Handout 1
Intro to Bayesian approach. Decision Theory. *

Instructor: Vira Semenova Note author: Vira Semenova

1 Definitions and Notation

Frequentist world: θ is fixed parameter, X is a random draw (sample)

- X denotes a random variable, e.g. $X \sim U[0,1]$ whose specific realizations are x
- $\theta \in \Theta$ is a target parameter
- $f_X(x \mid \theta)$ is the likelihood function. Examples
  - (Discrete). A coin flip with success probability $\theta \in [0,1]$ has PMF
$$f_X(x \mid \theta) = \begin{cases} \theta, & x = 1, \\ 1-\theta & x = 0 \end{cases}$$
  - (Continuous) A $N(\theta, 1)$ has PDF
$$f_X(x \mid \theta) = \frac{1}{\sqrt{2\pi}} \exp^{-(\theta - x)^2/2}.$$
- Decision space $\mathcal{A}$ with actions $a \in \mathcal{A}$. Examples:
  - Estimation. Goal is to produce a that is a "best guess" of θ. The quality of the guess is evaluated by the loss function
$$L(a, \theta) = (a - \theta)^2$$
  - Testing. The null hypothesis is
$$H_0 : \theta \in \Theta_0$$
and the alternative is
$$H_1 : \mu \in \Theta_1.$$
The decision space is $\mathcal{A} = \{1, 0\}$, where $a = 0$ means accept and $a = 1$ means reject. The parameter space is $\{H_0$ is true $, H_1$ is true $\}$ The loss function is
$$L(a, H) = \begin{cases} 0 & a = j \text{ correct} \\ c_1 & a = 1 \text{ and } H_0 \text{ is true; type 1 error} \\ c_2 & a = 0 \text{ and } H_1 \text{ is true; type 2 error} \end{cases}$$
  - Decision rule is a mapping from the space $\mathcal{X}$ (the space of sample realizations) into $\mathcal{A}$. Examples:
    - Estimation:
      - $\delta_0(X) = 0$ (no-data rule)

*We thank Prof. Anna Mikusheva and numerous former GSIs for their course notes.

  - $\delta_1(X) = X$ (identity rule)
  - $\delta(X) = \theta$ is NOT a decision rule (depends on the unknown parameter θ
* Testing. $\delta(X) = \begin{cases} |X| > 1.96, & a = 0 \\ |X| < 1.96 & a = 1 \end{cases}$
* Risk function is the ex-ante expected loss (before the sample realization takes place)

$$R(\delta, \theta) = \mathbb{E}\left[L(\delta(X), \theta)\right] = \int_{\mathcal{X}} L(\delta(x), \theta) f_X(x \mid \theta) dx.$$

* Ideal frequentist decision rule is infeasible

$$\delta^* = \arg\min_{\delta \text{ all rules}} \; R(\theta_0, \delta)$$

because θ_0 is unknown

Bayesian concepts:

- $\pi(\theta)$ is a prior weighting function on Θ. It summarizes the prior belief about θ before observing data.
- $\pi(\theta \mid X = x)$ is the posterior belief about θ after observing $X = x$.
- Bayes rule states

$$\pi(\theta \mid X = x) \sim \frac{f_X(x \mid \theta) \cdot \pi(\theta)}{\int_\Theta f_X(x \mid \theta) \cdot \pi(\theta) d\theta}$$

2 Examples of conjugate priors.

2.1 Beta-Binomial family.

- $X = (X_1, X_2, \ldots, X_n)$ is an i.i.d sample from Bernoulli (p) distribution with realizations $x = (x_1, \ldots, x_n)$.
- Prior $\pi(p) \sim U[0, 1]$. (All values of p are equally likely).
- The likelihood of observation i, $i = 1, 2, \ldots, n$ is

$$f(X_i = x_i | p) = \begin{cases} p, & x_i = 1 \\ 1 - p, & x_i = 0 \end{cases}$$

- The full likelihood of $X = x$ data is

$$f(X|p) = p^{\sum_i x_i} (1 - p)^{n - \sum_i x_i}$$

- The posterior is

$$\pi(p|X) \sim \frac{p^{\sum_i x_i} (1 - p)^{n - \sum_i x_i}}{\int p^{\sum_i x_i} (1 - p)^{n - \sum_i x_i} dp} = \frac{p^{\sum_i x_i} (1 - p)^{n - \sum_i x_i}}{B(\sum_i x_i + 1, n - \sum_i x_i + 1)}$$

