Лекция 1. Краткое напоминание математических понятий, которые понадобятся в следующих лекциях

А. С. Конушин[1] Д. П. Ветров[2] Д. А. Кропотов[3]
В. С. Конушин[1] О. В. Баринова[1]

[1]МГУ, ВМиК, лаб. КГ

[2]МГУ, ВМиК, каф. ММП

[3]ВЦ РАН

Спецкурс «Структурные методы анализа изображений и сигналов»

План

Задачи курса

Спецкурс направлен на

- бла-бла...
- еще бла-бла...
- и еще немножечко бла.

Актуальность курса

Современные алгоритмы компьютерного зрения активно используют следующие методы

- бла-бла...
- еще бла-бла...
- и еще немножечко бла.

Программа курса

- Спецкурс состоит из ... лекций
- В рамках спецкурса будет предложено ... заданий для самостоятельной работы
- Условия получения положительной оценки/система проведения экзамена

Матричная нотация

- При работе с многомерными величинами очень удобна матричная нотация, т.е. представление многих операций над векторами и числами в виде операций над матрицами
- Скалярное произведение двух векторов $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ принимает вид

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{i=1}^{d} x_i y_i = \boldsymbol{x}^T \boldsymbol{y},$$

т.е. вектора трактуются как частные случаи матриц

- Квадратичная форма

$$\langle A\boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{i=1}^{d} \sum_{j=1}^{d} x_i a_{ij} y_j = \boldsymbol{x}^T A \boldsymbol{y}$$

- Матричная нотация облегчает математические выкладки и позволяет реализовать вычисления на ЭВМ более эффективно

Пример использования

- Предположим нам надо решить несовместную систему линейных уравнений $A\boldsymbol{x} \approx \boldsymbol{b}$, $A \in \mathbb{R}^{m \times n}$
- Для этого будем минимизировать квадрат нормы невязки (система-то нерешаемая) $\|A\boldsymbol{x} - \boldsymbol{b}\|^2 \to \min_{\boldsymbol{x}}$
- Представляя норму в матричной виде, дифференцируя по вектору и приавнивая производную к нулю получаем известную формулу для псевдорешения СЛАУ

$$\|A\boldsymbol{x} - \boldsymbol{b}\|^2 = \langle A\boldsymbol{x} - \boldsymbol{b}, A\boldsymbol{x} - \boldsymbol{b}\rangle = (A\boldsymbol{x} - \boldsymbol{b})^T (A\boldsymbol{x} - \boldsymbol{b}) =$$

$$(A\boldsymbol{x})^T A\boldsymbol{x} - \boldsymbol{b}^T A\boldsymbol{x} - (A\boldsymbol{x})^T \boldsymbol{b} + \boldsymbol{b}^T \boldsymbol{b} =$$

$$\boldsymbol{x}^T A^T A\boldsymbol{x} - 2\boldsymbol{x}^T A^T \boldsymbol{b} + \boldsymbol{b}^T \boldsymbol{b}$$

$$\frac{\partial}{\partial \boldsymbol{x}} \left(\boldsymbol{x}^T A^T A\boldsymbol{x} - 2\boldsymbol{x}^T A^T \boldsymbol{b} + \boldsymbol{b}^T \boldsymbol{b}\right) = 2A^T A\boldsymbol{x} - 2A^T \boldsymbol{b} = 0$$

$$\boldsymbol{x} = \left(A^T A\right)^{-1} A^T \boldsymbol{b}$$

- Заметим, что если матрица A квадратная (число уравнений равно числу неизвестных) и невырожденная, то последняя формула переходит в формулу обычного решения СЛАУ $\boldsymbol{x} = A^{-1}\boldsymbol{b}$

Задача условной оптимизации

Пусть $f(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}$ — гладкая функция. Предположим, что нам необходимо найти ее экстремум:

$$f(\boldsymbol{x}) \to \operatorname*{extr}_{\boldsymbol{x}}$$

Для того, чтобы найти экстремум (решить задачу безусловной оптимизации), достаточно проверить условие стационарности:

