Лекция 11. Способы вывода в графических моделях. Методы Монте-Карло с марковскими цепями. Гауссовские процессы в машинном обучении.

Д. П. Ветров[1] Д. А. Кропотов[2]

[1]МГУ, ВМиК, каф. ММП

[2]ВЦ РАН

Спецкурс «Байесовские методы машинного обучения»

План лекции

План лекции

Случайные процессы

- Случайным процессом будем называть индексированное множество случайных величин $\xi(\omega) = \{\xi_t(\omega) | t \in T\}$
- Иногда используется нотация $\xi(\omega, t)$
- Первоначально $T \subset \mathbb{R}$, а переменная t ассоциировалась со временем

  Случайный процесс в этом случае удобно представлять как некоторую случайную величину, меняющуюся во времени
- Если $T \subset \mathbb{R}^d$, то случайный процесс обычно называют случайным полем

  Случайный процесс в этом случае удобно представлять как некоторую случайную величину, меняющуюся в пространстве

Двойственная природа случайного процесса

- При фиксированном времени $t = t_0$ процесс представляет собой обычную случайную величину

$$X(\omega) = \xi(\omega, t_0)$$

- При фиксированном элементарном событии $\omega = \omega_0$ процесс представляет собой функцию, называемую реализацией случайного процесса

$$f(t) = \xi(\omega_0, t)$$

- Таким образом, случайный процесс обладает как вероятностными, так и функциональными характеристиками
- В частности, можно говорить о математическом ожидании, дисперсии процесса в фиксированный момент времени, а также рассматривать производные и интегралы от реализаций процесса

Вероятностные характеристики случайного процесса

- Среднее значение процесса

$$m(t) = \mathbb{E}\xi(\omega, t)$$

- Ковариационная функция процесса

$$C(t_1, t_2) = \mathrm{Cov}(\xi(\omega, t_1), \xi(\omega, t_2)),$$

обладающая следующими свойствами

$$C(t, t) = \mathbb{D}\xi(\omega, t) \geq 0, \quad |C(t_1, t_2)| \leq \sqrt{C(t_1, t_1)C(t_2, t_2)}$$

- Процесс называется стационарным, если его вероятностные характеристики не зависят от времени, в частности

$$C(t, t+\tau) = C(0, \tau) = C(\tau), \quad \forall t$$

Большинство теорем в теории случайных процессов доказано для стационарных процессов

План лекции

Идея метода Монте-Карло

- Метод Монте-Карло применяется для решения задач численного моделирования, в частности взятия интегралов

$$\int_a^b f(x)dx \approx \frac{b-a}{n}\sum_{i=1}^n f(x_i) = \hat{f}, \quad x_i \sim U[a,b]$$

- Можно показать, что при весьма общих предположениях $\hat{f} \to \int_a^b f(x)dx$ при $n \to \infty$
- Точность оценки интегралов не зависит от размерности пространства d, а определяется исключительно дисперсией самой функции

$$\mathbb{D}\hat{f} = \frac{1}{n}\left[(b-a)\int f^2(x)dx - \left(\int f(x)dx\right)^2\right]$$

- Для численной оценки вероятностных интегралов необходимы специальные методы

Вероятностные интегралы

- В дальнейшем будем рассматривать интегралы вида

$$\mathbb{E}f = \int f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}$$

- К ним сводятся многие интегралы, возникающие при байесовском обучении, в частности обоснованность

$$Evidence = \mathbb{E}_{\boldsymbol{w}}p(\boldsymbol{t}|\boldsymbol{w}) = \int p(\boldsymbol{t}|\boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w}$$

и голосование по апостериорному распределению

$$p(t_{new}|\boldsymbol{t}) = \mathbb{E}_{\boldsymbol{w}}p(t_{new}|\boldsymbol{w}) = \int p(t_{new}|\boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{t})d\boldsymbol{w}$$

Особенности вероятностных интегралов

- Классическая выборка из равномерного распределения для взятия таких интегралов, т.е. формула

$$\int_D f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \approx \frac{|D|}{n}\sum f(\boldsymbol{x}_i)p(\boldsymbol{x}_i), \quad \boldsymbol{x} \sim U(D),$$