- Posterior distribution is Beta(s, f) with

$$s = \sum_i x_i, \quad f = n - (\sum_i x_i) + 1.$$

- Posterior mean is

$$\mathbb{E}\left[p \mid X = x\right] = \frac{\sum_i x_i}{n + 2} \to \widehat{p}_{MLE}, \quad n \to \infty$$

- Take-aways:
  - For Bernoulli likelihood, conjugate prior is
$$\pi(p) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{\mathrm{B}(\alpha,\beta)} \sim \mathrm{Beta}(\alpha,\beta)$$
  - Bayesian update to get the posterior is
$$\begin{aligned}\pi(p|X) &\sim p^{\sum_i x_i}(1-p)^{n-\sum_i x_i} \cdot \frac{p^{\alpha-1}(1-p)^{\beta-1}}{\mathrm{B}(\alpha,\beta)} \\ &\sim p^{\sum_i x_i+\alpha-1}(1-p)^{n-\sum_i x_i+\beta-1}\end{aligned}$$
    coincides with $\mathrm{Beta}(\alpha+s, \beta+f)$.
  - $U[0,1]$ is a special case of $\mathrm{Beta}(1,1)$. Note that
$$\pi(p) \sim p^{1-1}(1-p)^{1-1}1\{p \in [0,1]\}.$$

2.2 Normal-normal family.

The model is
$$Y_i = X_i'\theta + U_i$$
with $U_i \sim iidN(0,1)$. In matrix form we'll write
$$Y = X\theta + U,$$
where $Y = (Y_1, Y_2, \ldots, Y_n)$, $\quad U = (U_1, U_2, \ldots, U_n)$ and X is $n \times k$ matrix whose rows are
$$X_{i,\cdot} = (X_{i1}, X_{i2}, \ldots, X_{ik}), \quad i = 1, 2, \ldots, n.$$
The conditional likelihood is
$$f(Y|X,\theta) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2}(Y - X\theta)'(Y - X\theta)\right)$$
If we choose a normal prior, $\theta \sim N(0, \tau^2 I_k)$,
$$\pi(\theta) = (2\pi\tau^2)^{-k/2} \exp\left(\frac{-1}{2\tau^2}\theta'\theta\right)$$
. The posterior is
$$\begin{aligned}\pi(\theta|Y,X) &\propto \exp\left(-\frac{1}{2}\left[-Y'X\theta - \theta'X'Y + \theta'X'X\theta + \frac{1}{\tau^2}\theta'\theta\right]\right) \\ &\propto \exp\left(-\frac{1}{2}\left[-Y'X\theta - \theta'X'Y + \theta'(X'X + \frac{I_k}{\tau^2})\theta\right]\right) \\ &\propto \exp\left(-\frac{1}{2}\left[\left(\theta - (X'X + \frac{I_k}{\tau^2})^{-1}X'Y\right)'(X'X + \frac{I_k}{\tau^2})\left(\theta - (X'X + \frac{I_k}{\tau^2})^{-1}X'Y\right)\right]\right)\end{aligned}$$
so $\theta|Y,X \sim N(\tilde{\theta}, \tilde{\Sigma})$ with
$$\begin{aligned}\tilde{\theta} &= (X'X + \frac{I_k}{\tau^2})^{-1}X'Y \\ \tilde{\Sigma} &= (X'X + \frac{I_k}{\tau^2})^{-1}\end{aligned}$$
Also, we see that as $\tau \to \infty$ (uninformative prior), $\tilde{\theta} \to (X'X)^{-1}X'Y = \hat{\theta}^{ML}$, and as $\tau \to 0$, $\tilde{\theta} \to 0$, the prior dominates. Furthermore, if we fix τ and $T \to \infty$ with $\frac{X'X}{T} \to Q_{XX}$, then $\tilde{\theta} \to \theta_0$, the frequentist limit. So, prior vanishes asymptotically. This result is more general.

