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

Solutions to exercise 4, 9.2.2007

Problem 1.

i) Likelihood is irrelevant when considering the prior. The problem does not give any reason to distinguish θ_1 and θ_2. Therefore $p(\theta_1) = p(\theta_2) = 0.5$ is a reasonable choice for the prior distribution.

ii) Use $p(\theta_i|x = j) \propto p(x = j|\theta_i)p(\theta_i)$ for $i = 1, 2$. Then normalize the distribution.

$p(\theta_1|x = 0) \propto p(x = 0|\theta_1)p(\theta_1) = 0.8 \cdot 0.5$ and $p(\theta_2|x = 0) \propto 0.4 \cdot 0.5$, normalize $\rightarrow$
$p(\theta_1|x = 0) = 2/3, \quad p(\theta_2|x = 0) = 1/3$
$p(\theta_1|x = 1) \propto 0.2 \cdot 0.5$ and $p(\theta_2|x = 1) \propto 0.6 \cdot 0.5$, normalize $\rightarrow$
$p(\theta_1|x = 1) = 1/4, \quad p(\theta_2|x = 1) = 3/4$

iii) $p(\theta|x_1, \dots, x_n) \propto p(x_1, \dots, x_n|\theta)p(\theta)$

Since the data are independent given θ, the likelihood is $\prod_j p(x_j|\theta_i)$. Each term depends on the binary value of x_j but is independent of the index j. Therefore we may compute the sum $\overline{x} = \sum_j x_j$ and find that the likelihood is

$$\prod_j p(x_j|\theta_i) = (p(x = 1|\theta_i))^{\overline{x}}(p(x = 0|\theta_i))^{n-\overline{x}}$$

This can be computed for θ_1 and θ_2 to get the unnormalized posterior. Normalization gives the result.

The sum $\overline{x} = \sum_j x_j$ is the single variable that determines the answer. This kind of a variable is called a *sufficient statistic.*

Problem 2

Likelihood says that $x < \theta$ and prior says that $\theta \leq 1$, so compute the posterior with these conditions.

i) $p(\theta|x) \propto p(x|\theta)p(\theta) = 2x/\theta^2$

Normalization constant: $p(x) = \int p(x|\theta)p(\theta)d\theta = \int_x^1 2x\theta^{-2}d\theta = 2x(x^{-1} - 1)$

Result: $p(\theta|x) = \theta^{-2}/(x^{-1} - 1)$

ii) $p(\theta|x) \propto p(x|\theta)p(\theta) = 2x/\theta^2 \cdot 3\theta^2 = 6x$

This is independent of θ so the posterior is constant on the interval $(x, 1]$. This gives $p(\theta|x) = 1/(1 - x)$, $\theta \in (x, 1]$

iii) $E(\theta|x)$ can be directly computed using the posterior. Using the posterior in i),

$$E(\theta|x) = \int \theta p(\theta|x)d\theta = \int_x^1 \theta^{-1}d\theta/(x^{-1}-1) = (\log 1-\log x)/(x^{-1}-1) = (\log x)/(1-x^{-1})$$

Using the posterior in ii),

$$E(\theta|x) = \int_x^1 \theta/(1 - x)d\theta = (1 + x)/2$$

Problem 3.

i) The two Normal distributions are far enough from each other that they can be considered separately. Either $\theta = 0$ or $\theta = 4$ approximately maximizes the posterior. Compute the posterior at both points, ignoring the other Normal distribution:

$p(\theta = 0|D) \approx 0.9 * N(0|0, 1) = 0.9/(\sqrt{2\pi})$
$p(\theta = 4|D) \approx 0.1 * N(4|4, 0.1^2) = 0.1/(\sqrt{2\pi}0.1)$

Dividing by the the square-root, the first value is 0.9 and the second $0.1/0.1 = 1$. So the posterior is maximized around the value $\theta = 4$. The predicted value is approximately $y = \exp(4) \approx 55$. By looking at the prior probabilities of the Normal distributions you might think $\theta = 0$ is a better choice. But $\theta = 4$ has a smaller variance, so picking a point estimate results in choosing $\theta = 4$.

ii) θ is not normally distributed, but it is a mixture of two Normal distributions. However, we can use the fact that integration is a linear operation and we can thus calculate the integrals, i.e. expectations over the mixture component distributions separately. For both of the component distributions, y is lognormal, and by the given hint $E[y] = \exp(\mu+\sigma^2/2)$:

$$\mathrm{E}(y) = \mathrm{E}(\exp(\theta)) = 0.9 \exp(0 + \frac{1}{2} \cdot 1) + 0.1 \exp(4 + \frac{1}{2} \cdot 0.01) \approx 7$$

Now the mean value is closer to $\exp(1/2)$ than $\exp(4.005)$. In part i), the predicted value was $\exp(4) \approx 55$.

Problem 4.

i) In this problem we need to average models for different values of θ. Model averaging does not mean that the values $\tilde{y}$ predicted by the models are averaged: if this was so, then the average would be $\tilde{y} = 2$, but all models give zero probability for this value. Instead, the "probability mass" given by each model for each value of $\tilde{y}$ is averaged over the models, using the posterior probabilities of the models as weights. The average probability mass for a specific value $\tilde{y}$ is

$$p(\tilde{y}|D) = \sum_{i=1}^{3} p(\tilde{y}|\theta_i, D)p(\theta_i|D).$$

The value $\tilde{y} = 1$ gets probability mass $p(\tilde{y} = 1|\theta_1, D)p(\theta_1|D) = 1/2$ and $\tilde{y} = 3$ gets $p(\tilde{y} = 3|\theta_2, D)p(\theta_2|D)+p(\tilde{y} = 3|\theta_3, D)p(\theta_3|D) = 1/2$: all terms that are zero were omitted here.

ii) Using the above formula, we obtain

$$\begin{aligned} p(\tilde{y}=1|D) &= \sum_{i=1}^{3} p(\tilde{y}=1|\theta_i, D)p(\theta_i|D) \\ &= \frac{1}{2}\cdot 0.35 + \frac{1}{4}\cdot 0.5 + \frac{1}{4}\cdot 0.1 = 0.325. \end{aligned}$$

Similarly,

$$p(\tilde{y}=2|D) = \frac{1}{2}\cdot 0.3 + \frac{1}{4}\cdot 0.4 + \frac{1}{4}\cdot 0.4 = 0.35$$

and

$$p(\tilde{y}=3|D) = \frac{1}{2}\cdot 0.35 + \frac{1}{4}\cdot 0.1 + \frac{1}{4}\cdot 0.5 = 0.325.$$

Each model predicts that the most probable $\tilde{y}$ is either 1 or 3. However, the predictive distribution is maximized for $\tilde{y}=2$. Looking at the maximum of a distribution may thus be misleading!