ПРИКЛАДНАЯ СТАТИСТИКА

Каждой Маше по три медведя! *

23 февраля 2021 г.

Идет медведь по лесу, видит, машина горит. Сел в нее и сгорел.

Анекдот категории F

Упражнение 1 (Маша и медведи)

Маша прячется от Медведей в точке m на числовой прямой. Есть несколько Медведей, каждый из которых обнюхивает всю числовую прямую в поисках Маши. Медведю номер i кажется, что Машей сильней всего пахнет в точке y_i. Естественно, Медведи могут ошибаться, например, у них может быть заложен нос, поэтому **модель Медведя выглядит как:**

$$y_i \mid m \sim \mathcal{N}(m, 2^2).$$

При фиксированном m величины y_i независимы. Известно, что $y_1 = 0.5, y_2 = -1$. Априорно известно, что место, где спряталась Маша имеет нормальное распределение, $m \sim \mathcal{N}(1, 4^2)$. Нам нужно:

а. Найти апостериорную плотность распределения параметра m.

б. Найти апостериорные моду, медиану и математическое ожидание.

в. Найти $\mathbb{P}(m > 1 \mid y_1, y_2)$.

*Эта pdf-ка, по факту, представляет из себя кусочек недописанной виньетки по Байесовским методам: https://github.com/FUlyankin/book_about_bayes г. Найти $f(y_3 \mid y_1, y_2)$ и $\mathbb{E}(y_3 \mid y_1, y_2)$.

Решение:

Посмотрим немного подробнее на наше априорное мнение о том, где сидит Маша, $m \sim \mathcal{N}(1, 4^2)$. Значение 1 в данном случае — наше лучшее предположение о том, где она может находиться, а 4^2, в свою очередь, это наша степень доверия к этому предположению. Чем меньшее значение дисперсии мы берём в нашем априорном мнении, тем больше наше доверие к нему.

Делай раз! Апостериорная плотность Маши:

$$
\begin{aligned}
f(m \mid y_1, y_2) \propto f(y_1, y_2 \mid m) \cdot f(m) = \frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{(0.5-m)^2}{2\cdot 4}\right) \cdot \\
\cdot \frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{(-1-m)^2}{2\cdot 4}\right) \cdot \frac{1}{4\sqrt{2\pi}} \exp\left(-\frac{(m-1)^2}{2\cdot 16}\right)
\end{aligned}
$$

Воспользуемся магической силой уже привычного нам значка $\propto$ и для простоты расчётов пренебрежём кучей констант

$$
f(m \mid y_1, y_2) \propto \exp\left(-\frac{(0.5-m)^2}{2\cdot 4}\right) \cdot \exp\left(-\frac{(-1-m)^2}{2\cdot 4}\right) \cdot \exp\left(-\frac{(m-1)^2}{2\cdot 16}\right).
$$

Сольём всё,что находится под знаком экспоненты в единое целое и упростим

$$
\begin{aligned}
\frac{(0.5-m)^2}{2\cdot 4} + \frac{(-1-m)^2}{2\cdot 4} + \frac{(m-1)^2}{2\cdot 16} = \\
= \frac{4(m-0.5)^2 + 4(m+1)^2 + (m-1)^2}{32} = \frac{9m^2+2m+6}{32}
\end{aligned}
$$

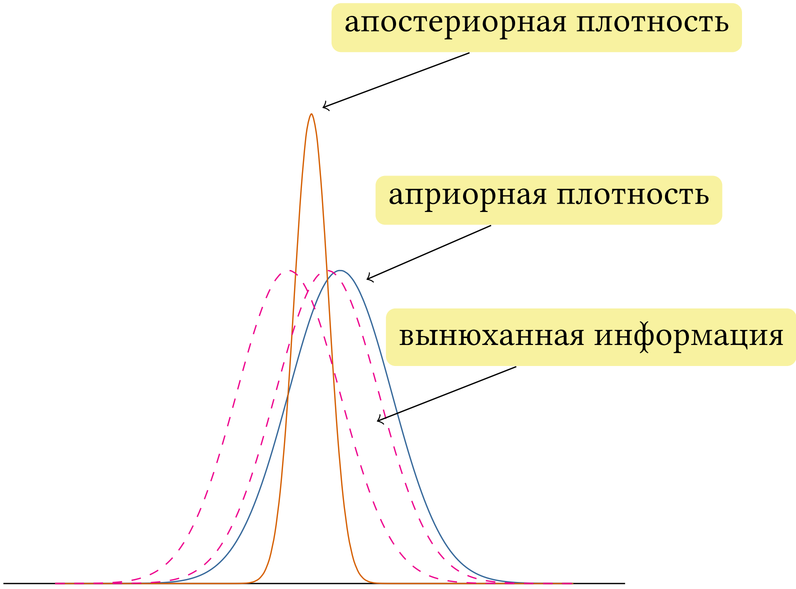
Используем двойную магию. С одной стороны пренебрегаем константой, с другой создаём новую для того, чтобы выделить полный квадрат. Не забываем перекинуть в знаменатель лишнюю девятку

