BML lecture #5: Foundations

http://github.com/rbardenet/bml-course

Rémi Bardenet

CNRS & CRIStAL, Univ. Lille, France

Misc

- You can still apply to the PhD position I advertise on my website with me and Subhro Ghosh (NUS Singapore).
- Stay tuned: we will announce projects on Monday, papers go on a first come first served basis.
- I'm late with writing up the solutions to the exercises, but they are coming!

Outline

1. Introduction
2. The likelihood principle
3. Subjective (ot ”personalist”) Bayesians
4. Objective Bayes
5. Frequentist Bayes
6. Most people are hybrid Bayesians
7. Discussion

Outline

What comes to *your* mind when you hear "Foundations "?

Recap: posterior expected utility

The subjective expected utility principle

1. Choose $\mathcal{S}, \mathcal{Z}, \mathcal{A}$ and a loss function $L(a, s)$,
2. Choose a distribution p over $\mathcal{S}$,
3. Take the the corresponding Bayes action

$$a^\star \in \arg\min_{a \in \mathcal{A}} \mathbb{E}_{s \sim p} L(a, s). \tag{1}$$

Corollary: minimize the posterior expected loss

If we partition $s = (s_o, s_\mathrm{u})$, then

$$a^\star \in \arg\min_{a \in \mathcal{A}} \mathbb{E}_{s_o} \mathbb{E}_{s_\mathrm{u} | s_o} L(a, s).$$

Equivalently to (1), given s_o, we choose

$$a^\star = \delta(s_o) = \arg\min_{a \in \mathcal{A}} \mathbb{E}_{s_\mathrm{u} | s_o} L(a, s).$$

Outline

The “formal” LP

Consider two statistical experiments

$$E_i = (X_i, \theta, \{p_i(\cdot|\vartheta), \vartheta \in \Theta\}), \quad i = 1, 2.$$

Assume that for some realizations x_1 and x_2,

$$p_1(x_1|\cdot) \propto p_2(x_2|\cdot).$$

If $\mathrm{Ev}(E, x)$ denotes the “evidence on θ arising from E and x”, then

$$\mathrm{Ev}(E_1, x_1) = \mathrm{Ev}(E_2, x_2).$$

Corollary

$\mathrm{Ev}(E, x)$ can depend on x solely through $p(x|\cdot)$.

An example: model-based classification

- Take $p_i(s_i) = p_i(x_i, \theta) = p_i(x_i|\theta)p(\theta) = Z p_i(\theta|x_i)$.
- Then for $a : \mathcal{S} \to \mathcal{Z}$,

$$\int L(a, s_1)\frac{p_1(x_1|\theta)p(\theta)}{Z}\mathrm{d}\theta \propto \int L(a, s_2)\frac{p_2(x_2|\theta)p(\theta)}{Z}\mathrm{d}\theta,$$

  so that Bayes actions coincide: $a^\star = \delta_1(x_1) = \delta_2(x_2)$.
- However, full expected utilities are different:

$$\begin{aligned}\int L(a, s_1)p_1(x_1|\theta)p(\theta)\mathrm{d}x_1\mathrm{d}\theta &= \int L(a, s_2)C(x_2)\, p_2(x_2|\theta)p(\theta)\mathrm{d}x_2\mathrm{d}\theta \\ &\neq \int L(a, s_2)p_2(x_2|\theta)p(\theta)\mathrm{d}x_2\mathrm{d}\theta.\end{aligned}$$

- The LP is compelling to many (Berger and Wolpert, 1988), but it has its downsides.
- It doesn't lead all the way to Bayes.
- I am (personally) uncomfortable with the stopping rule principle: it seems to good to be the right answer.
- It is hard to make fully formal: is $\mathrm{Ev}(E, x)$ even meaningful? See answer by LeCam to (Berger and Wolpert, 1988).
- It assumes well-specification: $x \sim p(\cdot|\theta^\star)$ for some $\theta^\star$. This is often false in ML.
- It separates the roles of the likelihood and the prior. For LP-abiding Bayesians, the prior is not allowed to depend on data.

