Bayesian Learning / Posterior Evolution / Bayesian Bandits

Source: Bandit Algorithms, L. & S., Chap. 34

Setting: Prior information on arms; prior characterized in terms of a given arm marginal distribution. The approach is to evolve the posterior distribution of the arms based on history of samples, and use this posterior to guide future actions.

Posterior dist: Conditional pdf/pmf of an arm reward distribution based on past samples.

(Several technical issues regarding existance, uniqueness. Please read Sec. 34.2 in text). In these notes, we will implicitly assume everything is well defined, and not worry about technical issues.

In a Bayesian setting, information about each arm fully characterized by the posterior dist. In general, computing the posterior is tricky. However, in several important cases, the form of the posterior distribution is the same as that of the prior,

only with new parameters. Such distributions are called conjugate pairs and the associated prior is a conjugate prior.

Conjugate Prior	Pairs

(1) Gaussian model: $\{X_i, i=1,2,3,\dots\}$ a conditionally iid sequence of Gaussian rvs with unknown mean but known variance. Specifically, let Θ be a proper random variable. Then,

$$P_{(X_1,\dots,X_n)|\Theta}(x_1,\dots x_n \mid \Theta=\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_s^2}} e^{-(x_i-\theta)^2/2\sigma_s^2}$$

σ_s^2: known variance

Further, Θ itself is a Gaussian r.v., with known mean and variance.

$\Theta \sim N(\mu_p, \sigma_p^2)$ the prior distribution

i.e., $$P_\Theta(\theta) = \frac{1}{\sqrt{2\pi\sigma_p^2}} e^{-(\theta-\mu_p)^2/2\sigma_p^2}$$

Then, the posterior distribution on the (unknown) mean Θ is given by:

$$P_{\Theta|X}(\theta|x) \sim N\left(\underbrace{\frac{\frac{\mu_p}{\sigma_p^2} + \frac{x}{\sigma_s^2}}{\frac{1}{\sigma_p^2} + \frac{1}{\sigma_s^2}}}_{\text{updated mean}}, \underbrace{\frac{1}{\frac{1}{\sigma_p^2} + \frac{1}{\sigma_s^2}}}_{\text{updated variance}} \right)$$

Since the prior and posterior are both Gaussian, and the posterior simply updates the mean and variance of Θ as above, the posterior evolution given samples $x_1, x_2, \cdots x_n$ is easy to describe in closed form. Specifically,

$$P_{\Theta|X_1,\cdots X_n}(\theta|x_1,\cdots x_n) \sim N\left(\frac{\frac{\mu_p}{\sigma_p^2} + \frac{\sum_{i=1}^{n} x_i}{\sigma_s^2}}{\frac{1}{\sigma_p^2} + \frac{n}{\sigma_s^2}}, \frac{1}{\frac{1}{\sigma_p^2} + \frac{n}{\sigma_s^2}} \right)$$

---

(2) Beta-Bernoulli Model: $\{X_i, i=1,2,\cdots\}$ are conditionally iid Bernoulli(Θ) r.vs., given $\Theta = \theta$ (i.e., Bernoulli with unknown mean).

Initial Prior (before seeing any samples):

$\Theta \sim \text{Uniform}[0,1]$

$$P(X_1 = x_1, X_2 = x_2, \cdots X_n = x_n \mid \Theta = \theta) \qquad x_i \in \{0,1\}, \ \theta \in [0,1]$$

$$= \prod_{i=1}^{n} P(X_i = x_i \mid \Theta = \theta)$$

Let $S_n = \sum_{i=1}^{n} X_i$ (S_n = # of 1s, F_n = # of 0s)

$F_n = (n - S_n)$.

$$= \theta^{S_n} (1-\theta)^{(n - S_n)} \qquad S_n = \sum_{i=1}^{n} x_i$$

Posterior Dist. of Θ given $(x_1, \cdots, x_n)$:

$$P_{\Theta | (X_1, \cdots X_n)}(\theta \mid x_1, \cdots x_n)$$

$$= \frac{P(X_1 = x_1, \cdots X_n = x_n \mid \Theta = \theta) \, P_\Theta(\theta)}{\int_0^1 P(X_1 = x_1, \cdots X_n = x_n \mid \Theta = z) \, P_\Theta(z) \, dz}$$

$$= C\,\theta^{s_n}(1-\theta)^{n-s_n} \longrightarrow (1)$$

C → normalizing constant.

