theta) d\theta. \end{aligned}$$

- Le risque integré associe un nombre réel à chaque estimateur, et non une fonction de θ. Il induit donc un ordre total sur l'ensemble des estimateurs et permet une comparaison directe entre ces estimateurs.

- Donc, l'approche bayésienne permet d'atteindre une décision efficace.

Théorème

Un estimateur minimisant le risque intégré $r(\pi, \delta)$ est obtenu par sélection, pour chaque $x \in \mathcal{X}$, de la valeur $\delta(x)$ qui minimise le coût moyen a posteriori, $\rho(\pi, \delta|x)$, puisque

$$r(\pi, \delta) = \int_{\mathcal{X}} \rho(\pi, \delta(x)|x) m(x) dx$$

où $m(x) = \int_{\Theta} f(x|\theta)\pi(\theta)$.

Fonction de Risque. III

- Soit π une distribution à priori. On définit le *risque intégré* : (qui est le risque fréquentiste moyenné sur les valeurs de θ selon π)

$$\begin{aligned} r(\pi, \delta) &= \mathbf{E}^{\pi}[L(\theta, \delta(X))] \\ &= \int_{\Theta} \int_{\mathcal{X}} L(\theta, \delta(x)) f(x|\theta) dx \pi(\theta) d\theta. \end{aligned}$$

- Le risque integré associe un nombre réel à chaque estimateur, et non une fonction de θ. Il induit donc un ordre total sur l'ensemble des estimateurs et permet une comparaison directe entre ces estimateurs.

- Donc, l'approche bayésienne permet d'atteindre une décision efficace.

Théorème

Un estimateur minimisant le risque intégré $r(\pi, \delta)$ est obtenu par sélection, pour chaque $x \in \mathcal{X}$, de la valeur $\delta(x)$ qui minimise le coût moyen a posteriori, $\rho(\pi, \delta|x)$, puisque

$$r(\pi, \delta) = \int_{\mathcal{X}} \rho(\pi, \delta(x)|x) m(x) dx$$

où $m(x) = \int_{\Theta} f(x|\theta)\pi(\theta)$.

Définition (Estimateur de Bayes)

Un estimateur de Bayes associé à une distribution à priori π et une fonction de coût L est un estimateur δ^π minimisant $r(\pi, \delta)$. Pour chaque $x \in \mathcal{X}$, ce dernier est donné par

$$\delta^\pi = \arg\min_d \rho(\pi, d|x).$$

La valeur $r(\pi) = r(\pi, \delta^\pi)$ est alors appelée risque de Bayes.

- Ce résultat est valable pour des a priori propres et impropres, du moment que le risque de Bayes $r(\pi)$ est fini.
- D'un point de vue strictement bayésien, seul le coût moyen à posteriori $\rho(\pi, d|x)$ compte, puisque le paradigme bayésien est fondé sur une approche conditionnelle. Faire la moyenne sur toutes les valeurs possibles de x, alors que nous connaissons la valeur observée de x, semble être une perte d'information.

Définition (Estimateur de Bayes)

Un estimateur de Bayes associé à une distribution à priori π et une fonction de coût L est un estimateur δ^π minimisant $r(\pi, \delta)$. Pour chaque $x \in \mathcal{X}$, ce dernier est donné par

$$\delta^\pi = \arg\min_d \rho(\pi, d|x).$$

La valeur $r(\pi) = r(\pi, \delta^\pi)$ est alors appelée risque de Bayes.

- Ce résultat est valable pour des a priori propres et impropres, du moment que le risque de Bayes $r(\pi)$ est fini.
- D'un point de vue strictement bayésien, seul le coût moyen à posteriori $\rho(\pi, d|x)$ compte, puisque le paradigme bayésien est fondé sur une approche conditionnelle. Faire la moyenne sur toutes les valeurs possibles de x, alors que nous connaissons la valeur observée de x, semble être une perte d'information.

Définition (Estimateur de Bayes)

Un estimateur de Bayes associé à une distribution à priori π et une fonction de coût L est un estimateur δ^π minimisant $r(\pi, \delta)$. Pour chaque $x \in \mathcal{X}$, ce dernier est donné par

$$\delta^\pi = \arg\min_d \rho(\pi, d|x).$$

La valeur $r(\pi) = r(\pi, \delta^\pi)$ est alors appelée risque de Bayes.