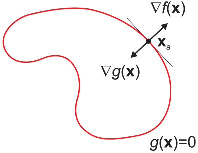
$$\nabla f(\boldsymbol{x}) = 0$$

Предположим, что нам необходимо найти экстремум функции при ограничениях:

$$f(\boldsymbol{x}) \to \operatorname*{extr}_{\boldsymbol{x}}$$

$$g(\boldsymbol{x}) = 0$$

Поверхность ограничения



Заметим, что $\nabla g(\boldsymbol{x})$ ортогонален поверхности ограничения $g(\boldsymbol{x}) = 0$. Пусть $\boldsymbol{x}$ и $\boldsymbol{x} + \boldsymbol{\varepsilon}$ — две близкие точки поверхности. Тогда

$$g(\boldsymbol{x} + \boldsymbol{\varepsilon}) \simeq g(\boldsymbol{x}) + \boldsymbol{\varepsilon}^T \nabla g(\boldsymbol{x})$$

Т.к. $g(\boldsymbol{x} + \boldsymbol{\varepsilon}) = g(\boldsymbol{x})$, то $\boldsymbol{\varepsilon}^T \nabla g(\boldsymbol{x}) \simeq 0$. При стремлении $\|\boldsymbol{\varepsilon}\| \to 0$ получаем $\boldsymbol{\varepsilon}^T \nabla g(\boldsymbol{x}) = 0$. Т.к. $\boldsymbol{\varepsilon}$ параллелен поверхности $g(\boldsymbol{x}) = 0$, то $\nabla g(\boldsymbol{x})$ является нормалью к этой поверхности.

Функция Лагранжа

Необходимым условием оптимальности является ортогональность $\nabla f(\boldsymbol{x})$ поверхности ограничения, т.е.:

$$\nabla f + \lambda \nabla g = 0 \tag{1}$$

Здесь $\lambda \neq 0$ — коэффициент Лагранжа. Он может быть любого знака.

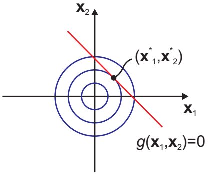
Функция Лагранжа

$$L(\boldsymbol{x}, \lambda) \triangleq f(\boldsymbol{x}) + \lambda g(\boldsymbol{x})$$

Тогда

$$\begin{aligned} \nabla_{\boldsymbol{x}} L = 0 \quad &\Rightarrow \quad \text{условие (1)} \\ \frac{\partial}{\partial \lambda} L = 0 \quad &\Rightarrow \quad g(\boldsymbol{x}) = 0 \end{aligned}$$

Функция Лагранжа. Пример.



$$f(x_1, x_2) = 1 - x_1^2 - x_2^2 \to \max_{x_1, x_2}$$

$$g(x_1, x_2) = x_1 + x_2 - 1 = 0$$

Функция Лагранжа:

$$L(\boldsymbol{x}, \lambda) = 1 - x_1^2 - x_2^2 + \lambda(x_1 + x_2 - 1)$$

Условия стационарности:

$$-2x_1 + \lambda = 0$$

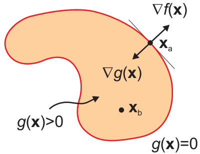
$$-2x_2 + \lambda = 0$$

$$x_1 + x_2 - 1 = 0$$

Решение: $(x_1^*, x_2^*) = (\frac{1}{2}, \frac{1}{2})$, $\lambda = 1$.

Ограничение в виде неравенства

$$f(\boldsymbol{x}) \to \max_{\boldsymbol{x}}$$
$$g(\boldsymbol{x}) \geq 0$$



Решение	Ограничение	Условие стационарности
Внутри области $g(\boldsymbol{x}) > 0$	неактивно	$\nabla f(\boldsymbol{x}) = 0$, $\nabla_{\boldsymbol{x}} L = 0$, $\lambda = 0$
На границе $g(\boldsymbol{x}) = 0$	активно	$\nabla f(\boldsymbol{x}) = -\lambda \nabla g(\boldsymbol{x})$, $\nabla_{\boldsymbol{x},\lambda} L = 0$, $\lambda > 0$

