Lecture 25

区别于 frequentist inference，Bayesian approach 认为 parameters 也为 random variable.

§1 Bayesian framework

1. **Definition:** Prior distribution

任何定义在 parameter space Θ 上的 distribution $\Lambda(\theta|\lambda)$ 被称为 prior distribution with hyperparameter λ

2. **Definition:** Posterior distribution

对于一个 sample X drawn from $f(x|\theta)$，posterior distribution $\pi(\theta|x,\lambda)$ 被定义为 θ conditional on observed $X=x$ 的(条件)概率，即

$$\pi(\theta|x,\lambda) = \frac{f(x|\theta)\Lambda(\theta|\lambda)}{\int f(x|\theta)\Lambda(\theta|\lambda)\,d\theta} \propto f(x|\theta)\Lambda(\theta|\lambda) = K(\theta|x,\lambda)\cdot h(x,\lambda)$$

其中 $K(\theta|x,\lambda)$ 被称为 kernel of posterior distribution

注：相较于直接计算 posterior distribution 的 exact form，求出 kernel 会更有效．因为 posterior 与 kernel 的比值仅是一些与 θ 无关的 constant，用于 normalize kernel

Prior distribution 的选取往往很重要，为了便于计算，有以下常用 prior.

3. **Definition:** conjugate distribution family 与 conjugate prior

考虑：① distribution family $\mathcal{F}=\{f(x|\theta),\ \theta\in\Theta\}$ (elements 定义在 $\mathcal{X}$ 上) (sample distribution)

② distribution family $\mathcal{G}=\{\Lambda(\theta|\lambda),\ \lambda\in\tilde{\Lambda}\}$ (elements 定义在 Θ 上) (prior distribution)

① $\mathcal{F}$ 和 $\mathcal{G}$ 被称为 conjugate distribution family，若对于 $\forall\ \Lambda(\theta|\lambda)\in\mathcal{G}$，有

$$\pi(\theta|x,\lambda) = \frac{f(x|\theta)\Lambda(\theta|\lambda)}{\int f(x|\theta)\Lambda(\theta|\lambda)\,d\theta} \in \mathcal{G} \quad \text{(posterior distribution)}$$

② 且 $\mathcal{G}$ 的任一 element 被称为 conjugate prior

注：换言之，conjugate prior 和 posterior 同属一个分布族

4. **Definition:** Jeffrey's prior

若 sample X drawn from $f(x|\theta)$，其 information matrix 为 $I(\theta)$．则 Jeffrey's prior 被定义为：

$$\Lambda(\theta) = (\det I(\theta))^{1/2}$$

注：① 在 one dimensional 情况下，Jeffrey's prior is invariant under reparameterization.

对于 reparameterization $\phi=h(\theta)$，有

$$\Lambda(\phi) = \Lambda(\theta)\cdot\left|\frac{\partial\theta}{\partial\phi}\right| = (I(\theta))^{1/2}\cdot\left|\frac{\partial\theta}{\partial\phi}\right| = (I(\phi))^{1/2}$$

换言之，若 θ 的 Jeffrey's prior 为 $\Lambda(\theta)$，则 $\phi=h(\theta)$ 的 Jeffrey's prior 为 $\Lambda(h(\theta))$

② 在 one dimensional 情况下，Jeffrey's prior 被称为 non-informative prior．因为 Jeffrey's prior maximizes prior 和 posterior 间的 KL-divergence．由于 $\Lambda(\theta|\lambda)$ 和 $\pi(\theta|x,\lambda)$ 间的 gap 是 x 的 information，Jeffrey's prior 可以视作保留了最多的 x 的 information，令 posterior 保留了最少的 θ 的 information

5. **Definition:** Improper prior

对于一个 sample X drawn from $f(x|\theta)$，一个 Θ 上的 function $\Lambda(\theta|\lambda)$ 被称为 improper prior，若

① $\Lambda(\theta|\lambda) \geq 0$ （类似于 pdf，非负）

② $\int_\Theta \Lambda(\theta|\lambda)\, d\theta = \infty$ （确保 $\Lambda(\theta|\lambda)$ 不是 pdf，且不能被 normalize 成 pdf）

③ $\int_\Theta f(x|\theta)\Lambda(\theta|\lambda)\, d\theta < \infty$ （确保 marginal 是 pdf，或能被 normalize 成 pdf）

注：尽管名字为 improper prior，它可以是一个 prior 的 proper choice.

e.g. ① 对于 location family with location parameter μ 且 $\Theta = R$，一个常用的 improper prior 为 $\Lambda(\mu) \equiv 1$

② 对于 scale family with scale parameter σ 且 $\Theta = (0, \infty)$，一个常用的 improper prior 为 $\Lambda(\sigma) = \frac{1}{\sigma}$

注意到我们仍需 specify hyperparameter λ. 一个常用的方法是 empirical Bayes

6. Hyperparameter λ 的选取 & parametric empirical Bayes

对于一个 sample X drawn from $f(x|\theta)$，令 prior 为 $G = \{\Lambda(\theta|\lambda), \lambda \in \tilde{\Lambda}\}$ 中的一个 element，则

- sample conditional on the hyperparameter 的 marginal distribution （likelihood）为

  $$\tilde{\pi}(\lambda|x) = \tilde{f}(x|\lambda) = \int f(x|\theta)\Lambda(\theta|\lambda)\, d\theta$$

- 因此，一个 λ 的 simple choice 便是 MLE，即

  $$\hat{\lambda} = \arg\max_{\lambda \in \tilde{\Lambda}} \tilde{\pi}(\lambda|x)$$

- 因此，基于 empirical Bayes approach 的 prior 为

  $$\pi(\theta|\hat{\lambda})$$

注：MLE 的求解也可以使用 profile likelihood / generalize profile likelihood.