- Ce résultat est valable pour des a priori propres et impropres, du moment que le risque de Bayes $r(\pi)$ est fini.
- D'un point de vue strictement bayésien, seul le coût moyen à posteriori $\rho(\pi, d|x)$ compte, puisque le paradigme bayésien est fondé sur une approche conditionnelle. Faire la moyenne sur toutes les valeurs possibles de x, alors que nous connaissons la valeur observée de x, semble être une perte d'information.

Outline

Estimateurs randomisés

On peut étendre l'espace de décision à l'ensemble des *estimateurs randomisés*, prenant leurs valeurs dans $\mathcal{D}^*$, l'espace des distributions de probabilité sur $\mathcal{D}$. Utiliser un estimateur randomisé δ^* signifie que *l'action est générée selon la distribution de densité de probabilité* $\delta^*(x,\cdot)$, une fois que l'observation x a été recueillie.

- Coût de l'estimateur randomisé δ^* :

$$L(\theta,\delta^*(x)) = \int_{\mathcal{D}} L(\theta,a)\delta^*(x,a)da.$$

- Cette modification de l'espace $\mathcal{D}$ ne modifie pas les réponses bayésiennes :

Théorème

Pour toute distribution a priori π sur Θ, le risque de Bayes pour l'ensemble des estimateurs randomisés est le même que celui pour l'ensemble des estimateurs non randomisés, soit

$$\inf_{\delta\in\mathcal{D}} r(\pi,\delta) = \inf_{\delta^*\in\mathcal{D}^*} r(\pi,\delta^*) = r(\pi).$$

Définition

On appelle risque minimax associé à la fonction de coût la valeur

$$\bar{R} = \inf_{\delta \in \mathcal{D}^*} \sup_{\theta} R(\theta, \delta) = \inf_{\delta \in \mathcal{D}^*} \sup_{\theta} \mathbf{E}_\theta [L(\theta, \delta(X))]$$

et estimateur minimax tout estimateur (éventuellement randomisé) δ_0 tel que

$$\sup_{\theta} R(\theta, \delta_0) = \bar{R}.$$

Les estimateurs minimax n'existent pas nécessairement.

Théorème

Si $\mathcal{D} \subset \mathbb{R}^k$ est convexe et compact et si $L(\theta, \delta)$ est continue et convexe en tant que fonction de δ, pour chaque $\theta \in \Theta$, il existe un estimateur minimax non randomisé.

Théorème

Le risque de Bayes est toujours plus petit que le risque minimax :

$$\underline{R} = \sup_{\pi} r(\pi) = \sup_{\pi} \inf_{\delta \in \mathcal{D}} r(\pi, \delta) \leq \bar{R} = \inf_{\delta \in \mathcal{D}^*} \sup_{\theta} R(\theta, \delta).$$

Minimaxité. II

- $\underline{R}$ est dite risque maximin et une distribution π^* telle que $r(\pi^*) = \underline{R}$ est appelée *distribution a priori la moins favorable*, quand de telles distributions existent.

Définition

Un problème d'estimation est dit admettre une valeur si $\underline{R} = \bar{R}$, c'est-à-dire quand

$$\sup_{\pi} \inf_{\delta \in \mathcal{D}} r(\pi, \delta) = \inf_{\delta \in \mathcal{D}^*} \sup_{\theta} R(\theta, \delta).$$

- Le principe minimax ne fournit pas toujours des estimateurs acceptables.

Lemma

Si δ_0 est un estimateur de Bayes pour π_0 et si $R(\theta, \delta_0) \leq r(\pi_0)$ pour tout θ dans le support de π_0, δ_0 est minimax et π_0 est la distribution la moins favorable.

Admissibilité. I

Ce deuxième critère fréquentiste induit un ordre partiel sur $\mathcal{D}^*$ en comparant les risques fréquentistes des estimateurs $R(\theta, \delta)$.

Définition (Estimateur admissible)

Un estimateur δ_0 est inadmissible s'il existe un estimateur δ_1 qui domine δ_0, c'est-à-dire tel que pour tout θ,

$$R(\theta, \delta_0) \geq R(\theta, \delta_1)$$

et, pour au moins une valeur θ_0 du paramètre,

$$R(\theta_0, \delta_0) > R(\theta_0, \delta_1).$$

Sinon, δ_0 est dit admissible.

- Les estimateurs inadmissibles ne devraient pas être considérés du tout, puisqu'ils peuvent être améliorés uniformément.
- Cependant, l'admissibilité seule n'est pas suffisante pour valider l'utilisation d'un estimateur. Par exemple, les estimateurs constants $\delta(x) = \theta_0$ sont en général admissibles parce qu'ils fournissent une valeur exacte pour $\theta = \theta_0$.
- D'un point de vue fréquentiste, il est donc important de chercher des estimateurs qui satisfassent les deux optimalité : minimaxité et admissibilité.

