

TD 1: Modèles bayésiens conjugués

Objectifs:

1. Découvrir des exemples d'application concrets des statistiques Bayésiennes.
2. Se familiariser avec les lois de probabilités classiques et leur utilisation pratique en modélisation.
3. Pratiquer, s'exercer et tester ses connaissances et sa compréhension de la méthode Bayésienne.

Nom	Domaine	Densité	Espérance	Variance
Bernoulli(p , $0 \leq p \leq 1$)	$\{0, 1\}$	$p^x(1-p)^{1-x}$	p	$p(1-p)$
Binomiale($n \in \mathbb{N}$, $0 \leq p \leq 1$)	$\{0, \dots, n\}$	$\binom{n}{x} p^x (1-p)^{n-x}$	np	$np(1-p)$
Poisson($\lambda > 0$)	\mathbb{N}	$\frac{\lambda^x e^{-\lambda}}{x!}$	λ	λ
Exponentielle($\lambda > 0$)	\mathbb{R}^+	$\lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Normale($\mu \in \mathbb{R}$, $\sigma^2 > 0$)	\mathbb{R}	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
Beta($\alpha, \beta > 0$)	$[0, 1]$	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Gamma($\alpha, \beta > 0$)	\mathbb{R}^+	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$
Inverse-Gamma($\alpha > 1, \beta > 0$)	\mathbb{R}^+	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\beta/x}$	$\frac{\beta}{\alpha-1}$	$\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$, si $\alpha > 2$
Student($\nu, \sigma^2 > 0, \mu \in \mathbb{R}$)	\mathbb{R}	$\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\sigma^2}} \left(1 + \frac{(x-\mu)^2}{\sigma^2\nu}\right)^{-\frac{\nu+1}{2}}$	μ	$\frac{\sigma^2\nu}{\nu-2}$, si $\nu > 2$

Table 1: Tableau des distributions de probabilité classiques

Exercice 1 (Modélisation de durée) – Un industriel étudie la durée de vie T_1, \dots, T_n (en mois) de composants électroniques identiques soumis à des conditions normales d'utilisation. On suppose que les temps de défaillance suivent une loi exponentielle :

$$T_i \mid \lambda \sim \text{Exp}(\lambda),$$

où les observations sont i.i.d conditionnellement à λ .

Les données suivantes (en mois) ont été observées sur $n = 8$ composants:

$$(t_1, \dots, t_8) = (42, 51, 39, 61, 47, 56, 44, 53).$$

Le constructeur affirme que la durée de vie moyenne attendue d'un composant est: 50 mois ± 10 mois. On adopte une approche bayésienne en modélisant le paramètre λ par une loi Gamma :

$$\lambda \sim \text{Gamma}(\alpha_0, \beta_0),$$

où α_0 et β_0 sont des hyperparamètres à préciser.

1. À partir des données observées, donnez l'estimateur fréquentiste de λ . Comment peut-on interpréter λ ?
2. En vous appuyant sur les informations fournies par le constructeur, proposez un choix raisonnable pour les hyperparamètres (α_0, β_0) de la loi a priori, et justifiez votre choix.
3. Déterminez la loi a posteriori de λ après observation des données.
4. Donnez l'expression de l'estimateur bayésien de λ (pour la perte quadratique). Cet estimateur est noté $\widehat{\lambda}_B$.
5. Comment évolue $\widehat{\lambda}_B$ lorsque le nombre d'observations $n \rightarrow 0$? lorsque $n \rightarrow +\infty$?
6. En vous inspirant de la formule de $\widehat{\lambda}_B$, comment peut-on interpréter les hyperparamètres de la loi a priori (α_0, β_0) ?

Exercice 2 (Modélisation des sinistres en assurance) – Une compagnie d'assurance automobile souhaite mettre à jour son estimation de la fréquence des sinistres d'un client A après n années. Le nombre de sinistres par an est souvent modélisé par une loi de Poisson car les sinistres sont des événements rares et indépendants. On note ce nombre par la variable aléatoire N qui suit une loi de Poisson $\mathcal{P}(\lambda)$. On rappelle sa densité:

$$\mathbb{P}(N = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k \in \mathbb{N}$$

Les observations des n dernières années du client A notées par N_1, \dots, N_n sont données par 1, 0, 3, 2, 0 (ici $n = 5$). Comme la moyenne de $\mathcal{P}(\lambda)$ est égale à λ , λ peut être interprétée comme le nombre moyen de sinistre par an. L'assurance a également des données historiques du nombre moyen de sinistres par an pour chacun de ses m clients: μ_1, \dots, μ_m dont la moyenne et l'écart type sont donnés par $\bar{\mu} = \frac{1}{m} \sum_{i=1}^m \mu_i = 3.4$ et $\sigma_\mu = \sqrt{\frac{1}{m} \sum_{i=1}^m (\mu_i - \bar{\mu})^2} = 2$.