  крайне неэффективна, так как в большей части области интегрирования плотность, а, следовательно, и подынтегральная функция близка к нулю
- Для взятия интегралов вида $\int f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}$ нужно уметь проводить выборку из распределения $p(\boldsymbol{x})$
- В этом случае интеграл может быть оценен конечной суммой

$$\int f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \approx \frac{1}{n}\sum f(\boldsymbol{x}_i), \quad \boldsymbol{x} \sim p(\boldsymbol{x})$$

Метод обратной функции

- В некоторых случаях можно свести задачу генерации выборки из некоторого распределения к генерации выборки из равномерного распределения
- Пусть $F(x) = P(X < x) = \int_{-\infty}^{x} p(\xi)d\xi$ — функция распределения случайной величины X
- Легко показать (Упр.), что $Y = F(X) \sim U(0, 1)$, тогда $X \sim F^{-1}(U(0, 1))$
- Так удается сгенерировать выборку из показательного распределения и распределения Коши (Упр.)

План лекции

Схема с весами

- В дальнейшем полагаем, что нам в каждой точке известна плотность распределения величины с точностью до множителя, т.е.

$$p(\boldsymbol{x}) = \frac{1}{Z_p}\tilde{p}(\boldsymbol{x}),$$

причем Z_p неизвестна, а $\tilde{p}(\boldsymbol{x})$ может быть легко подсчитана в любой точке
- Введем распределение $q(\boldsymbol{x})$, из которого легко сгенерировать выборку, тогда

$$\mathbb{E}_p f = \int f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} = \frac{1}{Z_p}\int f(\boldsymbol{x})\frac{\tilde{p}(\boldsymbol{x})}{q(\boldsymbol{x})}q(\boldsymbol{x})d\boldsymbol{x} \approx$$

$$\frac{1}{nZ_p}\sum_{i=1}^{n} f(\boldsymbol{x}_i)\frac{\tilde{p}(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)} = \frac{1}{n\sum_{i=1}^{n} r_i}\sum_{i=1}^{n} f(\boldsymbol{x}_i)r_i, \qquad \boldsymbol{x} \sim q(\boldsymbol{x})$$

- Если распределение $q(\boldsymbol{x})$ сильно отличается от $p(\boldsymbol{x})$, большинство весов r_i близки к нулю, и метод становится неустойчивым

Марковская цепь

- Методы Монте Карло, использующие Марковские цепи (Monte Carlo Markov chain, MCMC) являются более эффективными средствами получения выборки из заданного распределения
- При использовании MCMC каждая очередная точка выборки $\boldsymbol{x}_i$ зависит некоторым образом от предыдущей точки $\boldsymbol{x}_{i-1}$
- Методы этой группы позволяют «нащупать» области с высоким значением плотности и проводить выборку из них
- Полученная выборка $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ не является выборкой независимых одинаково распределенных случайных величин, но вполне подходит для взятия интеграла

Схема Метрополиса-Хастингса

Последовательная схема генерации точек.

Вход: Многомерное распределение $p(\boldsymbol{x})$, известное с точностью до нормировочной константы, т.е. $p(\boldsymbol{x}) = \frac{1}{Z}\tilde{p}(\boldsymbol{x})$; предложное распределение $q(\boldsymbol{x}|\boldsymbol{y})$, из которого можно легко генерировать точки $\boldsymbol{x}$.

