BAYESIAN LEARNING - LECTURE 4

Mattias Villani

Division of Statistics and Machine Learning
Department of Computer and Information Science
Linköping University

Lecture overview

- Prediction
  - Bernoulli model
  - Normal model
  - More complex examples
- Decision theory
  - The elements of a decision problem
  - The Bayesian way
  - Point estimation as a decision problem

Prediction

- Using the estimated model for **forecasting** a future observation $\tilde{y}$.
- **Posterior predictive distribution** (y denotes available data at the time of forecasting)

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

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

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

- The uncertainty that comes from not knowing θ is represented in $p(\tilde{y}|y)$ by **averaging over** $p(\theta|y)$.

Prediction - Bernoulli data

- Let $y = \sum_{i=1}^{n} y_i$ and $\tilde{y}$ the outcome of the next trial

$$
\begin{aligned}
p(\tilde{y} &= 1|y) = \int_\theta p(\tilde{y} = 1|\theta)p(\theta|y)d\theta \\
&= \int_\theta \theta p(\theta|y)d\theta = E_{\theta|y}(\theta) = \frac{\alpha + y}{\alpha + \beta + n}.
\end{aligned}
$$

- Uniform prior ($\alpha = \beta = 1$)

$$
p(\tilde{y} = 1|y) = \frac{y + 1}{n + 2}.
$$

Prediction - Normal data, known variance

- Under the uniform prior $p(\theta) \propto c$, then

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

where

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

Prediction - Normal data, known variance

- Under the uniform prior $p(\theta) \propto c$, then

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

where

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

1. Generate a posterior draw of θ ($\theta^{(1)}$) from $N(\bar{y}, \sigma^2/n)$
2. Generate a draw of $\tilde{y}$ ($\tilde{y}^{(1)}$) from $N(\theta^{(1)}, \sigma^2)$ (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

- $\theta^{(1)} = \bar{y} + \varepsilon^{(1)}$, where $\varepsilon^{(1)} \sim N(0, \sigma^2/n)$. (Step 1).
- $\tilde{y}^{(1)} = \theta^{(1)} + \upsilon^{(1)}$, where $\upsilon^{(1)} \sim N(0, \sigma^2)$. (Step 2).
- $\tilde{y}^{(1)} = \bar{y} + \varepsilon^{(1)} + \upsilon^{(1)}$.
- $\varepsilon^{(1)}$ and $\upsilon^{(1)}$ are independent.
- The sum of two normal random variables follows a normal distribution, so $\tilde{y}$ follows a normal distribution with

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

- Note that the estimation uncertainty (σ^2/n) is typically much less important than the intrinsic population uncertainty, σ^2.

Predictive distribution - Normal model and normal prior

- It easy to see that the predictive distribution is normal.
- The mean can be obtained from

$$E_{\tilde{y}|\theta}(\tilde{y}) = \theta$$

and then remove the conditioning on θ by averaging over θ

$$E(\tilde{y}|y) = E_{\theta|y}(\theta) = \mu_n \text{ (Posterior mean of } \theta).$$

- The predictive variance of $\tilde{y}$ (conditional variance formula):

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

- In summary:

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

Bayesian prediction in more complex models

- **Autoregressive process**

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

- 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
  - $\tilde{y}_{T+1} \sim p(y_{T+1} | y_T, y_{T-1}, ..., y_{T-p}, \theta^{(1)})$
  - $\tilde{y}_{T+2} \sim p(y_{T+2} | \tilde{y}_{T+1}, y_T, ..., y_{T-p}, \theta^{(1)})$
  - and so on.
- Repeat for new θ draws.

Bayesian prediction in more complex models

- **Autoregressive process**

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

- 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
  - $\tilde{y}_{T+1} \sim p(y_{T+1} | y_T, y_{T-1}, ..., y_{T-p}, \theta^{(1)})$
  - $\tilde{y}_{T+2} \sim p(y_{T+2} | \tilde{y}_{T+1}, y_T, ..., y_{T-p}, \theta^{(1)})$
  - and so on.
- Repeat for new θ draws.

- **Regression trees**.
  - Uncertainty on which variables to split on, and the split point.
  - For given draw of splitting variables and split points, simulate a response. Repeat for many different draws.

