$$p(\Theta|D) = \frac{p(D|\Theta)p(\Theta)}{\int p(D|\Theta)p(\Theta)d\Theta}$$

Bayesian Learning 732A46: Lecture 4

Matias Quiroz[1,2]

[1]Division of Statistics and Machine Learning, Linköping University

[2]Research Division, Sveriges Riksbank

April 2016

Lecture overview

- **Prediction**
  - Normal model
  - Complex predictions by simulation
- **Decision theory**
  - The elements of a decision problem
  - The Bayesian way
  - Point estimation as a decision problem

Prediction/Forecasting

- **Posterior predictive distribution** for future $\tilde{y}$ given observed data y

$$p(\tilde{y}|y) = \int_\theta p(\tilde{y}, \theta|y)d\theta = \int_\theta p(\tilde{y}|\theta, y)p(\theta|y)d\theta.$$

- **Note**: Averages $p(\tilde{y}|\theta, y)$ over the posterior distribution $p(\theta|y) \implies$ predictions **take into account the parameter uncertainty**.

- **Simplified** if $p(\tilde{y}|\theta, y) = p(\tilde{y}|\theta)$ [not true for time series], then

$$p(\tilde{y}|y) = \int_\theta p(\tilde{y}|\theta)p(\theta|y)d\theta.$$

- **Easy** to simulate (marginalization by simulation)

$$\begin{aligned} \theta^{(i)} &\sim p(\theta|y) \\ \tilde{y}^{(i)}|\theta^{(i)} &\sim p(\tilde{y}|\theta^{(i)}) \end{aligned}$$

- **Histogram** (or **Kernel density estimate**) of $\tilde{y}^{(i)}$ is an approximation of $p(\tilde{y}|y)$.

Prediction - Normal data, known variance

- Our old friend

$$y_i|\theta \overset{iid}{\sim} \mathcal{N}(\theta, \sigma^2) \quad [\textbf{known } \sigma^2]$$

- The **posterior predictive**

$$p(\tilde{y}|y) = \int_\theta p(\tilde{y}|\theta)p(\theta|y)d\theta,$$

where, if $p(\theta) \propto c$ (**uniform** prior),

$$\begin{aligned} \theta|y &\sim \mathcal{N}(\bar{y}, \sigma^2/n) \\ \tilde{y}|\theta &\sim \mathcal{N}(\theta, \sigma^2) \end{aligned}$$

1. Generate a posterior draw of θ $[\theta^{(1)}]$ from $\mathcal{N}(\bar{y}, \sigma^2/n)$
2. Generate a draw of $\tilde{y}$ $[\tilde{y}^{(1)}]$ from $\mathcal{N}(\theta^{(1)}, \sigma^2)$ (note the mean)
3. **Repeat** Steps 1 and 2 a large number of times (N) with the result:
   - Sequence of posterior draws: $\theta^{(1)}, ..., \theta^{(N)}$
   - Sequence of predictive draws: $\tilde{y}^{(1)}, ..., \tilde{y}^{(N)}$.

Predictive distribution - Normal model and uniform prior

- In this simple model it is **easy to derive** $p(\tilde{y}|y)$ analytically.
- Note that

Step 1. $\theta^{(i)} = \bar{y} + \omega^{(i)}, \quad \omega^{(i)} \sim \mathcal{N}(0, \sigma^2/n)$

Step 2. $\tilde{y}^{(i)} = \theta^{(i)} + \varepsilon^{(i)}, \quad \varepsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$

- $\varepsilon^{(i)}$ and $\upsilon^{(i)}$ are independent.
- The sum of two normal r.v's is normal so $p(\tilde{y}|y)$ is normal,

$$
\begin{aligned}
E(\tilde{y}|y) &= \bar{y} \\
V(\tilde{y}|y) &= \frac{\sigma^2}{n} + \sigma^2 = \sigma^2\left(1 + \frac{1}{n}\right)
\end{aligned}
$$

$$
\tilde{y}|y \sim \mathcal{N}\left(\bar{y}, \sigma^2\left(1 + \frac{1}{n}\right)\right).
$$