Выход: Выборка из распределения $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$

Инициализация $\boldsymbol{x}_0 = \boldsymbol{\mu}$;

для $i = 1, \ldots, n$

Сгенерировать $\boldsymbol{x}_i$ из распределения $q(\boldsymbol{x}_i|\boldsymbol{x}_{i-1})$;

Принять точку $\boldsymbol{x}_i$ с вероятностью

$$A_i(\boldsymbol{x}_i, \boldsymbol{x}_{i-1}) = min\left(1, \frac{\tilde{p}(\boldsymbol{x}_i)q(\boldsymbol{x}_{i-1}|\boldsymbol{x}_i)}{\tilde{p}(\boldsymbol{x}_{i-1})q(\boldsymbol{x}_i|\boldsymbol{x}_{i-1})}\right)$$

Схема Метрополиса

Если распределение q симметричное, т.е. $q(\boldsymbol{x}|\boldsymbol{y}) = q(\boldsymbol{y}|\boldsymbol{x})$, то схема Метрополиса-Хастингса переходит в схему Метрополиса, в которой очередная точка принимается с вероятностью

$$A_i(\boldsymbol{x}_i, \boldsymbol{x}_{i-1}) = min\left(1, \frac{\tilde{p}(\boldsymbol{x}_i)}{\tilde{p}(\boldsymbol{x}_{i-1})}\right)$$

Таким образом, если $\tilde{p}(\boldsymbol{x}_i) \geq \tilde{p}(\boldsymbol{x}_{i-1})$, то очередная точка с гарантией принимается. В противном случае точка принимается с некоторой вероятностью.

Пример использования схемы Метрополиса

Нужно сгенерировать выборку из многомерного нормального распределения, при этом предложное распределение является нормальным с матрицей ковариации, пропорциональной единичной.

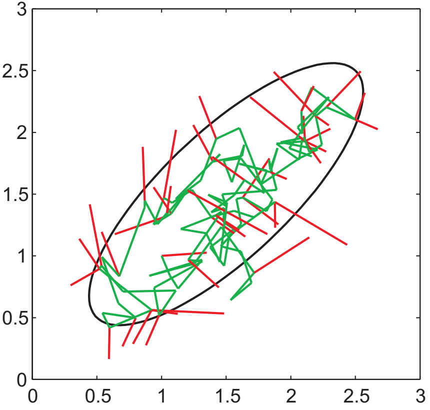


Схема Гиббса

Вход: Многомерное распределение $p(\boldsymbol{x})$;

Выход: Выборка из распределения $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$

1: Инициализация $\boldsymbol{x}_0 = (x_1^0, \dots, x_d^0)$;

2: для $i = 1, \dots, n$

3: Сгенерировать x_1^i из распределения $p(x_1 | x_2^{i-1}, x_3^{i-1}, \dots, x_d^{i-1})$;

4: Сгенерировать x_2^i из распределения $p(x_2 | x_1^i, x_3^{i-1}, \dots, x_d^{i-1})$;

...

5: Сгенерировать x_d^i из распределения $p(x_d | x_2^i, x_3^i, \dots, x_{d-1}^i)$;

6: $\boldsymbol{x}_i := (x_1^i, \dots, x_d^i)$;

План лекции

- Гибридные методы используют информацию не только о значении плотности $p(\boldsymbol{x})$, но и о градиенте ее логарифма $\frac{\partial \log p(\boldsymbol{x})}{\partial \boldsymbol{x}}$
- Для этого используются аналогии с аналитической механикой

  Аналитическая механика была разработана в первой половине 19 в. ирландским математиком Гамильтоном. В ее основе лежит идея замены одного дифференциального уравнения второго порядка во втором законе Ньютона на систему двух дифференциальных уравнений первого порядка
- Считая $\boldsymbol{x}$ переменными состояния, введем потенциальную энергию системы

$$E(\boldsymbol{x}) = -\log p(\boldsymbol{x}) + C$$

- Здесь используется принцип минимальной потенциальной энергии, гласящий, что состояние системы тем более вероятно, чем меньше ее потенциальная энергия

Аналитическая механика

- Введем дополнительные переменные, называемые моментами
$$\boldsymbol{r} = \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t}$$
- Кинетическая энергия системы является функцией моментов $K(\boldsymbol{r}) = 0.5\|\boldsymbol{r}\|^2$, а полная энергия системы (гамильтониан) равна
$$H(\boldsymbol{x}, \boldsymbol{r}) = E(\boldsymbol{x}) + K(\boldsymbol{r})$$
- Уравнения Гамильтона являются записью второго закона Ньютона через переменные состояния и моменты
$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \frac{\partial H}{\partial \boldsymbol{r}}$$
$$\frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}t} = -\frac{\partial H}{\partial \boldsymbol{x}}$$

