Лекція 5. Вибір за Гіббсом (семплування Гіббсом, або гібсовський семплер)

Необхідні відомості з теорії ймовірностей

Умовний розподіл дискретних випадкових величин

Нехай заданий ймовірнісний простір $(\Omega, \mathfrak{F}, P)$, $X : \Omega \to R^n$, $Y : \Omega \to R^m$ такі що, випадковий вектор $(X, Y)^T : \Omega \to R^{n+m}$ має дискретний розподіл із функцією ймовірності $p_{X,Y}(x, y)$, $x \in R^n, y \in R^m$. Нехай $y_0 \in R^m$ є таким, що $P(Y = y_0) > 0$. Тоді функція

$$p_{X,Y}(x, y) = P(X = x | Y = y_0) = \frac{p_{X,Y}(x, y_0)}{p_Y(y_0)},$$

де p_Y — функція розподілу випадкової величини Y, називається *умовною функцією ймовірності* випадкової величини X за умови, що $Y = y_0$. Розподіл, що задається умовною функцією ймовірності, називається *умовним розподілом*.

Умовний розподіл абсолютно неперервних випадкових величин

Нехай заданий ймовірнісний простір $(\Omega, \mathfrak{F}, P)$, $X : \Omega \to R^n$, $Y : \Omega \to R^m$ такі що, випадковий вектор $(X, Y)^T : \Omega \to R^{n+m}$ має абсолютно неперервний розподіл із щільністю ймовірності $f_{X,Y}(x, y)$, $x \in R^n, y \in R^m$. Нехай $y_0 \in R^m$ є таким, що , де f_Y — щільність випадкової величини Y. Тоді функція

$$f_{X,Y}(x, y) = \frac{f_{X,Y}(x, y_0)}{f_Y(y_0)}$$

називається *умовною функцією ймовірності* випадкової величини X за умови, що $Y = y_0$. Розподіл, що задається умовною функцією ймовірності, називається *умовним розподілом*.

Випадкова величина із абсолютно неперервним розподілом

Розподіл випадкової величини X називається абсолютно неперервним, якщо існує невід'ємна функція $f_X : R \to R^+$, така що $P(X \in B) = \int_B f_X(x)dx$. Функція f_X називається щільністю розподілу випадкової величини X.

Теорема 1. Якщо функція $f : R \to R$ задовольняє умови: $f(x) \geq 0, x \in R$ і $\int_{-\infty}^{\infty} f(x)dx = 1$, то існує такий розподіл, що функція f є його щільністю.

Теорема. Якщо f — неперервна щільність розподілу, а F — його функція розподілу, то $F'(x) = f(x)$ $\forall x \in R$ і $F(x) = \int_{-\infty}^{x} f(x)dx$.

Сумісний розподіл

Якщо $X_1, X_2, ..., X_n$ — випадкові величини, то кожний вектор $(X_1, X_2, ..., X_n)$ також є випадковою величиною. Його розподіл називається сумісним розподілом випадкових величин $X_1, X_2, ..., X_n$. Функцією розподілу випадкового вектора називається функція

$$F_{X_1, X_2, ..., X_n}(x_1, x_2, ..., x_n) = P(X_1 \le x_1, X_2 \le x_2, ..., X_n \le x_n).$$

Маргінальний розподіл

Для випадку двох дискретних випадкових величин X і Y маргінальна функція ймовірності має вигляд

$$P(X = x) = \sum_y P(X = x, Y = y) = \sum_y P(X = x | Y = y) P(Y = y),$$

де $P(X = x, Y = y)$ — сумісна функція розподілу, а $P(X = x | Y = y)$ — умовна функція розподілу X по Y.

Для неперервних випадкових величин маргінальна функція розподілу має вигляд

$$p_X(x) = \int_y p_{X,Y}(x, y) dy = \int_y p_{X|Y}(x|y) p_Y(y) dy.$$

Інакше кажучи, маргінальна ймовірність X обчислюється шляхом вивчення повної ймовірності X для певного значення Y, а потім усереднення цієї умовної ймовірності над розподілом усіх значень Y.

Марківський ланцюг.