Predictive distribution - Normal model and normal prior

- Assume still that σ^2 **is known**, but

$$p(\theta) = \mathcal{N}(\theta|\mu_0, \tau_0^2) \implies p(\theta|y) = \mathcal{N}(\theta|\mu_n, \tau_n^2)$$

Step 1. $\theta^{(i)} = \mu_n + \omega^{(i)}, \quad \omega^{(i)} \sim \mathcal{N}(0, \tau_n^2)$

Step 2. $\tilde{y}^{(i)} = \theta^{(i)} + \varepsilon^{(i)}, \quad \varepsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$

with (which **you know** by **heart** now!)

$$\frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2} \quad \text{and} \quad \mu_n = \left(\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}\right) \Big/ \frac{1}{\tau_n^2}.$$

- It easy to see that the **predictive distribution** is normal.
- **With mean** [**Tower Property** or **Law of total (conditional) expectation**]

$$E\left(\tilde{y}|y\right) = E_{\theta|y}\left(E_{\tilde{y}|\theta,y}\left(\tilde{y}|\theta, y\right)\right) = E_{\theta|y}\left(\underbrace{E_{\tilde{y}|\theta}\left(\tilde{y}|\theta\right)}_{\theta}\right) = \mu_n$$

- Note that $\tilde{y}$ and y are **conditionally independent given** θ.

Predictive distribution - Normal model and normal prior, cont

- **And variance** [**Law of total (conditional) variance** + $p(\tilde{y}|\theta, y) = p(\tilde{y}|\theta)$]

$$\begin{aligned} V(\tilde{y}|y) &= E_{\theta|y}[V_{\tilde{y}|\theta}(\tilde{y}|\theta)] + V_{\theta|y}[E_{\tilde{y}|\theta}(\tilde{y}|\theta)] \\ &= E_{\theta|y}(\sigma^2) + V_{\theta|y}(\theta) \\ &= \sigma^2 + \tau_n^2 \\ &= (\textbf{Population variance} + \textbf{Posterior variance} \text{ of } \theta). \end{aligned}$$

- In **summary**:

$$\tilde{y}|y \sim \mathcal{N}(\mu_n, \sigma^2 + \tau_n^2).$$

Bayesian prediction in a more complex model

- **Autoregressive process**

$$y_t = \phi_1(y_{t-1} - \mu) + ... + \phi_p(y_{t-p} - \mu) + \varepsilon_t, \ \varepsilon_t \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

- Note that $\tilde{y}$ and y are **not conditionally independent given** θ
- **Why not**?
  - **Conditional independence** means that **if I know** θ, I can simulate

    $$\tilde{y} \sim p(\tilde{y}|\theta, y) = p(\tilde{y}|\theta),$$

    i.e. **without caring** about y.
  - Let $p = 1$ and suppose we want $\tilde{y}_{T+1}$. Let $\theta = (\phi_1, \mu, \sigma)$ be given, then

    $$\tilde{y}_{T+1} = \phi_1(y_T - \mu) + \varepsilon_T, \quad \varepsilon_T \sim \mathcal{N}(0, \sigma^2)$$
  - We need $y_T \subset y$. **They can't be independent, even if we know** θ!
- **No worries**, we can still do predictions (slightly more to keep in mind).

Bayesian prediction in a more complex model, cont.

- **Autoregressive process**

$$y_t \;=\; \phi_1(y_{t-1}-\mu)+...+\phi_p(y_{t-p}-\mu)+\varepsilon_t,\; \varepsilon_t \overset{iid}{\sim} \mathcal{N}(0,\sigma^2)$$

- K-step ahead prediction of $\tilde{y}$ - **"roll simulation forward K-steps"**.
- **Simulate** a draw from $p(\phi_1,\phi_2,...,\phi_p,\mu,\sigma|y)$
  - Conditional on that draw $\theta^{(1)}=(\phi_1^{(1)},\phi_2^{(1)},...,\phi_p^{(1)},\mu^{(1)},\sigma^{(1)})$, simulate
  - $\tilde{y}_{T+1}\sim p(y_{T+1}|y_T,y_{T-1},...,y_{T+1-p},\theta^{(1)})$
  - $\tilde{y}_{T+2}\sim p(y_{T+2}|\tilde{y}_{T+1},y_T,...,y_{T+2-p},\theta^{(1)})$