Dans toutes les questions suivantes, donnez d'abord les formules en fonction des variables de l'énoncé avant d'utiliser les valeurs empiriques mentionnées.

1. Estimez λ en adoptant une approche fréquentiste. Cet estimateur est noté $\widehat{\lambda}_F$.
2. On souhaite à présent adopter une approche Bayésienne en prenant une loi a priori $\lambda \sim \text{Gamma}(\alpha, \beta)$. On note sa moyenne λ_{prior} . Comment peut-on choisir ses paramètres α, β ?
3. Déterminez la loi a posteriori $\lambda | N_1, \dots, N_n$.
4. À quoi correspond l'estimateur de Bayes dans ce cas ?
5. Écrivez la formule de l'estimateur de Bayes $\widehat{\lambda}_B$ comme combinaison convexe de la moyenne fréquentiste et de la moyenne a priori, c-à-d, trouvez $Z \in [0, 1]$ tel que:

$$\widehat{\lambda}_B = Z \widehat{\lambda}_F + (1 - Z) \lambda_{\text{prior}}$$

6. En théorie de crédibilité, Z est un score associé au client A. Il est appelé “facteur de crédibilité”. Pourquoi à votre avis ?

7. Expliquez l'influence du nombre d'années d'observation n sur l'estimation.

Exercice 3 (Bayesian A/B testing) – Google souhaite comparer l'efficacité de deux publicités (A et B) en menant un test A/B. Chaque publicité est affichée à un certain nombre d'utilisateurs n_A, n_B . On mesure le taux de conversion θ_A (resp. θ_B), c'est-à-dire la probabilité qu'un utilisateur clique sur la publicité A (resp. B) après l'avoir vue. On note X_A et X_B le nombre de clics respectivement obtenus pour les publicités A et B. Le but de l'étude est de comparer θ_A et θ_B en utilisant d'abord une approche fréquentiste, puis une approche bayésienne. On observe les chiffres suivants: $X_A = 40$, $X_B = 65$, $n_A = 1100$, $n_B = 1300$.

1. Approche Fréquentiste

1. Quel est le modèle approprié ?
2. Proposez un estimateur pour chaque paramètre θ_A et θ_B avec l'approche fréquentiste. On note ces estimateurs $\widehat{\theta}_{Af}$ et $\widehat{\theta}_{Bf}$.
3. Montrez que ces estimateurs sont asymptotiquement Gaussiens et déterminer les paramètres de leur limite Gaussienne.
4. On suppose que les données des deux publicités sont indépendantes et que n_A et n_B croissent de façon identique ($\lim_{n_A, n_B \rightarrow +\infty} \frac{n_A}{n_B} = 1$). En déduire la distribution asymptotique approximative de $\widehat{\theta}_{Af} - \widehat{\theta}_{Bf}$ en fonction de n_A et n_B .
5. On considère l'hypothèse $H_0 : \theta_A = \theta_B = \theta$, proposez un estimateur de θ . Déduire de la question précédente une statistique de la forme $W \stackrel{\text{def}}{=} \frac{\widehat{\theta}_{Af} - \widehat{\theta}_{Bf}}{Z}$ avec Z à déterminer telle que:

$$W \stackrel{n_A, n_B \rightarrow +\infty}{\sim} \mathcal{N}(0, 1).$$

6. En déduire un moyen de tester si l'hypothèse $\theta_A < \theta_B$ est vraie.

2. Approche Bayésienne

Google souhaite maintenant adopter une approche bayésienne en utilisant une loi Beta comme a priori :

$$\theta_A \sim \text{Beta}(\alpha_A, \beta_A), \quad \theta_B \sim \text{Beta}(\alpha_B, \beta_B)$$

7. Pour quelles valeurs de α, β obtiendrait-on des lois a priori non-informatives ?
8. Déterminez la loi a posteriori de θ_A et θ_B après observation des données.
9. Comment peut-on définir un estimateur bayésien de $\theta_A - \theta_B$?
10. Proposez une méthode de simulation empirique pour évaluer $\mathbb{P}(\theta_A < \theta_B)$.
11. Implémentez cette méthode et comparez le résultat à celui de l'approche fréquentiste.

12. Les entreprises en tech qui ont recours à la procédure du test A/B pour décider la meilleure version d'un produit, site-web, système de recommandation etc.. ont tendance à adopter l'approche Bayésienne. Comment pouvez-vous l'expliquer ?