Условие дополняющей нежесткости:

$$\lambda g(\boldsymbol{x}) = 0$$

Теорема Каруша-Куна-Таккера

Пусть $f_i : X \to \mathbb{R}, \;\; i = 0, 1, \ldots, m$ — выпуклые функции, отображающие нормированное пространство X в прямую, $A \in X$ — выпуклое множество. Рассмотрим следующую задачу оптимизации:

$$f_0(\boldsymbol{x}) \to \min; \quad f_i(\boldsymbol{x}) \le 0, \;\; i = 1, \ldots, m, \;\; \boldsymbol{x} \in A \tag{2}$$

Теорема

1. Если $\hat{\boldsymbol{x}} \in \operatorname{absmin}(2)$ — решение задачи, то найдется ненулевой вектор множителей Лагранжа $\boldsymbol{\lambda} \in \mathbb{R}^{m+1}$ такой, что для функции Лагранжа $L(\boldsymbol{x}) = \sum_{i=0}^{m} \lambda_i f_i(\boldsymbol{x})$ выполняются условия:
   a) стационарности $\min_{\boldsymbol{x} \in A} L(\boldsymbol{x}) = L(\hat{\boldsymbol{x}})$
   b) дополняющей нежесткости $\lambda_i f_i(\hat{\boldsymbol{x}}) = 0, \;\; i = 1, \ldots, m$
   c) неотрицательности $\lambda_i \ge 0$
2. Если для допустимой точки $\hat{\boldsymbol{x}}$ выполняются условия a)–c) и $\lambda_0 \ne 0$, то $\hat{\boldsymbol{x}} \in \operatorname{absmin}(2)$
3. Если для допустимой точки $\hat{\boldsymbol{x}}$ выполняются условия a)–c) и $\exists \tilde{\boldsymbol{x}} \in A : f_i(\tilde{\boldsymbol{x}}) < 0, i = 0, \ldots, m$ (условие Слейтера), то $\hat{\boldsymbol{x}} \in \operatorname{absmin}(2)$

План

Краткое напоминание основных вероятностных понятий

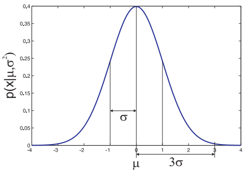
- $X : \Omega \to \mathbb{R}$ – случайная величина
- Вероятность попадания величины в интервал (a, b) равна
$$P(a \le X \le b) = \int_a^b p(x)dx,$$
где $p(x)$ – плотность распределения X,
$$p(x) \ge 0, \quad \int_{-\infty}^{\infty} p(x)dx = 1$$
- Если поведение случайной величины определяется некоторым параметром, возникают условные плотности $p(x|\theta)$. Если рассматривать условную плотность как функцию от параметра
$$f(\theta) = p(x|\theta),$$
то принято говорить о т.н. функции правдоподобия

Нормальное распределение

- Нормальное распределение играет важнейшую роль в математической статистике

$$X \sim \mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$\mu = \mathbb{E}X, \quad \sigma^2 = \mathbb{D}X \triangleq \mathbb{E}(X-\mathbb{E}X)^2$$



- Из центральной предельной теоремы следует, что сумма независимых случайных величин с ограниченной дисперсией стремится к нормальному распределению
- На практике многие случайные величины можно считать приближенно нормальными

Многомерное нормальное распределение

- Многомерное нормальное распределение имеет вид

$$X \sim \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{2\pi}^n \sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right),$$

  где $\boldsymbol{\mu} = \mathbb{E}X$, $\Sigma = \mathbb{E}(X-\boldsymbol{\mu})(X-\boldsymbol{\mu})^T$ — вектор математических ожиданий каждой из n компонент и матрица ковариаций соответственно
- Матрица ковариаций показывает, насколько сильно связаны (коррелируют) компоненты многомерного нормального распределения

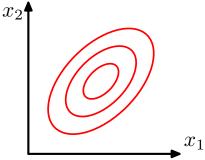
$$\Sigma_{ij} = \mathbb{E}(X_i - \mu_i)(X_j - \mu_j) = \mathrm{Cov}(X_i, X_j)$$

