Advanced Statistics

Aurélien Garivier

ParisTech

September 22, 2008

Outline

A random parameter

Theorem (Bayes)

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\,\mathbb{P}(A)}{\mathbb{P}(B)}$$

- Probablities as a *partial belief*

An example: mamographies

- 1% of women at age forty who participate in routine screening have breast cancer
- 80% of women with breast cancer will get positive mammographies
- 9.6% of women without breast cancer will also get positive mammographies
- What is the probability that a women with positive mammography in a routine screening actually has breast cancer?

An example: mamographies

- 1% of women at age forty who participate in routine screening have breast cancer
- 80% of women with breast cancer will get positive mammographies
- 9.6% of women without breast cancer will also get positive mammographies
- What is the probability that a women with positive mammography in a routine screening actually has breast cancer? $7,8\%$

An example: mamographies

- 1% of women at age forty who participate in routine screening have breast cancer
- 80% of women with breast cancer will get positive mammographies
- 9.6% of women without breast cancer will also get positive mammographies
- What is the probability that a women with positive mammography in a routine screening actually has breast cancer? $7,8\%$
- Two visions of the probabilities: from *outside* (frequency) and from *inside* (partial belief).

Generic formulation - prior and posterior

- model: experiment $X \sim dP(x|\theta) = p(x|\theta)d\mu(x)$, $\theta \in \Theta$
- *prior* distribution: $d\pi(\theta)$
  contains knowledge on θ anterior to the experiment
- *posterior* distribution $d\Pi(\theta)$ with density:

$$\Pi(\theta) = p(\theta|X) = \frac{\pi(\theta)p(X|\theta)}{\int_\theta \pi(\theta)p(X|\theta)d\theta} \propto \pi(\theta)p(X|\theta)$$

- Idea: *the experiment modifies the beliefs on θ*
- conjugate prior: $\pi(\theta)$ and $\Pi(\theta|X)$ have a common pattern
  Example: $X \sim \mathcal{N}(\theta, \sigma^2)$, $\theta \sim \mathcal{N}(m, \tau)$.
- Confidence interval, tests...

Example: Binomial variables

- Conjugate prior family: *Beta*(a, b)

$$\pi(\theta) = \frac{\theta^{a-1}\left(1-\theta\right)^{b-1}}{\beta(a, b)}$$

where

$$\beta(a, b) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}, \quad \Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$$

- $\mathbb{E}\left[Beta(a, b)\right] = \frac{a}{a+b}$, $\mathrm{Var}\left[Beta(a, b)\right] = \frac{ab}{(a+b)^2(a+b+1)}$
- Posterior distribution:

$$\Pi(\theta) \sim Beta\left(x+a, n-x+b\right)$$

$$\hat{\theta}_\pi(X) = \frac{X+a}{n+a+b}$$

Consistency result

Definition

The posterior distribution Π is said to be consistent at θ_0 if for every neighbourhood U of θ_0, $\Pi(U|X_1, \ldots, X_n)$ goes to 1 almost surely as $X_1, \ldots, X_n \sim^{iid} P_{\theta_0}$.

Theorem (Doob)

Suppose $\mathbb{P}(\cdot|\theta) \neq \mathbb{P}(\cdot|\theta')$ for $\theta \neq \theta'$. For any prior π, the posterior is consistent at every θ except possibly on a set of π-measure zero.

Theorem (Bernstein - Von Mises)

Under appropriate conditions [see Bickel-Docksum Section 5.5],

$$\mathcal{L}\left(\sqrt{n}\left(\hat{\theta}_\pi - \theta\right)|X_1, \ldots, X_n\right) \to \mathcal{N}\left(0, \mathrm{Var}[P_\theta]\right)$$

almost-surely under P_θ for all θ.

Risk, Bayesian and minimax approaches

Definition

- The (quadratic) *risk* of estimator $\hat{\theta}$ under paramater θ is:

$$R(\theta, \hat{\theta}) = \mathbb{E}_{X \sim P_\theta} \left\| \hat{\theta} - \theta \right\|^2$$

- Frequentist approach: worst case
  - ▷ *worst case risk*: $\overline{R}(\hat{\theta}) = \sup_{\theta \in \Theta} R(\theta, \hat{\theta})$
  - ▷ *minimax risk*: $\overline{R} = \inf_{\hat{\theta}} \overline{R}(\hat{\theta})$ ($\Longrightarrow$ *minimax estimator*)
- Bayesian approach: prior π
  - ▷ *average risk under* π: $\underline{R}(\pi, \hat{\theta}) = \mathbb{E}_{\theta \sim \pi} R(\theta, \hat{\theta})$
  - ▷ *bayesian risk under* π: $\underline{R}(\pi) = \inf_{\hat{\theta}} \underline{R}(\pi, \hat{\theta})$
  - ▷ *maximin risk*: $\underline{R} = \sup_\pi \underline{R}(\pi)$ ($\Longrightarrow$ *least favorable prior*)

Properties

Lemma

The bayesian risk is always smaller than the minimax risk: $\underline{R} \leq \overline{R}$

Lemma

The bayesian risk:

$$\underline{R}(\pi) \triangleq \min_{\hat{\theta}} \mathbb{E}_{\Pi} \left[\|\hat{\theta} - \theta\|^2 \right]$$

is reached by the the posterior mean $\hat{\theta}_\pi = \mathbb{E}[\Pi]$.