Outline

- Top requirement is internal coherence of decisions.
- Various attempts at proving that internally coherent decision-makers minimize some expected utility; see (Parmigiani and Inoue, 2009).

Figure: Bruno de Finetti (1906–1985) and L. Jimmie Savage (1917–1971)

Savage's axioms 1/2

- Start with the triple $(\mathcal{S}, \mathcal{Z}, \mathcal{A} \subset \mathcal{F}(\mathcal{S}, \mathcal{Z}))$ as in Wald, 1950.
- Savage's idea is to list what we expect from a binary relation $\prec$ on $\mathcal{A} \times \mathcal{A}$ describing a decision maker's preferences.

Savage's axioms 2/2

De Finetti's theorem (Hewitt-Savage form)

Theorem: exchangeable $\leftrightarrow$ conditionally i.i.d.; see(Sch95)

Let $X_1, X_2, \dots$ be a sequence of exchangeable random variables on $\mathcal{X}$, i.e.

$$X_1, \dots, X_n \sim X_{\pi(1)}, \dots, X_{\pi(n)}, \forall n, \forall \pi \in \mathfrak{S}_n.$$

Then there exists a probability distribution μ on the set of probability measures $\mathcal{P}(\mathcal{X})$ on $\mathcal{X}$ such that

$$\mathbb{P}(X_1 \in A_1, \dots, X_n \in A_n) = \int Q(A_1) \dots Q(A_n) \mathrm{d}\mu(Q).$$

Furthermore, if $Q \sim \mu$,

$$Q(A) = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} 1_A(X_i).$$

To a subjectivist, Savage's theorem says you should use SEU, and representation theorems like de Finetti's constrain your choice of p.

De Finetti's theorem and LDA

The Blackwell-McQueen urn scheme (aka the CRP)

Start with an urn containing a single black ball with weight α. Repeat: draw a ball from the urn with probability $\propto$ its weight. Then,

- If the ball is black, return it to the urn along with another ball of weight 1, with a new color sampled from some base measure H.
- If the ball is colored, return it to the urn along with another ball of weight 1 of the same color.

Denote by $X_1, \dots$ the color of the ball added.

- Exercise: show that $X_1, X_2, ...$ are exchangeable.
- The corresponding prior μ on $\mathcal{P}(\mathcal{X})$ is the Dirichlet process with concentration α and base measure H.

Flaws in the subjectivist viewpoint

Outline

Objective (or consensual) Bayes

- A historical objection to Bayes is the need to choose a prior.
- By "objective", we mean that the prior is chosen by some external rule, and that this rule is relatively consensual.
- Take for instance, Jeffreys's "noninformative" priors.

Outline

Complete class theorems state that there is a good prior (but which one?)

A complete class theorem for estimation (J. O. Berger, 1985)

Under topological and Euclidean assumptions, if further

- $L(\theta, \cdot)$ is continuous,
- $\theta \mapsto \int L(\theta, \hat{\theta}) p(y_{1:n}|x_{1:n}, \theta) \mathrm{d}y_{1:n}$ is continuous for any $\hat{\theta}$,

then for any estimator $\tilde{\theta}$ there exists a prior and a corresponding Bayes estimator

$$\hat{\theta}_{\text{Bayes}} \in \arg\min_{\hat{\theta}} \mathbb{E}_{\theta|x_{1:n},y_{1:n}} L(\theta, \hat{\theta})$$

such that

$$\forall \theta, \quad \mathbb{E}_{y_{1:n}|x_{1:n},\theta} L(\theta, \hat{\theta}_{\text{Bayes}}) < E_{y_{1:n}|x_{1:n},\theta} L(\theta, \tilde{\theta}).$$

Bayesian estimators thus have good frequentist properties

But finding the "right" prior can be difficult. Frequentists typically use Bayesian derivations with particular (often data-dependent) priors; see e.g. empirical Bayes procedures (Efron, 2012).