- Если мы поделим ковариацию на корень из произведений дисперсий, то получим коэффициент корреляции

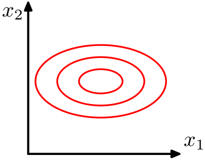
$$\rho(X_i, X_j) \triangleq \frac{\mathrm{Cov}(X_i, X_j)}{\sqrt{\mathbb{D}X_i \mathbb{D}X_j}} \in [-1, 1]$$

Особенности нормального распределения

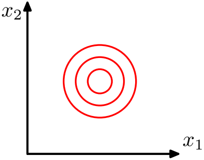
- Нормальное распределение полностью задается первыми двумя моментами (мат. ожидание и матрица ковариаций/дисперсия)
- Матрица ковариаций неотрицательно определена, причем на диагоналях стоят дисперсии соответствующих компонент
- Нормальное распределение имеет очень легкие хвосты: большие отклонения от мат. ожидания практически невозможны. Это обстоятельство нужно учитывать при приближении произвольных случайных величин нормальными



(a)



(b)



(c)

Основная задача мат. статистики

- Распределение случайной величины X известно с точностью до параметра θ
- Имеется выборка значений величины X, $\boldsymbol{x} = (x_1, \dots, x_n)$
- Требуется оценить значение θ
- Метод максимального правдоподобия

$$\hat{\theta}_{ML} = \arg\max f(\theta) = \arg\max p(\boldsymbol{x}|\theta) = \arg\max \prod_{i=1}^{n} p(x_i|\theta)$$

- Можно показать, что ММП (не путать с одноименной кафедрой) является асимптотически оптимальным при $n \to \infty$
- Обычно максимизируют не само правдоподобие, а его логарифм, т.к. это вычислительно проще (произведение плотностей по всем объектам переходит в сумму логарифмов плотностей)

Пример использования

- Пусть имееется выборка из нормального распределения $\mathcal{N}(x|\mu,\sigma^2)$ с неизвестными мат. ожиданием и дисперсией
- Выписываем логарифм функции правдоподобия

$$L(X|\mu,\sigma) = -\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2} - n\log\sigma - \frac{n}{2}\log(2\pi) \to \max_{\mu,\sigma}$$

$$\frac{\partial L}{\partial \mu} = -\sum_{i=1}^{n}\frac{(x_i-\mu)}{\sigma^2} = 0 \qquad \frac{\partial L}{\partial \sigma} = \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^3} - \frac{n}{\sigma} = 0$$

$$\mu_{ML} = \frac{1}{n}\sum_{i=1}^{n}x_i \qquad \sigma^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{j=1}^{n}x_j\right)^2$$

План

Условная вероятность

- Пусть X и Y — случайные величины с плотностями $p(x)$ и $p(y)$ соответственно
- В общем случае их совместная плотность $p(x,y) \neq p(x)p(y)$. Если это равенство выполняется, величины называют независимыми
- Условной плотностью называется величина

$$p(x|y) = \frac{p(x,y)}{p(y)}$$

- Смысл: как факт $Y = y$ влияет на распределение X. Заметим, что $\int p(x|y)dx \equiv 1$, но $\int p(x|y)dy$ не обязан равняться единице, т.к. относительно y это не плотность, а функция правдоподобия
- Очевидная система тождеств $p(x|y)p(y) = p(x,y) = p(y|x)p(x)$ позволяет легко переходить от $p(x|y)$ к $p(y|x)$

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

Правило суммирования вероятностей

- Все операции над вероятностями базируются на применении всего двух правил
- Правило суммирования: Пусть $A_1, \ldots, A_k$ взаимоисключающие события, одно из которых всегда происходит. Тогда

$$P(A_i \cup A_j) = P(A_i) + P(A_j) \qquad \sum_{i=1}^{k} P(A_i) = 1$$

- Очевидное следствие (формула полной вероятности): $\forall B$ верно $\sum_{i=1}^{k} P(A_i|B) = 1$, откуда