Интегрирование уравнений Гамильтона

- При динамическом изменении замкнутой системы гамильтониан H является постоянным по времени (закон сохранения энергии)
- Изменение системы описывается функциями $\boldsymbol{x}(t)$ и $\boldsymbol{r}(t)$, связанными уравнениями Гамильтона
- При численном решении уравнений получаем

$$\boldsymbol{r}(t+\varepsilon/2) = \boldsymbol{r}(t) - \frac{\varepsilon}{2}\frac{\partial E}{\partial \boldsymbol{x}}(\boldsymbol{x}(t))$$

$$\boldsymbol{x}(t+\varepsilon) = \boldsymbol{x}(t) + \varepsilon \boldsymbol{r}(t+\varepsilon/2)$$

$$\boldsymbol{r}(t+\varepsilon) = \boldsymbol{r}(t+\varepsilon/2) - \frac{\varepsilon}{2}\frac{\partial E}{\partial \boldsymbol{x}}(\boldsymbol{x}(t+\varepsilon))$$

- Полученные решения приблизительно описывают одну из линий уровня функции Гамильтона

Графическая иллюстрация эволюции системы

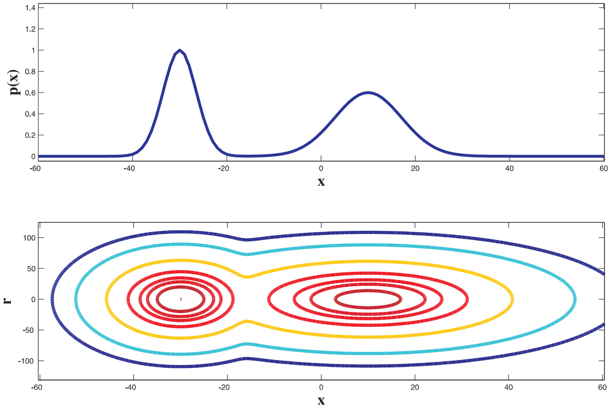


Схема генерации выборки

- Точки $(\boldsymbol{x}(t_1), \dots, \boldsymbol{x}(t_n))$ представляют собой равномерную выборку из множества $\{\boldsymbol{x} | p(\boldsymbol{x}) \geq C_0\}$
- Чтобы получить выборку из распределения $p(\boldsymbol{x})$ через каждые $m \ll n$ итераций значение моментов берется из распределения $p(\boldsymbol{r}) = \frac{1}{Z_r} \exp(-K(\boldsymbol{r})) = \mathcal{N}(\boldsymbol{r}|0, I)$
- Такая схема генерации выборки позволяет быстро найти области с большим значением $p(\boldsymbol{x})$ и получить репрезентативную выборку из этих областей

План лекции

Гауссовские процессы

- Гауссовским процессом называется случайный процесс, все конечномерные распределения которого нормальные

$$p(\xi(\omega, x_1), \dots, \xi(\omega, x_n)) = \mathcal{N}(\boldsymbol{\xi}|\boldsymbol{\mu}, \Sigma)$$

В дальнейшем символ ω будем опускать

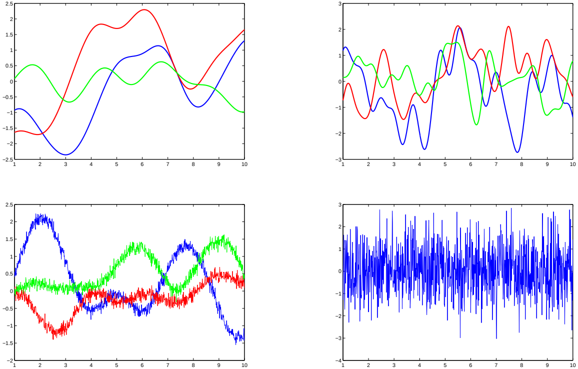
- Гауссовский процесс является обобщением многомерной гауссианы и полностью задается функцией среднего значения и ковариационной функцией
- Далее будем рассматривать стационарные гауссовские поля $\xi(\boldsymbol{x})$

$$\mu(t) = m, \qquad C(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{y}) = C(\boldsymbol{y})$$