PAC bounds; see e.g. (Shalev-Shwartz and Ben-David, 2014)

Let $(x_{1:n}, y_{1:n}) \sim \mathbb{P}^{\otimes n}$, and independently $(x, y) \sim \mathbb{P}$, we want an algorithm $g(\cdot; x_{1:n}, y_{1:n}) \in \mathcal{G}$ such that if $n \geqslant n(\delta, \varepsilon)$,

$$\mathbb{P}^{\otimes n}\left[\mathbb{E}_{(x,y)\sim\mathbb{P}} L(a_g, s) \leqslant \varepsilon\right] \geqslant 1 - \delta.$$

McAllester's bound for 0-1 loss (Chapter 31 of the above book)

For any two distributions P, Q on $\mathcal{G}$, with $\mathbb{P}^{\otimes n}$-probability $1 - \delta$,

$$\mathbb{E}_{g\sim Q}\mathbb{P}(g(x) \neq y) \leqslant \mathbb{E}_{g\sim Q}\frac{1}{n}\sum_{i=1}^{n} 1_{g(x_i)\neq y_i} + \sqrt{\frac{\mathrm{KL}(Q, P) + \log(n/\delta)}{2(n-1)}}.$$

This suggests taking the "posterior" Q to be in

$$\arg\min \mathbb{E}_{g\sim Q}\frac{1}{n}\sum_{i=1}^{n} 1_{g(x_i)\neq y_i} + \sqrt{\frac{\mathrm{KL}(Q, P) + \log(n/\delta)}{2(n-1)}}.$$

Outline

One possible hybrid view, e.g. (Robert, 2007)

- The starting point is Wald's decision setting, adding integration with respect to a prior.
- It is simple, widely applicable, has good frequentist properties.
- It satisfies the likelihood principle.
- It is tempting to interpret it as follows: beliefs are
  - represented by probabilities,
  - updated using Bayes' rule,
  - integrated when making decisions.
- It is easy to communicate your uncertainty
  - Simply give your posterior.
  - When making a decision, make sure that the priors of everyone involved would yield the same decision.
  - Alternately, perform a prior sensitivity analysis.

Outline

What kind of Bayesian are you?

- I've only scratched the surface. See e.g. (Mayo, 2018).
- Posterior expected utility is conceptually simple and unifying. Beyond that, many intepretations get (partial) philosophical support.
- The role of the likelihood, the prior, your update mechanism, etc. depend on the intepretation that you choose.
- ☹ Many people do not care.
- Hybrid views have become common among statisticians (Robert, 2007; Gelman et al., 2013), but this arguably makes the role of priors fuzzy.
- In ML, the development of Bayesian nonparametrics is reviving the subjectivist view, while objective approaches like PAC-Bayes are also increasingly popular.
- A great entry on subjective Bayes is (Parmigiani and Inoue, 2009).

[1] James O Berger. *Statistical decision theory and Bayesian analysis*. Springer, 1985.

[2] Berger and R. L. Wolpert. *The likelihood principle: A review, generalizations, and statistical implications*. Vol. 6. Institute of Mathematical Statistics, 1988.

[3] B. Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*. Vol. 1. Cambridge University Press, 2012.

[4] A. Gelman et al. *Bayesian data analysis*. 3rd. CRC press, 2013.

[5] D. G. Mayo. *Statistical inference as severe testing: How to get beyond the statistics wars*. Cambridge University Press, 2018.

[6] G. Parmigiani and L. Inoue. *Decision theory: principles and approaches*. Vol. 812. John Wiley & Sons, 2009.

[7] C. P. Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media, 2007.

References II

[8] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[9] A. Wald. *Statistical decision functions*. Wiley, 1950.