

Bayesian ML

Lecture # 4: Why would you want to be Bayesian?

Rémi Bardenet

CNRS & CRIStAL, Univ. Lille, France
<http://rbardenet.github.io>

- 1 Because you abide by the likelihood principle
- 2 Because you place coherence above all things: subjective Bayes
- 3 Because you like coherence and consensus: objective Bayes
- 4 Because you want to be a good (Waldian) frequentist
- 5 Most modern Bayesians are hybrid Bayesians
- 6 Discussion

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

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

Corollary: minimize the posterior expected loss

If we partition $s = (s_o, s_u)$, then

$$a^* = a_{g^*} \in \arg \min_{a \in \mathcal{A}} \mathbb{E}_{s_o} \mathbb{E}_{s_u | s_o} L(a, s).$$

Equivalently to (1), given s_o , we choose

$$g^*(s_o) = \arg \min_{a \in \mathcal{A}} \mathbb{E}_{s_u | s_o} L(a, s).$$

- 1 Because you abide by the likelihood principle
- 2 Because you place coherence above all things: subjective Bayes
- 3 Because you like coherence and consensus: objective Bayes
- 4 Because you want to be a good (Waldian) frequentist
- 5 Most modern Bayesians are hybrid Bayesians
- 6 Discussion

The likelihood principle (Berger and Wolpert, 1988)

The “formal” LP of Berger and Wolpert, 1988

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

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

Corollary

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

► Take $p_i(s_i) = p_i(x_i, \theta) = p_i(x_i|\theta)p(\theta) = Z p_i(\theta|x_i)$, $i = 1, 2$.

► Then for $a : \mathcal{S} \rightarrow \mathcal{Z}$,

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

so that Bayes actions coincide: $a^* = \delta_1(x_1) = \delta_2(x_2)$.

► However, full expected utilities are different in general:

$$\int L(a, s_1) p_1(x_1|\theta) p(\theta) dx_1 d\theta \neq \int L(a, s_2) p_2(x_2|\theta) p(\theta) dx_2 d\theta.$$

The stopping rule principle follows from the LP

- ▶ The LP is compelling to many (Berger and Wolpert, 1988), but it has its downsides.
- ▶ Being Bayesian is only one way to abide by the LP.
- ▶ I am personally uncomfortable with the stopping rule principle, probably because my frequentist intuition is still too strong.
- ▶ It is hard to make fully formal: is $Ev(E, x)$ even meaningful? See answer by LeCam to (Berger and Wolpert, 1988).
- ▶ It assumes we want to specify a likelihood, this prevents model-free Bayesianism.
- ▶ It separates the roles of the likelihood and the prior. For LP-abiding Bayesians, **the prior is not allowed to depend on data.**

- 1 Because you abide by the likelihood principle
- 2 Because you place coherence above all things: subjective Bayes**
- 3 Because you like coherence and consensus: objective Bayes
- 4 Because you want to be a good (Waldian) frequentist
- 5 Most modern Bayesians are hybrid Bayesians
- 6 Discussion

The subjectivistic viewpoint

- ▶ 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)

- ▶ 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.

Theorem; see (Schervish, 2012, Theorem 1.49)

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

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

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

- ▶ Axiomatic derivations are powerful, and shed light on what coherence requires. In particular, coherence leads to SEU.
- ☹ ... Yet all axiomatic systems have a bit that is difficult to swallow.
- ▶ Priors should be elicited by expert knowledge, and should be *bona fide* probability distributions.
- ▶ Representation theorems can help design the joint distribution over states (Orbanz and Roy, 2014).
- ▶ Bayesian nonparametrics has revived the subjectivist viewpoint.

- 1 Because you abide by the likelihood principle
- 2 Because you place coherence above all things: subjective Bayes
- 3 Because you like coherence and consensus: objective Bayes**
- 4 Because you want to be a good (Waldian) frequentist
- 5 Most modern Bayesians are hybrid Bayesians
- 6 Discussion

- ▶ 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.

Jeffrey's prior does not satisfy the LP

- 1 Because you abide by the likelihood principle
- 2 Because you place coherence above all things: subjective Bayes
- 3 Because you like coherence and consensus: objective Bayes
- 4 Because you want to be a good (Waldian) frequentist**
- 5 Most modern Bayesians are hybrid Bayesians
- 6 Discussion

A complete class theorem for estimation (Berger, 1985)

Under topological and Euclidean assumptions, if further

- ▶ $L(\theta, \cdot)$ is continuous,
- ▶ $\theta \mapsto \int L(\theta, \hat{\theta}) p(y_{1:n} | x_{1:n}, \theta) dy_{1:n}$ 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

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

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).

