10-602: Statistical Approaches to Learning and Discovery

Problem Set 2 Solutions

---

1. (a)

$$
\begin{aligned}
Pr(z \le \hat{z}) &= Pr(\ln(x) \le \hat{z}) = Pr(x \le e^{\hat{z}}) && (1) \\
p(z = \hat{z}) &= p(x = e^{\hat{z}})e^{\hat{z}} = \begin{cases} 0.5 & -1 \le \hat{z} \le 1 \\ 0 & \text{otherwise} \end{cases} && (2)
\end{aligned}
$$

(b)

$$
\begin{aligned}
Pr(y \le \hat{y}) &= Pr(1/x \le \hat{y}) = Pr(x \ge 1/\hat{y}) && (3) \\
p(y = \hat{y}) &= \frac{1}{\hat{y}^2}p(x = 1/\hat{y}) = \begin{cases} 0.5/\hat{y} & 1/e \le \hat{y} \le e \\ 0 & \text{otherwise} \end{cases} && (4)
\end{aligned}
$$

The mode of $p(y = \hat{y})$ is 1/e, which is also the mode of $p(x = \hat{x})$.

(c) It is tempting to argue that every continuous value has probability zero and therefore there cannot be a most likely one. But we can define the mode in terms of infinitesimal intervals and escape this argument. A better argument is suggested by the previous exercise: The mode cannot be interpreted as the most likely value because then x and y would have contradictory most likely values. This contradiction cannot be escaped by using intervals.

The only interpretation we can give to the mode is that it is the value chosen by decision theory under zero-one loss. The example also demonstrates that if you are interested in y, you should not apply decision theory to x first—you should apply it directly to the distribution of y. Otherwise you may get exactly the opposite answer.

2. (a) Let $\mathbf{y} = \mathbf{Ax}$.

$$
\begin{aligned}
\text{cov}(\mathbf{y}) &= E[(\mathbf{y} - E[\mathbf{y}])(\mathbf{y} - E[\mathbf{y}])^{\mathrm{T}}] && (5) \\
&= E[(\mathbf{Ax} - \mathbf{Am})(\mathbf{Ax} - \mathbf{Am})^{\mathrm{T}}] && (6) \\
&= \mathbf{A}E[(\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^{\mathrm{T}}]\mathbf{A}^{\mathrm{T}} && (7) \\
&= \mathbf{A\Sigma A}^{\mathrm{T}} && (8)
\end{aligned}
$$

(b) $\text{tr}(\mathbf{AB}) = \sum_{ij} a_{ij}b_{ji}$ which is symmetric in $\mathbf{A}$ and $\mathbf{B}$.

(c)

$$
\begin{aligned}
E[\mathbf{x}^{\mathrm{T}}\mathbf{Ax}] &= E[\text{tr}(\mathbf{x}^{\mathrm{T}}\mathbf{Ax})] = E[\text{tr}(\mathbf{Axx}^{\mathrm{T}})] && (9) \\
&= \text{tr}(\mathbf{A}E[\mathbf{xx}^{\mathrm{T}}]) = \text{tr}(\mathbf{A}(\mathbf{\Sigma} + \mathbf{mm}^{\mathrm{T}})) && (10) \\
&= \text{tr}(\mathbf{A\Sigma}) + \mathbf{m}^{\mathrm{T}}\mathbf{Am} && (11)
\end{aligned}
$$

3. (a) Each independent sample has an equal chance of being less than θ or greater than θ. The probability that both are less than θ is 1/4, the probability that both are greater than θ is 1/4, so the remaining probability, that θ is between the samples, is 1/2.

(b) The distribution of the two samples is

$$
p(y_1|\theta)p(y_2|\theta) = \begin{cases} \frac{1}{4} & \text{if } y_1 \ge \theta - 1 \text{ and } y_2 \le \theta + 1 \\ 0 & \text{otherwise} \end{cases} \qquad (12)
$$

The posterior for θ is $U(y_2 - 1, y_1 + 1)$, which is centered on $(y_1 + y_2)/2$ and has width $2 - |y_2 - y_1|$. The first half of this range, namely $(y_2 - 1, (y_1 + y_2)/2)$, is a valid 50% interval, as are many others.