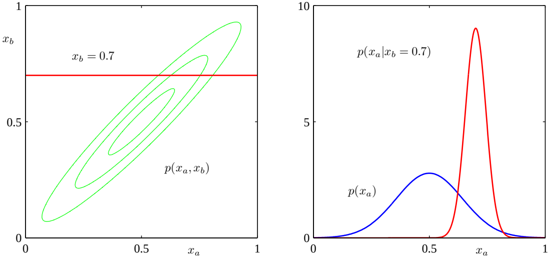
$$\sum_{i=1}^{k} \frac{P(B|A_i)P(A_i)}{P(B)} = 1 \qquad P(B) = \sum_{i=1}^{k} P(B|A_i)P(A_i)$$

- В интегральной форме

$$p(b) = \int p(b,a)da = \int p(b|a)p(a)da$$

Условное и маргинальное распределения

- Плотность распределения интересующей нас компоненты x_a многомерной случайной величины (x_a, x_b) можно получить двумя способами в зависимости от имеющейся информации
- Если нам неизвестны значения остальных компонент, мы маргинализуем плотность по ним: $p(x_a) = \int p(x_a, x_b) dx_b$
- Если значения других компонент нам известны, то мы обуславливаем плотность по ним: $p(x_a|x_b) = \frac{p(x_a, x_b)}{p(x_b)}$



Правило произведения вероятностей

- Правило произведения гласит, что любую совместную плотность всегда можно разбить на множители

$$p(a,b) = p(a|b)p(b) \quad P(A,B) = P(A|B)P(B)$$

- Аналогично для многомерных совместных распределений

$$p(a_1,\ldots,a_n) = p(a_1|a_2,\ldots,a_n)p(a_2|a_3,\ldots,a_n)\ldots p(a_{n-1}|a_n)p(a_n)$$

- Можно показать (Jaynes, 1995), что правила суммирования и произведения вероятностей являются единственными возможными операциями, позволяющими рассматривать вероятности как промежуточную ступень между истиной и ложью

Априорные и апостериорные суждения

- Предположим, мы пытаемся изучить некоторое явление
- У нас имеются некоторые знания, полученные до (лат. a priori) наблюдений/эксперимента. Это может быть опыт прошлых наблюдений, какие-то модельные гипотезы, ожидания
- В процессе наблюдений эти знания подвергаются постепенному уточнению. После (лат. a posteriori) наблюдений/эксперимента у нас формируются новые знания о явлении
- Будем считать, что мы пытаемся оценить неизвестное значение величины θ посредством наблюдений некоторых ее косвенных характеристик $x|\theta$

Формула Байеса

- Знаменитая формула Байеса (1763 г.) устанавливает правила, по которым происходит преобразование знаний в процессе наблюдений
- Обозначим априорные знания о величине θ за $p(\theta)$
- В процессе наблюдений мы получаем серию значений $\boldsymbol{x} = (x_1, \dots, x_n)$. При разных θ наблюдение выборки $\boldsymbol{x}$ более или менее вероятно и определяется значением правдоподобия $p(\boldsymbol{x}|\theta)$
- За счет наблюдений наши представления о значении θ меняются согласно формуле Байеса

$$p(\theta|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\theta)p(\theta)}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|\theta)p(\theta)}{\int p(\boldsymbol{x}|\theta)p(\theta)d\theta}$$

- Заметим, что знаменатель не зависит от θ и нужен исключительно для нормировки апостериорной плотности

Точечные оценки при использовании метода Байеса

- Хотя строгое применение байесовского оценивания подразумевает работу со всем апостериорным распределением, на практике его часто заменяют на точечную оценку, взятую в точке максимума апостериорной плотности

$$\hat{\theta}_{MP} = \arg\max P(\theta|\boldsymbol{x}) = \arg\max P(\boldsymbol{x}|\theta)P(\theta) =$$

$$\arg\max \left(\log P(\boldsymbol{x}|\theta) + \log P(\theta)\right)$$

- Это фактически регуляризация метода максимального правдоподобия, обспечивающая компромисс между априорными ограничениями и наблюдениями