The Lasso

$$\hat{\theta}_{\text{Lasso}} \in \arg \min_{\theta} \frac{1}{2} \|y - X\theta\|^2 + \lambda \|\theta\|_1.$$

- ▶ If $X|\theta \sim \mathcal{N}(X\theta, I)$ and $p(\theta) \propto e^{-\lambda \|\theta\|_1}$, the MAP is $\hat{\theta}_{\text{Lasso}}$.
- ▶ But the posterior is not sparse.

Spike-and-slab priors

$$p(\theta|w) = (1 - w)\delta_0 + wq(\theta).$$

The horseshoe prior (Carvalho, Polson, and Scott, 2010)

$$\begin{aligned} \lambda_1, \dots, \lambda_d, \tau &\sim C^+(0, 1), \\ \theta_j | \lambda_j, \tau &\sim \mathcal{N}(0, \tau^2 \lambda_j^2), \quad j = 1, \dots, d. \end{aligned}$$

The horseshoe prior (Carvalho, Polson, and Scott, 2010)

$$\begin{aligned}\lambda_1, \dots, \lambda_d, \tau &\sim C^+(0, 1), \\ \theta_j | \lambda_j, \tau &\sim \mathcal{N}(0, \tau^2 \lambda_j^2), \quad j = 1, \dots, d. \\ y_{1:n} | x_{1:n}, \theta &\sim \mathcal{N}(X\theta, \sigma^2 I)\end{aligned}$$

Theorem (van der Pas, Kleijn, and van der Vaart, 2014)

As $n, p_n \rightarrow \infty$, with $\tau = (p/n)^\alpha$, $\alpha \geq 1$,

$$\sup_{\|\theta\|_0^2 = p_n} \mathbb{E} \|g^*(x_{1:n}, y_{1:n}) - \theta\|^2 \asymp p_n \log \frac{n}{p_n}.$$

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 $n \geq n(\delta, \varepsilon)$,

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

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 $1 - \delta$,

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

This suggests taking the “posterior” Q to be in

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

- 1 Because you abide by the likelihood principle
- 2 Because you place coherence above all things: subjective Bayes
- 3 Because you like coherence and consensus: objective Bayes
- 4 Because you want to be a good (Waldian) frequentist
- 5 Most modern Bayesians are hybrid Bayesians**
- 6 Discussion

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** when priors do not depend on data.
- ▶ 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**.

- 1 Because you abide by the likelihood principle
- 2 Because you place coherence above all things: subjective Bayes
- 3 Because you like coherence and consensus: objective Bayes
- 4 Because you want to be a good (Waldian) frequentist
- 5 Most modern Bayesians are hybrid Bayesians
- 6 Discussion**

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 interpretations get partial philosophical support.**
- ▶ The role of the likelihood, the prior, your update mechanism, etc. **depend** on the interpretation that you choose.
- ☹ **Many people do not care.**
- ▶ Hybrid views have become common (Robert, 2007; Gelman et al., 2013). 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 door to subjective Bayes is (Parmigiani and Inoue, 2009).

- [1] J. O. Berger. *Statistical decision theory and Bayesian analysis*. Springer, 1985.
- [2] J. O. Berger and R. L. Wolpert. *The likelihood principle: A review, generalizations, and statistical implications*. Vol. 6. Institute of Mathematical Statistics, 1988.
- [3] C. M. Carvalho, N. G. Polson, and J. G. Scott. “The horseshoe estimator for sparse signals”. In: *Biometrika* 97.2 (2010), pp. 465–480.
- [4] B. Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*. Vol. 1. Cambridge University Press, 2012.
- [5] A. Gelman et al. *Bayesian data analysis*. 3rd. CRC press, 2013.
- [6] D. G. Mayo. *Statistical inference as severe testing: How to get beyond the statistics wars*. Cambridge University Press, 2018.

- [7] P. Orbanz and D. M. Roy. “Bayesian models of graphs, arrays and other exchangeable random structures”. In: *IEEE transactions on pattern analysis and machine intelligence* 37.2 (2014), pp. 437–461.
- [8] G. Parmigiani and L. Inoue. *Decision theory: principles and approaches*. Vol. 812. John Wiley & Sons, 2009.
- [9] C. P. Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media, 2007.
- [10] M. J. Schervish. *Theory of statistics*. Springer Science & Business Media, 2012.
- [11] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.
- [12] S. L. van der Pas, B. J. K. Kleijn, and A. W. van der Vaart. “The horseshoe estimator: Posterior concentration around nearly black vectors”. In: *Electronic Journal of Statistics* 8.2 (2014), pp. 2585–2618.

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