3 Frequentist Decision theory.

3.1 Admissibility

Definition 1. *A decision rule, δ, is* admissible *if there exists no $\tilde{\delta}$ such that*

$$R(\delta, \theta) \geq R(\tilde{\delta}, \theta) \forall \theta$$

with strict inequality for some θ_0. Otherwise, δ is inadmissible

Example 1. *Decision set $\mathcal{A} = \mathrm{R}$. Loss function is $0-1$ loss is*

$$L(\theta, a) = 0 \text{ if } \theta = a \text{ and } L(\theta, a) = 1 \text{ otherwise}$$

D.g.p. model is

$$f_\theta(X = \theta - 1) = f_\theta(X = \theta + 1) = 0.5, \theta \in \mathrm{R}$$

Decision rule δ_0:

$$\delta_0(x_1, x_2) = \frac{x_1 + x_2}{2}$$

Decision rule δ_1:

$$\delta_1(x_1, x_2) = x_1 - 1$$

The frequentist risk $R(\theta, \delta_0)$ of the rule δ_0 is

$$R(\theta, \delta_0) = 0.5 \quad \forall \theta$$

- *$\delta_0(x_1, x_2)$ is correct if $x_1 \neq x_2$. $L(\theta, \delta_0(x_1, x_2) | x_1 \neq x_2) = 0$*
- *$\delta_0(x_1, x_2)$ is wrong if $x_1 = x_2$. $L(\theta, \delta_0(x_1, x_2) | x_1 = x_2) = 1$*
- *The frequentist risk of δ_0 is $0.5 \cdot 0 + 0.5 \cdot 1 = 0.5$*

The frequentist risk $R(\theta, \delta_1)$ of the rule δ_0 is

$$R(\theta, \delta_1) = \Pr(x_1 - 1 \neq \theta) = \Pr(x_1 \neq \theta + 1) = 0.5$$

The decision rules δ_0 and δ_1 are inadmissible because both of them are dominated by

$$\delta^\pi(x_1, x_2) = \begin{cases} \dfrac{x_1 + x_2}{2}, & x_1 \neq x_2 \\ x_1 - 1, & x_1 = x_2 \end{cases}$$

whose risk $R(\theta, \delta^\pi) = 1/4$.

Example 2. *Decision set $\mathcal{A} = \mathrm{R}$. Loss function is $L(a, \theta) = (a - \theta)^2$. Observed $X \sim Dis(\mu, \sigma^2)$ where σ^2 is known and $\theta := \mu$ is the unknown (target) parameter. The class of rules is*

$$\mathcal{F}' = \{\beta X, \quad \beta \in \mathrm{R}\}.$$

Examples:

- *$\beta = 0, \quad \delta_0(X) = 0$: no data rule.*
- *$\beta = 1, \quad \delta_1(X) = X$: identity map*

- $\beta = -1, \quad \delta_1(X) = -X$*: exotic rule*

The risk of rule $\delta(X) = \beta X$ *is*

$$R(\delta, \theta) = (\beta - 1)^2\mu^2 + \beta^2\sigma^2.$$

The optimal thing to do is shrink X *towards zero*

$$\beta^*(\mu, \sigma) = \frac{\mu^2}{\mu^2 + \sigma^2},$$

where the shrinkage coefficient is determined by signal-to-noise ratio. However,

$$\delta = \frac{\mu^2}{\mu^2 + \sigma^2} X$$

is not a rule because μ *is unknown! However,* $\frac{\mu^2}{\mu^2 + \sigma^2} \in [0, 1]$*, which allows us (verify yourself) that the set of* ***admissible*** *decision rules is*

$$\mathcal{F} = \{\beta X, \quad \beta \in [0, 1]\}.$$

3.2 Minimaxity

Definition 2. *The worst -case risk is*

$$\bar{R}(\delta) = \sup_\theta R(\delta, \theta)$$

Minimax decision rule is

$$\delta^* = \arg\min_\delta \bar{R}(\delta) = \arg\min_\delta \sup_\theta R(\delta, \theta)$$

Theorem 1 (Relationship between admissibility and minimaxity)**.** *If* δ^* *is admissible with constant risk, then it is a minimax decision rule.*

Proof by contraposition. Suppose δ^* is admissible with constant risk but not minimax. Then, for some δ',

$$\bar{R}(\delta') < \bar{R}(\delta^*).$$

Then,

$$R(\delta', \theta') \leq \sup_\theta R(\delta', \theta) <^a \sup_\theta R(\delta, \theta) =^b R(\delta^*, \theta'),$$

which contradicts admissibility. Here, we used admissibility in (a) and constant risk assumption in (b).