Theorem ("a Bayes rule with constant risk is minimax")

If $\hat{\theta}_\pi$ *is a Bayes estimator with respect to a prior* π *and if* $R(\theta, \hat{\theta}_\pi) = \underline{R}(\pi)$ *for all* θ*, then* $\hat{\theta}_\pi$ *is minimax and* π *is a least favorable prior.*

Minimax estimator in the Binomial setup

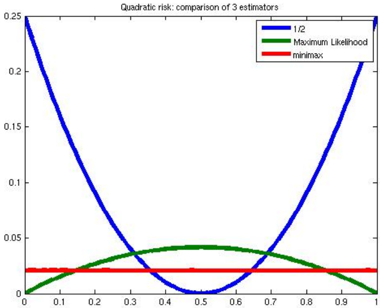
Theorem

Let $X \sim \mathcal{B}(n, \theta), \theta \in [0, 1]$

- *The minimax minimax estimator is*

$$\hat{\theta}_n(X) = \frac{X + \sqrt{n}/2}{n + \sqrt{n}}$$

- *It has quadratic loss* $\bar{R}_n = \frac{1}{4\left(1+\sqrt{n}\right)^2}$
- *The least favorable prior is* $\pi_n = Beta\left(\sqrt{n}/2, \sqrt{n}/2\right)$



Outline

"pulling yourself up by your own bootstraps"

What to do when the classical testing or estimating procedure can't be trusted? When the distribution is strongly non-gaussian? When the amount of data is not sufficient to assume normality?

Idea: create new data *from* the data available !

Principle: resampling

- Given a sample $X_1, \ldots, X_n$ from a distribution P, let P_n be the empirical distribution
- **plug-in**: to estimate a functional $T(P)$, estimate $T(P_n)$!
- Let $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ be an estimator of $T(P)$
- for k from 1 to N ($N \gg 1$), repeat
    1. **Resampling:** sample $\tilde{X}_1^k, \ldots, \tilde{X}_n^k$ from P_n i.e. from $\{X_1, \ldots, X_n\}$ *with replacement*
    2. compute an estimator $\hat{\theta}_k = \hat{\theta}_k\left(\tilde{X}_1, \ldots, \tilde{X}_n\right)$ of $T(P_n)$.
- **Bootstrap idea**: the empirical distribution of the $\left(\hat{\theta}_k\right)_k$ is close to the distribution of $\hat{\theta}$ given $X_1, \ldots, X_n$

Berry-Esseen Theorem

Theorem

Let $X_1, \ldots, X_n$ be iid with $\mathbb{E}[X_i] = 0, \mathbb{E}[X_i^2] = \sigma^2$ and $\mathbb{E}\left[|X_i|^3\right] = \rho < \infty$. If F_n is the distribution of $(X_1 + \cdots + X_n)/\sigma\sqrt{n}$ and $\mathcal{N}$ is the standard normal distribution, then

$$|F_n(x) - \mathcal{N}(x)| \leq \frac{3\rho}{\sigma^3\sqrt{n}}$$

Properties of the Bootstrap distribution

- **shape:** because it approximates the sampling distribution, the bootstrap distribution can be used to check normality of the latter.
- **spread:** the standard deviation of the sampling distribution $\mathrm{Var}[P_n]^{1/2}$ is approximately the standard error of the statistic $\hat{\theta}_n$.
- **center:** the bias of the bootstrap distribution mean $T(P_n)$ from the value of the statistic on the sample $\hat{\theta}$ is the same as the bias of $\hat{\theta}_n$ from $T(P)$

Bootstrap Confidence Intervals

- **Bootstrap t-confidence interval:** instead of $\hat{\sigma}/\sqrt{n}$, use the standard deviation of the bootstrap distribution to estimate the deviation of the sampling distribution.
  Requests that its shape is nearly gaussian.
- **Bootstrap percentile confidence interval:** keep as a α-confidence interval the central $1-\alpha$ values of $\left(\hat{\theta}_k\right)_k$
- Example: confidence interval for the mean, regression.

Comparing of two groups

Given independent samples $X_1, \dots, X_n$ and $Y_1, \dots, Y_m$,

1. Draw a resample of size n of the first sample $X_1, \dots, X_n$, and a separate resample of size m of the second sample $Y_1, \dots, Y_m$.
2. Compute the statistic that compares the two groups, such as the difference between the two sample means
3. Repeat the first two steps 10000 times
4. Construct the bootstrap distribution of the statistic.
5. Inspect the shape, bias, and bootstrap error (i.e., the standard deviation of the bootstrap distribution).
6. Other possibility: permutation test.