(c) When $|y_2 - y_1|$ is large, the posterior for θ is narrow—we have a lot of information about θ. Thus the Bayesian confidence interval is small. But the frequentist confidence interval is large. When $|y_2 - y_1|$ is small, we have little information about θ, and the Bayesian interval is large, but the frequentist interval is small. In this problem it is hard to justify using the frequentist interval.

4. (a) Define $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$.

$$p(m_1, m_2|D_1, D_2) \propto p(D_1|m_1)p(D_2|m_2)p(m_1)p(m_2) \quad (13)$$

$$\propto \exp(-\frac{1}{2v_1}\sum_i (x_i - m_1)^2)\exp(-\frac{1}{2v_2}\sum_i (x_i - m_2)^2)\exp(-\frac{m_1^2 + m_2^2}{2v_0}) \quad (14)$$

This distribution factors, so m_1 and m_2 are independent:

$$p(m_1|D_1) = \mathcal{N}(\bar{x}, v_1/n) \quad (15)$$

$$p(m_2|D_2) = \mathcal{N}(\bar{y}, v_2/n) \quad (16)$$

(b) Since m_1 and m_2 are independent and Gaussian given the data, b is also Gaussian given the data:

$$p(b|D_1, D_2) \sim \mathcal{N}(\bar{x} - \bar{y}, v_1/n + v_2/n) \quad (17)$$

The posterior for (m_1, m_2) summarizes everything the data has to say about (m_1, m_2)—we do not have to refer back to the data in order to reason about b. The probability that $m_1 < m_2$ is the probability that $b < 0$. Define g so that

$$b = g\sqrt{v_1/n + v_2/n} + \bar{x} - \bar{y} \quad (18)$$

The distribution of g is $\mathcal{N}(0, 1)$. Now

$$p(b < 0|D_1, D_2) = p\left(g < \frac{\bar{y} - \bar{x}}{\sqrt{v_1/n + v_2/n}}\right) = \phi\left(\frac{\bar{y} - \bar{x}}{\sqrt{v_1/n + v_2/n}}\right) \quad (19)$$

This function is close to 1 when $\bar{x} << \bar{y}$, as we expect. The distance between $\bar{y}$ and $\bar{x}$ is normalized by the noise variance, which is also reasonable.

(c) All of the previous equations were implicitly conditioned on v_1 and v_2. If v_1 and v_2 are unknown with posterior distribution $p(v_1, v_2|D_1, D_2)$, we average the result over their values:

$$p(b < 0|D_1, D_2) = \int_0^\infty \int_0^\infty p(b < 0|v_1, v_2, D_1, D_2)p(v_1, v_2|D_1, D_2)dv_1 dv_2 \quad (20)$$

$$= \int_0^\infty \int_0^\infty \phi\left(\frac{\bar{y} - \bar{x}}{\sqrt{v_1/n + v_2/n}}\right) p(v_1, v_2|D_1, D_2)dv_1 dv_2 \quad (21)$$

This can be written even more succinctly by realizing that only the sum $v_1 + v_2$ appears in the integrand. Define $w = v_1 + v_2$, whose posterior distribution $p(w|D_1, D_2)$ can be determined from $p(v_1, v_2|D_1, D_2)$. Then

$$p(b < 0|D_1, D_2) = \int_0^\infty \phi(\frac{\bar{y} - \bar{x}}{\sqrt{w/n}})p(w|D_1, D_2)dw \quad (22)$$

This form makes it easy to apply numerical integration routines.