Admissibilité. II

Proposition

S'il existe un unique estimateur minimax, cet estimateur est admissible.

Proposition

Si δ_0 est admissible de risque constant, δ_0 est l'unique estimateur minimax.

La notion d'admissibilité est fortement liée au paradigme bayésien : dans la plupart des problèmes statistiques, les estimateurs de Bayes engendrent la classe des estimateurs admissibles, c'est-à-dire que ces derniers peuvent être écrits comme des estimateurs de Bayes ou comme limites d'estimateurs de Bayes.

Admissibilité. III

Proposition

Si la distribution a priori π est strictement positive sur Θ, de risque de Bayes fini, et la fonction de risque $R(\theta, \delta)$ est une fonction continue de θ pour tout δ, l'estimateur de Bayes δ^π est admissible.

Proposition

Si l'estimateur de Bayes associé à une loi a priori π est unique, il est admissible.

Admissibilité. IV

Proposition

Si un estimateur de Bayes, δ^π, associé à une loi a priori (propre ou impropre) π, est tel que le risque de Bayes,

$$r(\pi) = \int_\Theta R(\theta, \delta^\pi)\pi(\theta)d\theta$$

soit fini, δ^π est admissible.

Outline

Le coût quadratique

Coût quadratique : $L(\theta, \delta) = \|\theta - \delta\|^2$.

Proposition

L'estimateur de Bayes δ^π associé à la loi a priori π et au coût quadratique est la moyenne a posteriori

$$\delta^\pi(x) = \mathbf{E}^\pi[\theta|x] = \frac{\int_\Theta \theta f(x|\theta)\pi(\theta)d\theta}{\int_\Theta f(x|\theta)\pi(\theta)d\theta}.$$

Corollary

Quand $\Theta \subset \mathbb{R}^p$, l'estimateur de Bayes δ^π associé à π et au coût quadratique,

$$L(\theta, \delta) = (\theta - \delta)' Q(\theta - \delta)$$

est la moyenne a posteriori, $\delta^\pi(x) = \mathbf{E}^\pi[\theta|x]$, pour toute matrice Q $p \times p$ symétrique définie positive.

Le coût absolu

Coût absolu : $L(\theta, \delta) = |\theta - \delta|$ ou, plus généralement, une fonction linéaire par morceaux :

$$L_{k_1,k_2}(\theta, \delta) = \begin{cases} k_2(\theta - \delta) & \text{si } \theta > \delta, \\ k_1(\delta - \theta) & \text{sinon.} \end{cases}$$

Proposition

L'estimateur de Bayes associé à la loi a priori π et à la fonction de coût linéaire par morceaux est le fractile $(k_2/(k_1 + k_2))$ de $\pi(\theta|x)$.

Le coût $0-1$

Coût $0-1$:

$$L(\theta,\delta)=\begin{cases}1-\delta & \text{si } \theta\in\Theta_0,\\ \delta & \text{sinon.}\end{cases}$$

Proposition

L'estimateur de Bayes associé à π et au coût $0-1$ est

$$\delta^\pi(x)=\begin{cases}1 & \textit{si } P(\theta\in\Theta_0|x)>P(\theta\notin\Theta_0|x),\\ 0 & \textit{sinon},\end{cases}$$

donc $\delta^\pi(x)$ vaut 1 si et seulement si $P(\theta\in\Theta_0|x)>1/2$.

Coûts intrinsèques

Il peut arriver que certains problèmes soient tellement non informatifs que non seulement la fonction de coût soit inconnue, mais aussi que le modèle n'admette pas une paramétrisation naturelle.

Il semble naturel d'utiliser des coûts comparant directement les distributions $f(\cdot|\theta)$ et $f(\cdot|\delta)$ associées au vrai paramètre θ et l'estimateur δ. De telles fonctions de coût,

$$L(\theta,\delta) = d(f(\cdot|\theta), f(\cdot|\delta))$$

sont effectivement indépendantes de la paramétrisation.

- Distance entropique (divergence de Kullback-Leibler) :

$$L_e(\theta,\delta) = \mathbf{E}_\theta \left[\log \left(\frac{f(x|\theta)}{f(x|\delta)} \right) \right]$$

- Distance de Hellinger :

$$L_H(\theta,\delta) = \frac{1}{2}\mathbf{E}_\theta \left[\left(\sqrt{\frac{f(x|\delta)}{f(x|\theta)}} - 1 \right)^2 \right].$$