$$
\begin{aligned}
\exp\left(-\frac{9m^2+2m+6}{32}\right) \propto \exp\left(-\frac{9m^2+2m}{32}\right) = \\
= \exp\left(-\frac{m^2+\frac{2}{9}m}{\frac{32}{9}}\right) = \exp\left(-\frac{m^2+2\cdot\frac{1}{9}m+\frac{1}{81}-\frac{1}{81}}{\frac{32}{9}}\right) \propto \\
\propto \exp\left(-\frac{m^2+2\frac{1}{9}m+\frac{1}{81}}{\frac{32}{9}}\right) = \exp\left(-\frac{(m+\frac{1}{9})^2}{2\cdot({}^{4}/{}_{3})^2}\right)
\end{aligned}
$$

Видим, что параметр m имеет нормальное апостериорное распределение

$$m \mid y_1, y_2 \sim \mathcal{N}(-{}^1/_9, ({}^4/_3)^2).$$

При желании можно восстановить константу. Обратите внимания, что после того как Медведи попытались вынюхать, где находится Маша, самое вероятное её положение изменилось, а дисперсия её положения уменьшилась.



Картинка 1: Информация о Маше

Новая информация сместила априорную плотность влево и вытянула её вверх, в силу того, что Медведи вынюхали похожие вещи.

Делай два! Мода и медиана для нормального распределения совпадают с математическим ожиданием. Мы можем использовать эти величины в качестве точечных оценок.

Делай три! Обратите внимание, что до запуска Медведей, $\mathbb{P}(m > 1) = 0.5$. После запуска, эта вероятность уменьшится,так как распределение очень сильно съедет влево.

$$\mathbb{P}(m > 1 \mid y_1, y_2) = 1 - \mathbb{P}(m \leqslant 1 \mid y_1, y_2) = \\ = 1 - \mathbb{P}\left(\frac{m + {}^1/_9}{{}^4/_3} \leqslant \frac{1 + {}^1/_9}{{}^4/_3} \mid y_1, y_2\right) = 1 - \Phi\left(\frac{10}{12}\right) \approx 0.2.$$

Значение функции $\Phi(z)$ можно получить, воспользовавшись таблицами для стандартной нормально распределённой случайной величины. Либо её можно найти с помощью компьютера.

Делай четыре! Найдём $f(y_3 \mid y_1, y_2)$ и $\mathbb{E}(y_3 \mid y_1, y_2)$. Будем делать это под слоганом: «Каждой Маше по три Медведя!»:

$$f(y_3 \mid y_1, y_2) = \int_{-\infty}^{+\infty} f(y_3, m \mid y_1, y_2)\, dm = \int_{-\infty}^{+\infty} f(y_3 \mid y_1, y_2, m) \cdot f(m \mid y_1, y_2)\, dm.$$

Под знаком интеграла мы получаем произведение модели и апостериорного распределения. Чтобы найти плотности распределение y_3, мы должны провести свёртку по двум нормальным распределениям

$$\begin{aligned} f(y_3 \mid y_1, y_2) &= \int_{-\infty}^{+\infty} \mathcal{N}(m, 4) \cdot \mathcal{N}\left(-\frac{1}{9}, \frac{16}{9}\right) \, dm = \\ &= \int_{-\infty}^{+\infty} \frac{1}{2\sqrt{2\pi \cdot 4^2}} \, exp\left(-\frac{(y-m)^2}{2 \cdot 4^2}\right) \cdot \frac{1}{2\sqrt{2\pi \cdot (^{16}/_9)^2}} \, exp\left(-\frac{(m + ^1/_9)^2}{2 \cdot (^{16}/_9)^2}\right) \, dm \propto \\ &\propto \int_{-\infty}^{+\infty} exp\left(-\frac{(y-m)^2}{2 \cdot 4^2} - \frac{(m + ^1/_9)^2}{2 \cdot (^{16}/_9)^2}\right) \, dm \end{aligned}$$

Если аккуратно взять этот интеграл и восстановить константу, можно получить, что $y_3 \mid y_1, y_2 \sim \mathcal{N}\left(-\frac{1}{9}, \frac{52}{9}\right)$. Если нам необходим точечный прогноз и в качестве функции потерь выбрано MSE, мы можем выбрать число $-\frac{1}{9}$.

Теперь, когда нюхательные способности третьего Машиного Медведя предсказаны, вы можете попробовать проделать всё то же самое самое, предположив, что вам вообще ничего неизвестно и $m \sim \mathcal{U}(-\infty; +\infty)$. В таком случае в качестве плотности нужно будет взять $f(m) = 1$. Стоит отметить, что результат у вас, при этом, получится похожим на случай нормального априорного распределения с большой дисперсией. Эти два распределения обладают довольно большой энтропией. Из-за этого получаются схожие результаты. Также попытайтесь провернуть процедуру байесовского вывода для нормального распределения в общем случае. Формулы выйдут довольно громоздкими. Если запутаетесь, загляните в решебник.