T-61.5040 Oppivat mallit ja menetelmät
T-61.5040 Learning Models and Methods
Pajunen, Viitaniemi

Exercises 11, 13.4.2007

Problem 1.

You want to know the weight θ of an object and you have a scale which reports the weight $\hat{\theta}$ with a Normally distributed error, where the likelihood is $N(\hat{\theta}|\theta, 1)$. For simplicity, assume all priors are constant so you can directly use likelihoods as posteriors.

i) What is the posterior of θ if you measure the object 100 times and 9 times the scale reports 'no value'? Assume that the probability for reporting 'no value' is unknown, but independent of the missing values.

ii) Change the 'no value' report to 'over 100 kg' report. What is the posterior of θ?

iii) Repeat part ii), but this time you don't know how many missing values there are. You only have 91 observed values.

Problem 2.

What is the missing-data mechanism in the following situations? Is the mechanism MAR, MCAR, or OAR? Comment on the ignorability of the missing data mechanism in each case.

i) You observe 50 results out of 100 independent coin flips. Denote the probability of heads with θ. The decision not to observe the result is random (with probability ϕ) and independent of the coin flips.

ii) 100 people are asked if they drink alcohol, and their gender (male/female) is also recorded. Assume that men are somewhat more likely to refuse to answer. The probability to refuse to answer is directly defined by the gender. Now the measurement y has two components: a (alcohol) and x (gender).

iii) Let the data y consist of independent samples of $N(y_i|\theta)$. Suppose θ is either 0.2 or 0.4. Each sample is observed randomly with probability $(1-\theta)$, otherwise it remains missing. In a particular experiment you observe 20 data samples and 80 observations are left missing.

iv) 100 people are asked how much money they earn yearly. Those that have an income less than the national average are more likely to refuse to answer.

Problem 3.

Assume that data $y = (y_{obs}, y_{mis})$ with missing components is Normally distributed, i.e. $y(i) \sim N(\mu, \Sigma)$ and $y(i) \in \mathbb{R}^d$. The data may contain vectors $y(i)$ with any number of missing components. Write the Data Augmentation/Gibbs sampler algorithm for simulating $p(y_{mis}, \mu, \Sigma | y_{obs}, I)$. Assume that you have already solved the Gibbs sampler for the same model *without missing data*, and that the missing data mechanism is ignorable and $p(I|y, \phi) = p(I|y_{obs}, \phi)$.

Hint: recall that if $p(u, v)$ is a Normal distribution, then $p(u)$, $p(v)$, and $p(u|v)$ are Normal distributions and

$$\begin{aligned} \mathrm{E}(u|v) &= E(u) + \mathrm{Cov}(v, u)\mathrm{Var}(v)^{-1}(v - \mathrm{E}(v)) \\ \mathrm{Var}(u|v) &= \mathrm{Var}(u) - \mathrm{Cov}(v, u)\mathrm{Var}(v)^{-1}\mathrm{Cov}(u, v) \end{aligned}$$