Летняя школа по байесовским методам в глубинном обучении

Семинар 1. Байесовские методы

26 августа 2017 г.

1. В результате медицинского обследования один из тестов выявил у человека серьезное заболевание. Данный тест имеет высокую точность: вероятность позитивного ответа при наличии заболевания составляет 99%, вероятность отрицательного ответа при отсутствии заболевания также составляет 99%. Однако, выявленное заболевание является достаточно редким и встречается только у одного человека на 10000. Вычислить вероятность того, что у обследуемого человека действительно есть выявленное заболевание.

2. На школу было отобрано N_a участников от академии и N_b участников от индустрии. Кому-то больше нравится решать задачи, а кому-то меньше. Поэтому в первой группе на семинар приходят с вероятностью α, во второй – с вероятностью β. Первые успешно решают задачи с вероятностью γ_a, а вторые – с вероятностью γ_b. Каждую из N задач суммарно решило $c_1, \ldots, c_N$ людей, но при опросе руку подняло лишь $d_1, \ldots, d_N$ человек. Рассмотрим следующую вероятностную модель для описанной ситуации:

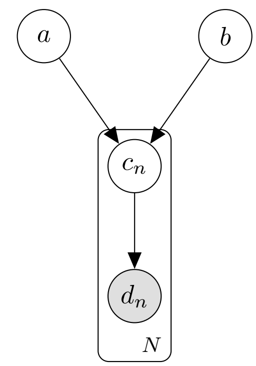
$$p(a, b, c_1, \ldots, c_N, d_1, \ldots, d_N) = p(a)p(b) \prod_{n=1}^{N} p(d_n|c_n)p(c_n|a, b),$$

$$d_n|c_n \sim \mathrm{Bin}(c_n, \delta),$$

$$c_n|a, b \sim \mathrm{Bin}(a, \gamma_a) + \mathrm{Bin}(b, \gamma_b),$$

$$a \sim \mathrm{Bin}(N_a, \alpha),$$

$$b \sim \mathrm{Bin}(N_b, \beta).$$



Здесь через $\mathrm{Bin}(N, q)$ обозначено биномиальное распределение с N испытаниями и вероятностью успеха q.

Зная результаты опросов $d_1, \ldots, d_N$, мы хотим оценить число пришедших на семинар представителей академии a и индустрии b. Предложите формулы для подсчета $p(a|d_1, \ldots, d_N)$ и $p(b|a, d_1, \ldots, d_N)$ с помощью элементарных распределений $p(d_n|c_n), p(c_n|a, b), p(a), p(b)$. Постарайтесь избежать суммирований по нескольким группам переменных одновременно.

3. Пусть $x_1, \ldots, x_N$ — результаты подбрасывания монетки, выпадающей орлом с вероятностью q. Найти оценку максимального правдоподобия для q.т

Пусть помимо этого параметр q имеет априорное бета-распределение $p(q|\alpha, \beta) = \mathrm{Beta}(\alpha, \beta)$:

$$p(q|\alpha, \beta) = \frac{q^{\alpha-1}(1-q)^{\beta-1}}{B(\alpha, \beta)}, \ \alpha, \beta > 0, \ Eq = \frac{\alpha}{\alpha + \beta};$$

$$p(x_1, \ldots, x_N, q|\alpha, \beta) = p(q|\alpha, \beta) \prod_{i=1}^{N} p(x_i|q).$$

Здесь через $B(\alpha, \beta)$ обозначена бета-функция. Требуется найти апостериорное распределение на q $p(q|x_1, \ldots, x_N, \alpha, \beta)$.

4. Апостериорное распределение может также быть использовано для поиска точечной оценки параметров. Зафиксируем функцию потерь L и будет выбирать параметр $\hat{\theta}$ согласно правилу

$$\hat{\theta} = \mathrm{argmin}_{\hat{\theta}} \mathbb{E}_{p(\theta|x_1, \ldots, x_N)} L(\theta, \hat{\theta}). \tag{1}$$

Какие оценки даст апостериорное распределение параметра монетки q из предыдущей задачи для квадратичной функции потерь $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$? Для модуля $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$?

5. Вычислите обоснованность $p(x_1, \ldots, x_N|\alpha, \beta)$ модели из задачи 3.

6. Пусть случайный вектор ε имеет многомерное стандартное нормальное распределение. Известно, что случайный вектор $z = A\varepsilon + b$ также имеет нормальное распределение. Здесь A и b — некоторые матрица и вектор соответственно. Требуется определить параметры распределения для z (мат. ожидание и матрицу ковариации).

   Обратно, пусть случайная величина z имеет нормальное распределение

   $$p(z) = \mathcal{N}(z|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2}\det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(z-\mu)^T\Sigma^{-1}(z-\mu)\right).$$

   Для каких пар (A, b) вектор z представим в виде $z = A\varepsilon + b$ с помощью вспомогательной стандартной нормальной случайной величины ε? Какова сложность поиска такого представления?

7. Рассмотрим задачу регрессии. Пусть имеется выборка $(X, y) = \{x_i, y_i\}_{i=1}^N$, состоящая из N объектов. Здесь $x_i \in \mathbb{R}^d$ – признаковое описание i-го объекта, $y_i \in \mathbb{R}$ – значение его целевой переменной. Рассмотрим байесовскую модель линейной регрессии:

   $$\begin{aligned}
   &p(y, w|X, \alpha, \beta) = p(y|w, X, \beta)p(w|\alpha),\\
   &p(y|w, X, \beta) = \mathcal{N}(y|Xw, \beta I),\\
   &p(w|\alpha) = \mathcal{N}(w|0, \alpha I).
   \end{aligned}$$

   Здесь $w \in \mathbb{R}^d$ — вектор весов линейной регрессии, β – параметр шума, α – параметр регуляризации. Требуется найти апостериорное распределение на веса $p(w|X, y, \alpha, \beta)$.

8. Вычислите обоснованность $p(y|X, \alpha, \beta)$ модели из задачи 7.