    $\vdots$
  - $\tilde{y}_{T+K}\sim p(y_{T+K}|\tilde{y}_{T+K-1},\tilde{y}_{T+K-2},...,y_{T+K-p},\theta^{(1)})$ [if $K\leq p$, otherwise $\sim$]
- **Repeat** for new θ draws.

Decision Theory

- **Brief** introduction. See the **excellent** Berger (2013) book.
- Let $\theta \in \Theta$ be an **unknown quantity**. **State of nature**.
  Examples: *Future inflation, Global temperature, Disease*.
- Let $a \in \mathcal{A}$ be an **action**. **Examples**: *Interest rate, Energy tax, Surgery*.
- **Choosing action** a (=decision) when state of nature turns out to be θ gives **utility**

$$U(a, \theta)$$

- Alternatively **loss** $L(a, \theta) = -U(a, \theta)$.
- **Loss table**:

	θ_1	θ_2
a_1	$L(a_1, \theta_1)$	$L(a_1, \theta_2)$
a_2	$L(a_2, \theta_1)$	$L(a_2, \theta_2)$

- **Example**:

	Rainy	Sunny
Umbrella	20	10
No umbrella	50	0

- **The decision problem**: Choose **an action** a that **minimizes the loss**.

Decision Theory, cont.

- Example **loss functions** when both a and θ are continuous:
  - Linear: $L(a, \theta) = |a - \theta|$
  - Quadratic: $L(a, \theta) = (a - \theta)^2$
  - Lin-Lin:

$$L(a, \theta) = \begin{cases} c_1 \cdot |a - \theta| & \text{if } a \leq \theta \\ c_2 \cdot |a - \theta| & \text{if } a > \theta \end{cases}$$

- **Example**:
  - θ is the **number of items** demanded of a product
  - a is the **number of items** in stock
  - Loss

$$L(a, \theta) = \begin{cases} 10 \cdot (\theta - a) & \text{if } a \leq \theta \text{ [too little stock]} \\ 1 \cdot (a - \theta) & \text{if } a > \theta \text{ [too much stock]} \end{cases}.$$

  - We are **punished** by a factor of 10 for keeping **too little** in stock.

Optimal decision

- Bayesian choice: maximize the **posterior expected utility**:

$$a_{bayes} = \mathrm{argmax}_{a \in \mathcal{A}} \; E_{\theta|y} \left(U(a, \theta) \right),$$

where $E_{\theta|y}$ denotes the **posterior expectation,**

$$E_{\theta|y} \left(U(a, \theta) \right) = \int_{\theta \in \Theta} U(a, \theta) p(\theta|y) d\theta$$

- **Easy** to estimate by simulation (**LLN**):

$$E_{\theta|y} \left(U(a, \theta) \right) \approx \frac{1}{N} \sum_{i=1}^{N} U(a, \theta^{(i)}) \quad \theta^{(i)} \sim p(\theta|y)$$

- **Note**: we could have **minimized** the **posterior expected loss** .

Choosing a point estimate is a decision

- Choosing a **point estimator** is a decision problem.
- Possible **action space**

$$\mathcal{A} = \{\theta_{\text{median}}, \theta_{\text{mode}}, \theta_{\text{mean}}\}.$$

- Which one is the **optimal choice**?
- **It depends on the loss function**:
  - **Linear loss** $\rightarrow$ **Posterior median** is optimal
  - **Quadratic loss** $\rightarrow$ **Posterior mean** is optimal
  - **Lin-Lin loss** $\rightarrow$ $c_2/(c_1 + c_2)$ **posterior quantile** is optimal
  - **Zero-one loss** $\rightarrow$ **Posterior mode** is optimal

References

Berger, J. (2013). *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media.