Predicting auction prices on eBay

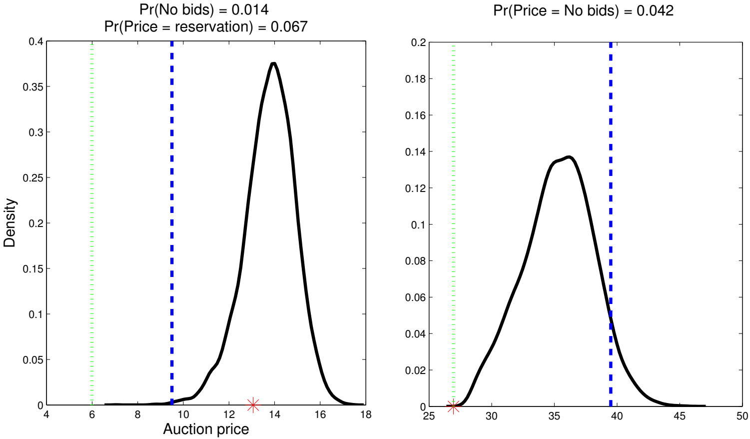
- Problem: Predicting the auctioned price in eBay coin auctions.
- Data: Bid from 1000 auctions on eBay. The highest bid is not observed. The lowest bids are also not observed because of the seller's reservation price.
- Covariates: auction-specific, e.g. Book value from catalog, seller's reservation price, quality of sold object, rating of seller, powerseller, verified seller ID etc
- Buyers are strategic. Their bids does not fully reflect their valuation. Game theory needed in the econometric model. **Very** complicated likelihood.

Simulating auction prices on eBay, cont.

- A draw from the posterior predictive distibution of an auction's price:

1. Simulate a draw $\theta^{(1)}$ from the posterior of the model parameters θ
2. Simulate the number of bidders conditional on θ (which contains the intensity parameter of a Poisson process)
3. Simulate a complete auction bid sequence, $\mathbf{b}^{(1)}$, conditional on $\theta = \theta^{(1)}$, for the bidders generated in Step 2.
4. For the bid sequence $\mathbf{b}^{(1)}$, return the next to largest bid (eBay's proxy bidding system).

Predicting auction prices on eBay, cont.



Decision Theory

- Let θ be an unknown quantity. State of nature. Examples: Future inflation, Global temperature, Disease.
- Let $a \in \mathcal{A}$ be an action. Ex: Interest rate, Energy tax, Surgery.
- Choosing action a 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	0	-10
No umbrella	-50	20

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
  - Utility
  $$U(a,\theta) = \begin{cases} p \cdot \theta - c_1(a-\theta) & \text{if } a > \theta \text{ [too much stock]} \\ p \cdot a - c_2(\theta - a)^2 & \text{if } a \leq \theta \text{ [too little stock]} \end{cases}$$

Optimal decision

- Ad hoc decision rules:
  - *Minimax*. Choose the decision that minimizes the maximum loss.
  - *Minimax-regret*: Choose the decision rule that gives you least regret when you eventually find out the true value of θ.
  - bla bla bla ...
- Bayesian **theory**: Just maximize the **expected utility**:

$$a_{bayes} = \mathrm{argmax}_{a \in \mathcal{A}} \ E_{p(\theta|y)}[U(a,\theta)],$$

where $E_{p(\theta|y)}$ denotes the posterior expectation.

- Using simulated draws $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(N)}$ from $p(\theta|y)$:

$$E_{p(\theta|y)}[U(a,\theta)] \approx N^{-1} \sum_{i=1}^{N} U(a, \theta^{(i)})$$

- **Separation principle**: The analysis of uncertainty (i.e. the posterior of θ) is completely separated from the utilities of the choices.

Choosing a point estimate is a decision

- Choosing a point estimator is a decision problem.
- Which to choose: posterior median, mean or mode?
- It depends on your loss function:
  - **Linear loss** $\rightarrow$ Posterior median is optimal
  - **Quadratic loss** $\rightarrow$ Posterior mean is optimal
  - **Lin-Lin loss** $\rightarrow$ $c_2/(c_1 + c_2)$ quantile of the posterior is optimal
  - **Zero-one loss** $\rightarrow$ Posterior mode is optimal