5. (a) The probability of the data is

$$p(a_1, a_2|\theta)p(b_1, b_2|\theta) = \frac{1}{2\pi v_a}\exp(-\frac{1}{2v_a}\sum_i (a_i - \theta)^2)\frac{1}{2\pi v_b}\exp(-\frac{1}{2v_b}\sum_i (b_i - \theta)^2) \quad (23)$$

Define

$$\bar{a} = \frac{1}{2}\sum_i a_i \qquad\qquad \bar{b} = \frac{1}{2}\sum_i b_i \tag{24}$$

$$S_a = \sum_i (a_i - \bar{a})^2 \qquad\qquad S_b = \sum_i (b_i - \bar{b})^2 \tag{25}$$

The maximum-likelihood equations are

$$\hat{\theta} = \frac{2\bar{a}/\hat{v}_a + 2\bar{b}/\hat{v}_b}{2/\hat{v}_a + 2/\hat{v}_b} \tag{26}$$

$$\hat{v}_a = \frac{1}{2}\sum_i (a_i - \hat{\theta})^2 \tag{27}$$

$$\hat{v}_b = \frac{1}{2}\sum_i (b_i - \hat{\theta})^2 \tag{28}$$

which cannot be solved in closed form. As an approximation, let

$$\hat{v}_a = S_a/2 = 0.16 \tag{29}$$

$$\hat{v}_b = S_b/2 = 0.36 \tag{30}$$

With these estimates, the posterior distribution for θ is

$$p(\theta|\hat{v}_a, \hat{v}_b, D) = \mathcal{N}(\hat{\theta}, \frac{1}{2/\hat{v}_a + 2/\hat{v}_b}) = \mathcal{N}(2.1154, 0.0554) \tag{31}$$

(The exact MLEs, by the way, can be found by fixed-point iteration of (26)-(28):

$$\hat{v}_a = 0.1662 \tag{32}$$

$$\hat{v}_b = 4.0509 \tag{33}$$

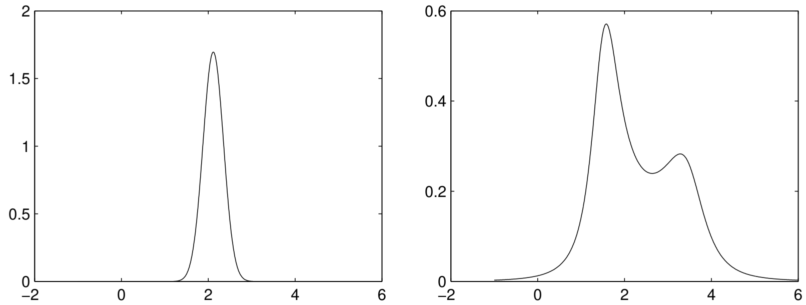
$$p(\theta|\hat{v}_a, \hat{v}_b, D) = \mathcal{N}(1.5788, 0.0798) \tag{34}$$

This solution essentially ignores sensor b.)

(b) The posterior distribution for θ is

$$p(\theta|D) \propto \left(\int_0^\infty p(a_1, a_2|\theta, v_a)p(v_a)dv_a\right)\left(\int_0^\infty p(b_1, b_2|\theta, v_b)p(v_b)dv_b\right) \tag{35}$$

$$\propto \frac{1}{2(\theta-\bar{a})^2 + S_a}\,\frac{1}{2(\theta-\bar{b})^2 + S_b} \tag{36}$$



(a) Approximate posterior for θ (b) Exact posterior for θ

(c) The exact posterior has two modes, one near $\bar{a}$ (which is 1.5) and the other near $\bar{b}$ (which is 3.5). They correspond to a belief that v_b is large (sensor 2 is unreliable) or v_a is large (sensor 1 is unreliable). By contrast, the approximate posterior in part (a) assumes that v_a and v_b are approximately equal, i.e. the sensors are equally reliable, leading a very different inference about θ (not only in mean, but variance as well). This is the essential dilemma in sensor fusion: should we weight the sensors equally or should we trust some more than others? The approximate posterior weights each sensor solely according to its measurement variance. The exact posterior is more intelligent—it considers how much the sensors agree. If they seem to agree, it will weight them equally and take a linear combination. (You can see this by choosing data which is less separated.) If the sensors seem to disagree, it will consider the possibility that one is unreliable. With more than two sensors, things get more interesting, with groups of sensors essentially voting against other sensors.