Розглянемо систему, еволюція якої описується випадковим процесом $\{X_n, n = 1, 2, ...\}$. Значення x_n випадкової величини X_n у момент часу t_n називається *миттєвим станом* системи. Простір усіх можливих значень цих випадкових величин називається *простором станів*. Якщо структура стохастичного процесу $\{X_n, n = 1, 2, ...\}$ є такою, що значення залежить виключно від значення x_n, то такий процес називається *марківським ланцюгом*:

$$P(X_{n+1} = x_{n+1} | X_n = x_n, ..., X_1 = x_1) = P(X_{n+1} = x_{n+1} | X_n = x_n).$$

Перехід із стану i в стан j задається *імовірністю переходу*

$$p_{ij}(n) = P(X_{n+1} = j | X_n = i).$$

Оскільки $p_{ij}(n)$ — ймовірності переходу, вони мають задовольняти такі умови:

$$p_{ij}(n) > 0 \quad \forall i, j$$

$$\sum_j p_{ij}(n) = 1 \quad \forall i.$$

Ланцюг Маркова називається *однорідним*, якщо матриця ймовірностей переходу $\{p_{ij}\}$ не залежить від n. Ланцюг Маркова називається *нерозкладним*, якщо з будь-якого його стану можна досягти будь-який інший стан за скінченну кількість кроків. Стан називається *ергодичним*, якщо він аперіодичний та позитивно рекурентним, тобто його період дорівнює 1 і час повернення має скінченне математичне сподівання. Якщо всі стани в нерозкладному ланцюгу Маркова є ергодичними, то такий ланцюг називається *ергодичним*.

Алгоритм Гіббса

Розглянемо *K*-вимірний випадковий вектор X, що складається із компонентів $X_1, X_2, ..., X_K$. Припустимо, що ми знаємо умовний розподіл X_k при заданих значеннях усіх інших компонентів $k = 1, 2, ..., K$. Ми хочемо отримати числову оцінку маргінальної щільності випадкової величини X_k для довільного k. Починаючи із довільної конфігурації $\{x_1(0), x_2(0), ..., x_K(0)\}$, на першій ітерації згенеруємо

$x_1(1)$ із розподілу X_1 при заданих $x_2(0), x_3(0), ..., x_K(0)$;

$x_2(1)$ із розподілу X_2 при заданих $x_1(1), x_3(0), ..., x_K(0)$;

…

$x_k(1)$ із розподілу X_k при заданих $x_1(1), ..., x_{k-1}(1), x_{k+1}(0), ..., x_K(0)$;

…

$x_K(1)$ із розподілу випадкової величини X_K при заданих $x_1(1), x_3(1), ..., x_{K-1}(1)$;

На усіх інших ітераціях діємо аналогічно. Слід підкреслити такі властивості цієї схеми.

1. Кожний компонент випадкового вектору X обходиться природним способом, і в результаті на кожній ітерації генеруються K нових значень.
2. Нове значення компоненту X_{k-1} одразу використовується при генеруванні нового значення $X_k, k = 2, 3, ..., K.$

Після n ітерацій ми отримаємо K випадкових величин $X_1(n), X_2(n), ..., X_K(n)$. Для алгоритму Гіббса виконуються три теореми.

Теорема 1 (про збіжність). Випадкова величина $X_k(n)$ збігається по розподілу до істинного розподілу ймовірності X_k для $k = 1, 2, ..., K$ при $n \to \infty$, тобто

$$\lim_{n\to\infty} P\left(X_k^{(n)} \le x \middle| x_k(0) \right) = P_{X_k}(x) \text{ для } k = 1, 2, ..., K,$$

де P_{X_k} — маргінальна функція розподілу випадкової величини X_k.

Теорема 2 (про швидкість збіжності). Сумісний розподіл випадкових величин $X_1(n), X_2(n), ..., X_K(n)$ збігається істинного сумісного розподілу випадкових величин $X_1, X_2, ..., X_K$ із швидкістю геометричної прогресії.

