Bayes theorem for dominated models

Alessandra Guglielmi

October 2016

Those notes are based on the technical report Regazzini (1996) and material in Chapter 1 in Schervish (2012) in the References. Please, report any mistake in these notes to the instructor.

1 The Bayesian paradigm

Consider a sequence of random variables (r.v.'s) $X_1, X_2, \ldots, X_n, \ldots$ on $(\Omega, \mathcal{F}, \mathbb{P})$, a probability space. Let $P_\theta^{(n)}$ be the distribution of $\boldsymbol{X} := (X_1, \ldots, X_n)$ for any n, i.e. $P_\theta^{(n)}$ is a probability measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$; measures $\{P_\theta^{(n)}, n \geq 1\}$ are consisently defined, i.e. $P_\theta^{(n)}$ is the marginal distribution of $P_\theta^{(n+1)}$. Moreover, we drop the index n in $P_\theta^{(n)}$. Summing up:

$$\boldsymbol{X}|\theta \sim P_\theta, \quad \theta \in \Theta \subset \mathbb{R}^p.$$

By **Bayesian approach** we mean the statistical setting where θ itself is a random element, distributed according to π, which is a probability measure on $(\Theta, \mathcal{B}(\Theta))$, and it is called *prior distribution*. By *posterior distribution* we mean the conditional law of θ, given $\boldsymbol{X}$.

2 Bayes Theorem for dominated models

In case of dominated models, posterior distribution can be derived by **Bayes Theorem**:

$$\mathbb{P}(\theta \in B|\boldsymbol{X} = \boldsymbol{x}) \overset{a.s.}{=} \frac{\int_B f(\boldsymbol{x}|\theta)\pi(d\theta)}{\int_\Theta f(\boldsymbol{x}|\theta)\pi(d\theta)} \quad \forall B \in \mathcal{B}(\Theta),$$

where $f(\boldsymbol{x}|\theta)$ is a density of P_θ with respect to (w.r.t.) $\lambda^{(n)}$, a σ-finite measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. See the proof to understand which is the probability measure ruling the *a.s.*-equality above. Before seeing the proof, let us recall the definition of absolutely continuity (of measures) and the Radon-Nikodym theorem (see, for instance, Billingsley, 1986).

Definition: If μ and ν are measures on a measure space $(S, \mathcal{S})$, the measure ν is *absolutely continuous* w.r.t. μ if $\mu(A) = 0$ implies $\nu(A) = 0$. The relation is indicated by $\nu \ll \mu$.

Radon-Nikodym Theorem. If μ and ν are σ-finite measures on $(S, \mathcal{S})$ and $\nu \ll \mu$, then there exists a non-negative f, called density, such that $\nu(A) = \int_A f d\mu$ for all A in $\mathcal{S}$. Two such densities f and g are such that $\mu\{s : f(s) \neq g(s)\} = 0$.

PROOF OF BAYES THEOREM. Denote by $f(\boldsymbol{x}|\theta)$ a density of P_θ w.r.t. the σ-finite measure $\lambda^{(n)}$ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ (for instance, assume that $\lambda^{(n)}$ is the Lebesgue measure or the counting measure on $\mathbb{R}^n$), then

$$P_\theta(A) = \int_A f(\boldsymbol{x}|\theta)\lambda^{(n)}(d\boldsymbol{x}) \quad A \in \mathcal{B}(\mathbb{R}^n). \tag{1}$$

Moreover, we define a product measure γ on $\mathbb{R}^n \times \Theta$, that is the *joint distribution* of $(\boldsymbol{X}, \theta)$:

$$\begin{aligned} \gamma(A \times B) &:= \int_B P_\theta(A)\pi(d\theta) = \int_B \int_A f(\boldsymbol{x}|\theta)\lambda^{(n)}(d\boldsymbol{x})\pi(d\theta) \\ &= \int_A \left(\int_B f(\boldsymbol{x}|\theta)\pi(d\theta) \right) \lambda^{(n)}(d\boldsymbol{x}) \quad A \in \mathcal{B}(\mathbb{R}^n), B \in \mathcal{B}(\Theta), \end{aligned} \tag{2}$$

where last equality follows from Fubini's theorem. By μ_n we denote the marginal law of γ on $\mathbb{R}^n$, representing the *marginal distribution* of $\boldsymbol{X}$:

$$\mu_n(A) := \gamma(A \times \Theta) = \int_A \left(\int_\Theta f(\boldsymbol{x}|\theta)\pi(d\theta) \right) \lambda^{(n)}(d\boldsymbol{x}),\ A \in \mathcal{B}(\mathbb{R}^n). \tag{3}$$