Если дополнительно известно, что ковариационная функция зависит только от нормы разности $C(\boldsymbol{y}) = C(\|\boldsymbol{y}\|)$, то процесс называют изотропным

Примеры гауссовских процессов

Гауссовские процессы (ГП) являются довольно гибким средством описания данных, а степень «гладкости» процесса определяется видом ковариационной функции



Использование случайных полей в задачах восстановления регрессии

- Рассмотрим задачу восстановления регрессии по обучающей выборке $(X, \boldsymbol{t})$, $t \in \mathbb{R}$
- Значения t_i можно интерпретировать как значения реализации случайного процесса (поля) в соответствующей точке $\boldsymbol{x}_i$
- Возникает задача прогноза значения поля t в новой точке $\boldsymbol{x}$ при условии, что в точках обучающей выборки поле имело значения $\boldsymbol{t}$

$$p(\xi(\boldsymbol{x})|\xi(\boldsymbol{x}_1) = t_1, \ldots, \xi(\boldsymbol{x}_n) = t_n) =?$$

Конечномерные распределения поля

- Заметим, что по определению гауссовского случайного процесса (поля)

$$p(\xi(\boldsymbol{x}_1), \dots, \xi(\boldsymbol{x}_n), \xi(\boldsymbol{x})) = \mathcal{N}((\boldsymbol{\xi}, \xi)|\boldsymbol{0}, \hat{C}),$$

где

$$\hat{C} = \begin{pmatrix} C & \boldsymbol{k} \\ \boldsymbol{k}^T & C(\boldsymbol{x}, \boldsymbol{x}) \end{pmatrix},$$

$$C = (C(\boldsymbol{x}_i, \boldsymbol{x}_j)), \quad \boldsymbol{k} = (C(\boldsymbol{x}_1, \boldsymbol{x}), \dots, C(\boldsymbol{x}_n, \boldsymbol{x}))^T$$

- Также по определению $p(\xi(\boldsymbol{x}_1), \dots, \xi(\boldsymbol{x}_n)) = \mathcal{N}(\boldsymbol{\xi}|\boldsymbol{0}, C)$

Формула Андерсона

- Учитывая, что

$$p(\xi(\boldsymbol{x})|\xi(\boldsymbol{x}_1),\ldots,\xi(\boldsymbol{x}_n)) = \frac{p(\xi(\boldsymbol{x}_1),\ldots,\xi(\boldsymbol{x}_n),\xi(\boldsymbol{x}))}{p(\xi(\boldsymbol{x}_1),\ldots,\xi(\boldsymbol{x}_n))},$$

легко показать (Упр.), что

$$p(\xi(\boldsymbol{x})|\xi(\boldsymbol{x}_1),\ldots,\xi(\boldsymbol{x}_n)) = \mathcal{N}(\xi|\mu,\sigma^2)$$

- Прогноз поля имеет нормальное распределение с параметрами

$$\mu = \boldsymbol{k}^T C^{-1}\boldsymbol{t}$$

$$\sigma^2 = C(\boldsymbol{x},\boldsymbol{x}) - \boldsymbol{k}^T C^{-1}\boldsymbol{k} = s^2 - \boldsymbol{k}^T C^{-1}\boldsymbol{k},$$

где $s^2 = \mathbb{D}\xi$ — дисперсия случайного поля

План лекции

Задача классификации

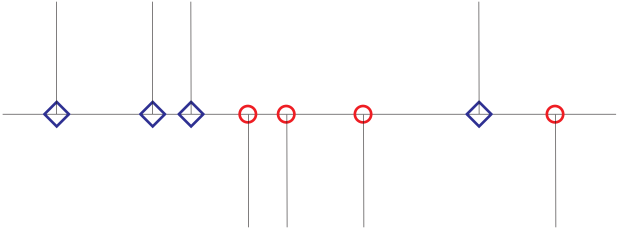
- В задаче классификации ситуация сложнее
- Значение реализации процесса в точках обучающей выборки неизвестно, да и интересует нас лишь знак прогноза, т.е.