Exercice 4 (Stock returns) – Un analyste financier veut étudier la rentabilité moyenne μ d'une action. On note le prix de l'action à la fin de chaque journée i par P_i . On définit le rendement logarithmique de l'action de la journée i par $X_i = \log(P_i/P_{i-1})$. On suppose que les rendements passés (logarithmiques) X_1, \dots, X_n suivent une loi normale:

$$X_i | \mu \sim \mathcal{N}(\mu, \sigma^2) \quad (i.i.d)$$

Partie 1. σ^2 connue.

On suppose σ^2 connue. L'analyste a une croyance a priori sur μ et modélise cette incertitude par une loi normale :

$$\mu \sim \mathcal{N}(\mu_0, \tau_0^2)$$

1. Donnez l'estimateur de μ selon une approche fréquentiste.
2. Déterminez la loi a posteriori de μ après observation des rendements passés.
3. Déduisez l'estimateur bayésien de μ .
4. Expliquez comment cet estimateur prend en compte l'information a priori et les données observées.

Partie 2. σ^2 inconnue.

On suppose maintenant que la variance σ^2 est inconnue et qu'on la modélise avec une loi a priori Inverse-Gamma :

$$\sigma^2 \sim \text{IG}(\alpha_0, \beta_0)$$

et que la loi conditionnelle de μ sachant σ^2 est une loi normale:

$$\mu | \sigma^2 \sim \mathcal{N}(\mu_0, \frac{\sigma^2}{\kappa_0})$$

Et que les rendements passés sont i.i.d conditionnellement à μ et σ^2 :

$$X_i | \mu, \sigma^2 \sim \mathcal{N}(\mu, \sigma^2) \quad (i.i.d)$$

5. Quelle est la loi a priori jointe de (μ, σ^2) ? Cette loi est appelée *Gaussian-inverse-gamma*.
6. Montrez que la loi jointe a posteriori de (μ, σ^2) après observation des rendements est une loi Gaussian-inverse-gamma. En particulier:

$$\mu | \sigma^2, X_1, \dots, X_n \sim \mathcal{N}(\mu_n, \frac{\sigma^2}{\kappa_0 + n})$$

$$\sigma^2 | X_1, \dots, X_n \sim \text{IG}(\alpha_n, \beta_n)$$

avec $\mu_n = \frac{\kappa_0 \mu_0 + n \bar{X}}{\kappa_0 + n}$ et $\alpha_n = \alpha_0 + \frac{n}{2}$ et $\beta_n = \beta_0 + \frac{1}{2} \sum_{i=1}^n (X_i - \bar{X})^2 + \frac{\kappa_0 n}{2(\kappa_0 + n)} (\bar{X} - \mu_0)^2$.

7. Comment peut-on interpréter le rôle de κ_0 ? Celui de β_n ?
8. On considère les données des 5 derniers jours de l'index SP500 : $X_1 = -0.013$, $X_2 = 0.002$, $X_3 = 0.006$, $X_4 = 0.021$, $X_5 = -0.008$. L'index SP500 augmente annuellement de 8% et a une volatilité quotidienne de 1.5% en moyenne. Nous avons seulement 5 observations, pour compenser cette faiblesse on utilise un a priori sur μ avec l'équivalent de 10 observations fictives égale à μ_0 . Pour faire simple, on peut également choisir $\alpha_0 = 5$. Précisez des valeurs raisonnables pour μ_0 , κ_0 , β_0 .
9. Comparez les valeurs obtenues avec une approche fréquentiste.

Partie 3. Predictive distribution. On considère toujours le même modèle de la partie 2. On veut prédire le rendement de l'action pour la journée $n + 1$ en quantifiant l'incertitude sur les valeurs futures du rendement. On s'intéresse donc à la loi de $X_{n+1}|X_1, \dots, X_n$. Si on arrive à caractériser cette loi, on pourra en déduire toutes les plages de valeurs possibles de X_{n+1} avec leur probabilités. Nous allons déterminer cette loi en utilisant la formule des probabilités totales (PT) deux fois. D'abord, on cherchera à identifier la loi de conditionnelle de $X_{n+1}|\sigma^2, X_1, \dots, X_n$. Puis celle de $X_{n+1}|X_1, \dots, X_n$

10. Appliquez la formule PT en conditionnant par rapport à μ pour montrer que $X_{n+1}|\sigma^2, X_1, \dots, X_n \sim \mathcal{N}(\mu_n, \sigma^2(1 + \frac{1}{\kappa_0+n}))$.
11. Appliquez la formule PT en conditionnant par rapport à σ^2 pour montrer que $X_{n+1}|X_1, \dots, X_n \sim \text{Student}(\nu_n, \mu_n, s_n^2)$ avec $\nu_n = 2\alpha_n$ et $s_n^2 = (1 + \frac{1}{\kappa_0+n})\frac{\beta_n}{\alpha_n}$.
12. Comment peut-on calculer la probabilité que le rendement de l'action pour la journée $n + 1$ soit supérieur à 0.001 ?