Observe that, for any $B \in \mathcal{B}(\Theta)$, $\gamma(\cdot \times B)$ and μ_n are finite measures (therefore they are σ-finite as well). Moreover, it is easy to verify that $\gamma(\cdot \times B) \ll \mu_n(\cdot)$, since

$$A \times B = (A \times \Theta) \cap (\mathbb{R}^n \times B) \Rightarrow \gamma(A \times B) \leq \gamma(A \times \Theta) = \mu_n(A);$$

hence, when $\mu_n(A) = 0$, $\gamma(A \times B) = 0$ as well. By Radon-Nikodym Theorem, for any $B \in \mathcal{B}(\Theta)$, there exists a measurable function $\boldsymbol{x} \mapsto \pi(\boldsymbol{x}; B)$ such that

$$\gamma(A \times B) = \int_A \pi(\boldsymbol{x}; B)\mu_n(d\boldsymbol{x}) = \int_A \pi(\boldsymbol{x}; B) \left(\int_\Theta f(\boldsymbol{x}|\theta)\pi(d\theta) \right) \lambda^{(n)}(d\boldsymbol{x}) \quad A \in \mathcal{B}(\mathbb{R}^n). \tag{4}$$

However we also require that $\{\pi(\boldsymbol{x}; B)\}$ represents a *regular version of the conditional probability* of $\theta \in B$ given $\boldsymbol{X} = \boldsymbol{x}$, that is

i) $B \mapsto \pi(\boldsymbol{x}; B)$ is a probability measure on $(\Theta, \mathcal{B}(\Theta))$ for any $\boldsymbol{x}$;

ii) $\boldsymbol{x} \mapsto \pi(\boldsymbol{x}; B)$ is $\mathcal{B}(\mathbb{R}^n)$-measurable for any $B \in \mathcal{B}(\Theta)$ [it is a result of Radon-Nikodym theorem];

iii) $\pi(\boldsymbol{x}; B) = \mathbb{P}(\theta \in B | \boldsymbol{X} = \boldsymbol{x})$ a.s.-$\boldsymbol{x}$ for all $B \in \mathcal{B}(\Theta)$.

In order to verify $iii)$, we have to check the integral equation defining conditional probability, i.e.

$$\mathbb{P}(\boldsymbol{X} \in A, \theta \in B) = \int_A \mathbb{P}(\theta \in B|\boldsymbol{X} = \boldsymbol{x})\mu_n(d\boldsymbol{x}) = \int_A \pi(\boldsymbol{x}; B)\mu_n(d\boldsymbol{x}) = \gamma(A \times B).$$

This equality follows by the definition itself of $\pi(\boldsymbol{x}; B)$ - see (4). As far as $i)$ is concerned, and more generally the existence of the conditional probability of θ given $\boldsymbol{X}$, it follows since θ and $\boldsymbol{X}$ are random vectors (see Ash, 1972).

Therefore $\pi(\boldsymbol{x}; \cdot) = \mathbb{P}(\theta \in \cdot|\boldsymbol{X} = \boldsymbol{x})$ a.s.-μ_n rappresents the posterior distribution of θ. Now let us show how we can compute it.