Aside: Beta Distribution: $Y \sim \text{Beta}(\alpha, \beta)$ if

support $(Y) = [0,1]$ and

$$p_Y(y) = \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right) y^{\alpha-1}(1-y)^{\beta-1}$$

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \qquad \Gamma(m) = (m-1)!$$

Comparing (1) with above,

Note: $\text{Beta}(1,1) = \text{Unif}[0,1]$

$$p_{\theta|(x_1,\ldots x_n)}(\theta \mid x_1,\ldots x_n) \sim \text{Beta}(s_n+1, f_n+1).$$

$s_n = \sum_{i=1}^{n} x_i$ = # of 1's

$f_n = n - \sum_{i=1}^{n} x_i$ = # of 0's

---

(3) Exponential family of Distributions (both of the above are a special case of this). Read sec. 34.5.1 in book for a discussion of this family.

Bayesian Bandit Environment

k-armed bandit environment: $(\mathcal{E}, \mathcal{G}, Q, P)$

- $\mathcal{E}$: the unknown parameter space over k arms.
- Q: prior dist on parameters.
- P: reward dist. for a fixed environment, i.e., fixed $\theta = (\theta_1, \cdots \theta_k)$.

ex: in Bernoulli bandit with unknown means $(\theta_1, \theta_2, \cdots \theta_k)$,

$\mathcal{E} = [0,1]^k$

$\mathcal{G} =$ Borel-σ-algebra $[0,1]^k$

Example is: $Q \sim \text{Beta}(\alpha, \beta)$, $\alpha, \beta \geq 1$

- $(\mathcal{E}, \mathcal{G})$ the parameter space.
- Q a prob. dist. over $(\mathcal{E}, \mathcal{G})$, and is the prior distribution. For any fixed $\nu \in \mathcal{E}$, $P(\nu \in B) = Q(B)$.
  - B: subset of environments (e.g. subset of $[0,1]^k$ for Bernoulli bandit).
- Policy for playing arms $\pi = (\pi_t, t = 1, 2, \cdots)$ generating actions $\{A_1, A_2, \cdots\}$ and corresponding arm rewards $(X_1, X_2, \cdots)$, where the conditional dist. of A_t given $\nu, A_1, X_1, A_2, X_2, \cdots, A_{t-1}, X_{t-1}$ is:

$$\pi_t(\cdot \mid A_1, X_1, A_2, X_2, \dots, A_{t-1}, X_{t-1})$$

↳ depends on actions and observed rewards

- $P = (P_{\nu i}, \nu \in \mathcal{E}, i \in \{1, 2, \dots, k\})$ a family of arm reward distributions, for each fixed environment $\nu \in \mathcal{E}$ and arm $i \in \{1, 2, \dots, k\}$. The dist. of the reward given $\nu, A_1, X_1, \dots, A_{t-1}, X_{t-1}$ is

$$X_t \sim P_{\nu A_t}(\cdot).$$

↳ depends on environment ν and action A_t

Posterior dist. on (Action, Reward) Sequence:

$(\mathcal{E}, G, Q, P)$ bandit with policy π.

$$P_{\nu\pi}(a_1, x_1, a_2, x_2, \dots, a_n, x_n)$$

$$= \prod_{t=1}^{n} \pi_t(a_t \mid a_1, x_1, \dots, a_{t-1}, x_{t-1}) \cdot P_{\nu a_t}(x_t)$$

Bayesian Regret

Recall (frequentist / environment-dependent) regret is:

$$R_n(\pi, \nu) = n\mu^* - E\left[\sum_{i=1}^{n} X_i\right]$$

$\mu^* = \max_{1 \le i \le K} \mu_i$

expectation over randomization due to policy and arm rewards for a FIXED environment ν.

$$BR_n(\pi, Q) = \int_{\nu \in \mathcal{E}} R_n(\pi, \nu)\, dQ(\nu)$$

Bayesian Regret ≡ Average over environments w.r.t prior Q.

$$= E_{\nu \sim Q}\left[R_n(\pi, \nu)\right]$$