Теорема 3 (про ергодичність). Для будь-якої вимірної функції g від випадкових величин $X_1, X_2, ..., X_K$, у якої існує математичне сподівання,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} g\left(X_1(i), X_2(i), ..., X_K(i) \right) \to E\left[g\left(X_1, X_2, ..., X_K \right) \right]$$

із ймовірністю 1, тобто

$$P\left(\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} g\left(X_1(i), X_2(i), ..., X_K(i) \right) = E\left[g\left(X_1, X_2, ..., X_K \right) \right] \right) = 1.$$

Приклад

Нехай є спільний розподіл $p(x_1, x_2, x_3)$ трьох випадкових величин, кожна з який знаходиться в діапазоні від 0 до 10. Приймемо, що первісне значення вектора, від якого почнеться ітераційний процес, буде $X = \{5, 2, 7\}$. Далі фіксуємо x_2 і x_3, після чого розраховуємо по відомій заздалегідь формулі умовну імовірність $p(x_1 | x_2, x_3)$, тобто $p(x_1 | x_2 = 2, x_3 = 7)$, одержуючи деякий графік щільності імовірності від змінної x_1. Те, що спочатку x_1 ми поклали рівним 5, забуваємо, більше це значення не знадобиться.

Тепер необхідно виконати семплування — згенерувати нове випадкове значення для x_1 відповідно до отриманої щільності імовірності. Семплірування можна зробити, наприклад, по алгоритму вибору з відхиленням (sampling with recject). Для цього генерується випадкове число з рівномірним розподілом від 0 до 10, після чого для цього згенерованого числа обчислюється його імовірність за графіком щільності імовірності $p(x_1 | x_2 = 2, x_3 = 7)$.

Наприклад, нехай згенерувалося випадкове число 4 і за графіком щільності його імовірність дорівнює 0.2. Тоді, відповідно до алгоритму вибору з відхиленням, ми приймаємо це згенероване число з імовірністю 0,2. А для цього, у свою чергу, генеруємо ще одне випадкове число від 0 до 1 з рівномірним розподілом, і, якщо згенерувалося число менше 0.2, то ми приймаємо число 4 як успішне. Інакше повторюємо спочатку — генеруємо ще одне число (наприклад випадає 3), для нього знаходимо імовірність (наприклад, 0.3), для нього генеруємо ще число від 0 до 1 (наприклад, 0,1) і тоді вже приймаємо остаточно, що на цій ітерації $x_1 = 3$.

Далі необхідно повторити всі дії вище з величиною x_2, причому x_1 ми уже використовуємо “нове” — у нашому прикладі рівне 3. Так, розраховуємо щільність імовірності $p(x_2 | x_1 = 3, x_3 = 7)$ генеруємо знову випадкове число на роль кандидата нового значення x_2, робимо вибірку з відхиленням і повторюємо її у випадку, якщо значення “відхилене”.

Аналогічно дії повторюються для x_3 з новими значеннями x_1 і x_2. Перша ітерація алгоритму семплування по Гиббсу довершена. Через кілька сотень/тисяч таких ітерацій випадкові значення повинні прийти до максимуму своєї щільності, що може бути розташований досить далеко від нашого першого наближення $X = \{5, 2, 7\}$ і семплуватися в тій області. Подальша тисяча ітерацій може уже використовуватися по призначенню (для пошуку математичного сподівання, наприклад) як зразок значень шуканого розподілу, що не залежать від первісного вектора $X = \{5, 2, 7\}$.

Метод вибору за Гіббсом належить до сімейства методів Монте-Карло із марковськими ланцюгами (Markov Chain Vonte Carlo), тобто методів для моделювання невідомого розподілу ймовірностей, які генерують ергодичний ланцюг Маркова, стаціонарним розподілом якого є невідомий розподіл, що моделюється.

Список використаних джерел

1. Haykin S. Neural Networks and Learning Machines. — Pearson Education, 2009.
2. https://ru.wikipedia.org/wiki/Семплирование_по_Гиббсу.
3. Николенко С.И. Вероятностные модели: сэмплирование. https://habr.com/company/surfingbird/blog/226677
4. Николенко С.И. Методы Монте-Карло: сэмплинг. https://logic.pdmi.ras.ru/~sergey/teaching/mlbayes/05-montecarlo.pdf