From (2), (4) and the definition of conditional probability, we have that, for any $B \in \mathcal{B}(\Theta)$,

$$\pi(\boldsymbol{x}; B)\left(\int_\Theta f(\boldsymbol{x}|\theta)\pi(d\theta)\right) = \int_B f(\boldsymbol{x}|\theta)\pi(d\theta) \quad a.s. - \lambda^{(n)}(\text{anche } a.s. - \mu_n).$$

From the definition of μ_n it is straightforward to prove that

$$\mu(\{\boldsymbol{x} : \int_\Theta f(\boldsymbol{x}|\theta)\pi(d\theta) = 0\}) = 0,$$

and therefore

$$\pi(\boldsymbol{x}; B) = \mathbb{P}(\theta \in B|\boldsymbol{X} = \boldsymbol{x}) = \frac{\int_B f(\boldsymbol{x}|\theta)\pi(d\theta)}{\int_\Theta f(\boldsymbol{x}|\theta)\pi(d\theta)} \, a.s. - \mu_n, \quad B \in \mathcal{B}(\Theta) \quad \square.$$

In particular, when $X_1, \ldots, X_n \overset{iid}{\sim} f(x|\theta)$ density w.r.t. a measure λ on $\mathbb{R}$, then

$$\mathbb{P}(\theta \in B|X_1 = x_1, \ldots, X_n = x_n) = \frac{\int_B \prod_{i=1}^n f(x_i|\theta)\pi(d\theta)}{\int_\Theta \prod_{i=1}^n f(x_i|\theta)\pi(d\theta)} \, a.s. - \mu_n, \quad B \in \mathcal{B}(\Theta).$$

In this case the denominator is the density, w.r.t. the product measure $\lambda \times \cdots \times \lambda$ on $\mathbb{R}^n$, of the marginal distribution of $(X_1, \ldots, X_n)$:

$$m_{\boldsymbol{X}}(x_1, \ldots, x_n) = \int_\Theta \prod_{i=1}^n f(x_i|\theta)\pi(d\theta)$$

If, in addition, π has a density, w.r.t. a measure ν on Θ, that we still denote by $\pi(\theta)$, then the posterior distribution has a density too (w.r.t. ν), which is as follows:

$$\pi(\theta|x_1, \ldots, x_n) = \frac{\prod_{i=1}^n f(x_i|\theta)\pi(\theta)}{m_{\boldsymbol{X}}(x_1, \ldots, x_n)}, \quad \theta \in \Theta \quad a.s. - \mu_n.$$

3 Predictive distributions

Let $\{X_n, n \geq 1\}$ be the sequence of r.v.'s representing the available observations By **(posterior) predictive distributions** we denote the laws

$$\mathcal{L}(X_{n+1}, X_{n+2}, \ldots, X_{n+m} | X_1, \ldots, X_n).$$

In particular, the one-step-ahead posterior predictive distribution is the conditional law of X_{n+1}, given $X_1, \ldots, X_n$, and it represents the Bayesian forecast of X_{n+1} on the basis of data $X_1, \ldots, X_n$.

If $(X_1, \ldots, X_n)$, given θ, has density (w.r.t. the Lebesgue measure or the counting measure on $\mathbb{R}^n$) $f(\boldsymbol{x}|\theta)$, then the conditional law of X_{n+1}, given $X_1, \ldots, X_n$, has a density as well, and this density is given by the ratio of the joint densities:

$$m_{X_{n+1}|X_1,\ldots,X_n}(x; x_1, \ldots, x_n) = \frac{m_{X_1,\ldots,X_n,X_{n+1}}(x_1, \ldots, x_n, x)}{m_{X_1,\ldots,X_n}(x_1, \ldots, x_n)} = \frac{\int_\Theta f(\boldsymbol{x}, x|\theta)\pi(d\theta)}{\int_\Theta f(\boldsymbol{x}|\theta)\pi(d\theta)}$$

4 References

Ash (1972). *Real analysis and probability*. Academic Press.

Billingsley, P. (1986). *Probability and measure*. Wiley.

Regazzini, E. (1996). Impostazione non parametrica di problemi d'inferenza statistica bayesiana. Rapporto tecnico IAMI 96.21, `http://www.mi.imati.cnr.it/iami/abstracts/96-21.html`

Schervish, M. J. (2012). *Theory of Statistics*. Springer.