3.3 Bayesian approach

Suppose we have a loss function $L(a, \theta)$, where θ is a parameter and a is some action that we want to choose. For example, if we just want to estimate θ, we might have $a = \hat{\theta}$ and $L(a, \theta) = (a - \theta)^2$. Our goal is to come up with a decision rule $a(X)$ that depends on the sample X, and give a small loss. Let our expected loss(called *risk*) for a given value of θ be

$$R_a(\theta) = E_\theta L(a(X), \theta)$$

Note, that it is frequentist notion (expectation is taken over repeated samples)!!! In example above the risk is MSE. We would like $a(X)$ to minimize our expected loss. However, in general, the solution will depend on θ.

Definition 3. *A* Bayesian Decision Rule *minimizes a weighted risk (with weights $\pi(\theta)$):*

$$\begin{aligned}\min_a \int R(a,\theta)\pi(\theta)d\theta &= \min_a \int E_\theta L(\delta(X),\theta)\pi(\theta)d\theta \\ &= \min_a \int\int L(a(X),\theta)f(X|\theta)\pi(\theta)dXd\theta \\ &= \min_a \int \left[\int L(a(X),\theta)\pi(\theta|X)d\theta\right] p(X)dX \\ &= \int \left[\min_a \int L(a(X),\theta)\pi(\theta|X)d\theta\right] p(X)dX\end{aligned}$$

Thus, the Bayes decision function $\delta^*(x)$ is obtained by minimizing $\min_a \int L(a(X),\theta)\pi(\theta|X)d\theta$ for each realization of X!

Theorem 2. *All Bayesian decision rules are admissible. Also, under some conditions, all admissible decision rules are Bayesian.*

4 Summary: Reasons to be Bayesian

1. (Philosophical.) Two observations, $X_1, X_2 \sim$ iid.

   $$P_\theta(X_i = \theta - 1) = \frac{1}{2} = P_\theta(X_i = \theta + 1)$$

   Consider a confidence set:

   $$C(y_1, y_2) = \begin{cases} \frac{y_1+y_2}{2} & y_1 \neq y_2 \\ y_1 - 1 & y_1 = y_2 \end{cases}$$

   If we observe $y_1 \neq y_2$, which will happen 1/2 the time, we know $\theta = C(y_1, y_2)$. Otherwise, we have a probability of 1/2 that $\theta = C(y_1, y_2)$. From a frequentist perspective, then the coverage of this set is $1/2 + 1/2 * 1/2 = 75\%$. Now, suppose we observe $y_1 \neq y_2$. Then we know θ with certainty. Why would we then report a coverage of 75% (ex-ante coverage) rather than the ex-post accuracy of 100%? Frequentists average probabilities over all situations that may have been realized, but were not. Bayesians are conditioning on the realization. As this example shows, conditioning on observation may be justified.

2. (Prior vanishes asymptotically.)

   The prior vanishes asymptotically. All inference in "clean" situations asymptotically converge to frequentists'. This claim is based on the following theorem:

   Theorem 3. *(Geweke, p.93) Suppose*

   (a) Prior is absolutely continuous wrt to Lebesgue measure and prior puts positive probability on all sets with positive Lebesgue measure

   (b) Uniform convergence of likelihood $\frac{1}{n}\log f(X_i|\theta) \to^{a.s.} l(\theta)$ uniformly in θ

   (c) $l(\theta)$ is continuous and has a unique maximum at θ^*

   Then for any open neighborhood, $\varepsilon(\theta^*)$,

   $$\lim_{n\to\infty} \pi(\theta \in \varepsilon(\theta^*)|X) = 1 \text{ a.s.}$$

   Theorem says that the posterior concentrates around the asymptotic limit of frequentist MLE, and in any "reasonabl" situation Baeys estimate in large samples will be close to MLE. There's a similar theorem that shows asymptotic normality of the Bayesian estimator. These two theorems sort of say what happens when you use Bayesian methods in a frequentist world. One interpretation is that the prior vanishes asymptotically. However, you should be cautious about this theorem!
   Cautions:

- The theorem is about asymptotics. However, the prior can influence inferences in finite samples.
- Condition 3 is an identification condition. If you are not identified, then where the Bayesian estimator converges depends on your prior.
- Prior should not restrict parameter space (condition 1)
- Condition 2 is like a LLN, it may not be satisfied with non-stationarity

3. (Decision Theory). All Bayesian rules are admissible. Under some conditions, all admissible rules are Bayesian.