$$p(\mathrm{sign}(\xi(\boldsymbol{x}))|\mathrm{sign}(\xi(\boldsymbol{x}_1)) = t_1, \ldots, \mathrm{sign}(\xi(\boldsymbol{x}_n)) = t_n) =?$$

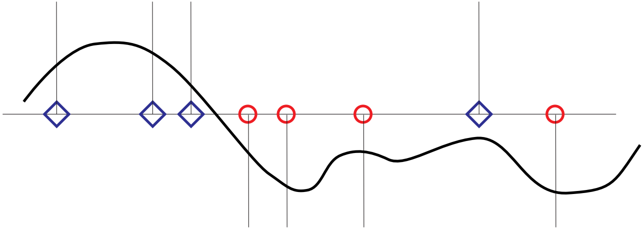
- Решение заключается в поиске наиболее вероятной реализации случайного процесса с учетом информации о знаках

Восстановление реализации ГП в задаче классификации

Восстановление реализации ГП в задаче классификации



Восстановление реализации ГП в задаче классификации



ГП классификатор

- Введем правдоподобие метки класса
$$p(t|\xi(\boldsymbol{x})) = \frac{1}{1+\exp(-t\xi(\boldsymbol{x}))}$$
- Тогда обозначив $\boldsymbol{\xi} = (\xi(\boldsymbol{x}_1), \dots, \xi(\boldsymbol{x}_n))$, получаем
$$p(\boldsymbol{\xi}|\boldsymbol{t}) \propto p(\boldsymbol{t}|\boldsymbol{\xi})p(\boldsymbol{\xi}) =$$
$$\prod_{i=1}^{n} \frac{1}{1+\exp(-t_i\xi(\boldsymbol{x}_i))} \frac{1}{\sqrt{(2\pi)^n \det(C)}} \exp\left(-\frac{1}{2}\boldsymbol{\xi}^T C^{-1}\boldsymbol{\xi}\right)$$
- Отсюда находим
$$\hat{\boldsymbol{\xi}} = \arg\max p(\boldsymbol{\xi}|\boldsymbol{t})$$
Для поиска $\hat{\boldsymbol{\xi}}$ можно воспользоваться методом IRLS (см. лекцию 3)
- Окончательный вид решающего правила для ГП классификатора
$$t_{new} = \mathrm{sign}(\boldsymbol{k}C^{-1}\hat{\boldsymbol{\xi}})$$

План лекции

Функционал качества для ковариационной функции

- В зависимости от вида ковариационной функции могут быть найдены различные реализации ГП
- !! Ковариационная функция является структурным параметром ГП!!
- Запишем правдоподобие ковариационной функции при данной реализации

$$p(\boldsymbol{\xi}|C(\boldsymbol{x},\boldsymbol{y})) = \frac{1}{\sqrt{(2\pi)^n \det(C)}} \exp\left(-\frac{1}{2}\boldsymbol{\xi}^T C^{-1}\boldsymbol{\xi}\right) \to \max_{C_{ij}=C(\boldsymbol{x}_i,\boldsymbol{x}_j)}$$

Заметим, что при этой оптимизации реализация $\boldsymbol{\xi}$ фиксирована

Обоснованность модели ГП

- Популярным параметрическим семейством ковариационных функций является

$$C_{A,\sigma,s}(\boldsymbol{x},\boldsymbol{y}) = A\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2s^2}\right) + \sigma^2 I_{\{\boldsymbol{x}=\boldsymbol{y}\}}$$

- При оптимизации $p(\boldsymbol{\xi}|C(\boldsymbol{x},\boldsymbol{y}))$ происходит поиск ковариационной функции, наиболее адекватной данной реализации
- Величина $p(\boldsymbol{\xi}|C(\boldsymbol{x},\boldsymbol{y}))$ является правдоподобием структурных параметров или обоснованностью модели ГП