   Theorem 4 (Complete class theorem)**.** *All Bayesian decision rules are admissible. Also, under some conditions, all admissible decision rules are Bayesian.*

4. (Nuisance parameters.) Let $\omega = h(\theta)$. Let $C(X)$ be a set such that $P(\omega \in C(X)|X) = 1 - \alpha$. It is very easy to go from $\pi(\theta|X)$ to $p(\omega|X)$. For example, suppose $\theta = (\theta_1, \theta_2)$ and $\omega = \theta_1$, then

$$\pi(\theta_1|X) = \int \pi(\theta_1, \theta_2|X)d\theta_2$$

   This example is especially relevant because there are many examples in econometrics where we want to eliminate nuisance parameters. Here, θ_2 would be the nuisance parameters. The Bayesian approach makes it very easy to deal with the nuisance parameters. Whereas in the frequentist world, nuisance parameters are an extremely difficult problem.

5 Exercises

Exercise 1. Derive the Bayesian decision rules for the following problems:

1. Quadractic loss function $L(a, \theta) = (a - \theta)^2$. Show that Bayesian decision rule is the **POSTERIOR MEAN**:

$$\arg\min \int_\Theta L(a, \theta)\pi(\theta \mid X = x)d\theta := \mathbb{E}\left[\theta \mid X = x\right]$$

   Hint: take FOC with respect to a and expand:

$$\begin{aligned}\int_\Theta L(a, \theta)\pi(\theta \mid X = x)d\theta &= a^2 \int_\Theta \pi(\theta \mid X = x)d\theta - 2a \int_\Theta \theta\pi(\theta \mid X = x)d\theta + \int_\Theta \theta^2\pi(\theta \mid X = x)d\theta. \\ &= a^2 \cdot 1 - 2a\mathbb{E}\left[\theta \mid X = x\right] + \int_\Theta \theta^2\pi(\theta \mid X = x)d\theta\end{aligned}$$

2. Absolute loss function $L(a, \theta) = |a - \theta|$. Show that Bayesian decision rule is the **POSTERIOR MEDIAN**:

$$\arg\min \int_\Theta L(a, \theta)\pi(\theta \mid X = x)d\theta := \delta^*(X)$$

   such that

$$\int_{\theta<\delta^*(X)} \pi(\theta \mid X = x)d\theta = 1/2, \quad \int_{\theta>\delta^*(X)} \pi(\theta \mid X = x)d\theta = 1/2.$$

3. Testing decision problem. The loss function is

$$L(a, H) = \begin{cases} 0 & a = j \text{ correct} \\ c_1 & a = 1 \text{ and } H_0 \text{ is true; type 1 error} \\ c_2 & a = 0 \text{ and } H_1 \text{ is true; type 2 error} \end{cases}$$

Show that Bayesian decision rule is

$$\delta^*(X) = \begin{cases} 1, & \Pr(\theta \in \Theta_1 \mid X = x) \geq \dfrac{c_1}{c_1 + c_2} \\ 0, & \text{otherwise} \end{cases}$$

Hint: Bayesian risk is

$$\begin{aligned} &0 + \delta(x)c_1 \Pr(\theta \in \Theta_0 \mid X = x) + (1 - \delta(x))c_2 \Pr(\theta \in \Theta_1 \mid X = x) \\ &= \delta(x)c_1(1 - \Pr(\theta \in \Theta_1 \mid X = x)) + (1 - \delta(x))c_2 \Pr(\theta \in \Theta_1 \mid X = x) \\ &= c_1\delta(x) - (c_1 + c_2)\delta(x) \Pr(\theta \in \Theta_1 \mid X = x) + c_2 \Pr(\theta \in \Theta_1 \mid X = x). \end{aligned}$$

Exercise 2.

1. You observe $X \sim N(\theta, 1)$
2. You impose the prior $\theta \sim N(0, \tau^2)$
3. For any decision rule δ, calculate: (a) the posterior risk $\rho(\delta, x) := \mathbb{E}\left[L(\delta(x), \theta) \mid X = x\right]$ and (b) Bayesian decision rule $\delta^*(x) = \arg\min \rho(\delta, x)$.

Exercise 3.

1. You observe $X_i \in \{1, 2, \ldots, k\}$ with $Pr(X_i = j) = \theta_j$
2. You impose the Dirichlet prior $\pi(\theta) \sim \prod_{j=1}^{k} \theta_j^{\alpha_j - 1}$
3. For any decision rule δ, calculate: (a) the posterior distribution $\pi(\theta \mid X = x)$ and (b) the posterior mean $\mathbb{E}\left[\theta \mid X = x\right]$. (See Wikipedia!)