П.В. Голубцов

Теоретические основы аналитики больших данных

Задания для домашней работы

Москва
Физический факультет МГУ им. М.В. Ломоносова
2024

Голубцов Петр Викторович

Теоретические основы аналитики больших данных
Задания для домашней работы

Пособие предназначено для студентов 5 курса физического факультета МГУ — слушателей курса “Теоретические основы аналитики больших данных”. Данный курс предполагает выполнение домашних заданий, которые, по сути, представляют собой теоретические (исследовательские) или практические (программные) проекты. Пособие содержит условия и подробные рекомендации по выполнению основных этапов этих заданий. Материал соответствует курсу лекций по спецкурсу “Теоретические основы аналитики больших данных”.

Содержание

1 Каноническая Информация для Выборочного Среднего и Ковариационной Матрицы

Пусть $(x_1, x_2, \ldots, x_n)$ последовательность векторов-столбцов:

$$x_i = \begin{bmatrix} x_i^1 \\ \vdots \\ x_i^m \end{bmatrix}, \qquad i = 1, \ldots, n.$$

В статистике часто приходится вычислять вектор **выборочного среднего**

$$X = \frac{1}{n} \sum_{i=1}^{n} x_i$$

и **выборочную ковариационную** матрицу

$$V = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - X)(x_i - X)^T,$$

где x^T - транспонирование столбца x.

1. Какую **каноническую** форму информации, Вы бы предложили для представления последовательности $(x_1, x_2, \ldots, x_n)$, чтобы иметь возможность вычислить вектор выборочного среднего и выборочную ковариационную матрицу?

Убедитесь в том, что все "желательные" свойства канонической информации выполнены:

(a) **Существование и единственность**. Любая последовательность исходных данных

$(x_1, x_2, \ldots, x_n)$ может быть представлена в канонической форме единственным образом. Зависит ли это представление от порядка векторов x_i в последовательности данных?

(b) **Полнота.** Каноническая информация должна сохранять всю информацию, которая присутствовала в исходных необработанных данных. В частности, вычисления, использующие каноническую информацию должны давать те же результаты, что и оригинальный алгоритм примененный к исходным необработанным данным.

Удовлетворяет ли Ваше каноническое представление данных этому требованию? Как вычислить X и V используя только собранную канонической информацию?

(c) **Элементарная** каноническая информация. Существует ли канонический информация для одного наблюдения?

(d) **Пустая** каноническая информация. Существует ли канонический информация для пустой последовательности наблюдений?

(e) **Объединение** (комбинация, композиция). Как бы Вы определили объединение информации в канонической форме? Выполняются ли аксиомы коммутативного моноида? (Коммутативность, ассоциативность, нейтральный элемент)

(f) **Обновление.** Как обновляется каноническая информация, когда поступает новый вектор наблюдений x?

(g) Каково **минимальное число** наблюдений n для которого определены X, V, каноническая информация?

(h) **Компактность и эффективность.** Что Вы можете сказать о компактности (или минимальности) и эффективности вашей канонической формы информации с точки зрения требований к хранению и сложности объединения, обновления и использования?

2. **(Необязательно, по желанию)** Какую **явную** форму информации Вы могли бы предложить, чтобы представить последовательность векторов $(x_1, x_2, \ldots, x_n)$? Это представление должно содержать X и V и, возможно, что-то еще.

   Проанализируйте все вопросы задачи 1 для предложенной явной информации.

2 Модель Распределённых Вычислений MapReduce

Напишите программу, имитирующую модель распределённых вычислений MapReduce для построения выборочного среднего и выборочной ковариационной матрицы из домашнего задания №1.

А именно, пусть $(x_1, x_2, \ldots, x_n)$ последовательность векторов:

$$x_i = \begin{bmatrix} x_i^1 \\ \vdots \\ x_i^m \end{bmatrix}, \qquad i = 1, \ldots, n.$$

Напишите программу в стиле MapReduce, которая вычисляет вектор **выборочного среднего:**

$$X = \frac{1}{n} \sum_{i=1}^{n} x_i$$

и **выборочную ковариационную** матрицу:

$$V = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - X)(x_i - X)^T,$$

где x^T - транспонирование столбца x.

Ключевой частью является разработка трех функций:

1. Функция, которая извлекает каноническую информацию из набора данных. Эта функция будет применена к списку наборов данных на этапе Map.

2. Функция, которая объединяет две части канонической информации в одну. Она будет использоваться для объединения всех частей канонической информации в одну на этапе Reduce.

3. Функция, которая вычисляет результат из объединенной канонической информации.

Рекомендации:

- Используйте каноническое представление информации, разработанное в домашнем задании №1.
- В качестве примера можно использовать код “Mean_MR.py”, иллюстрирующий вычисление среднего арифметического для набора чисел.
- Для генерации набора случайных столбцов с заданными средним и ковариационной матрицей можно использовать “MV_ArGen.py”.

3 Линейная Регрессия

Напишите программу, которая иллюстрирует линейную регрессию и реализует накопление канонической информации.

a) Для некоторых фиксированных параметров $a_1, \ldots, a_m$ сгенерируйте последовательность "наблюдений" (x_i, y_i):

$$y_i = f(x_i) + \varepsilon_i,$$

где

$$f(x) = a_1 f_1(x) + a_2 f_2(x) + \cdots + a_m f_m(x),$$

а $f_1(x), \ldots, f_m(x)$ - Ваш любимый набор линейно независимых функций, например,

$$f(x) = a_1 + a_2 x + a_3 x^2 + \cdots + a_m x^{m-1},$$

ε_i независимы и одинаково распределены с нулевым средним и дисперсией $\mathsf{E}\varepsilon_i^2 = \sigma^2$. Значения x_i также могут быть сгенерированы случайным образом.

b) Осуществите накопление канонической информации, т.е. на каждом шаге, когда производится новое наблюдение (x_i, y_i), обновляйте каноническую информацию.

c) Проиллюстрируйте истинную функцию $f(x)$ и ее оценку $\widehat{f(x)}$.

d) Проиллюстрируйте $\mathrm{Var}(\widehat{f(x)})$, при условии, что σ^2 известно, например, обозначьте коридор в одно стандартное отклонение $(= \sqrt{\mathrm{Var}(\widehat{f(x)})})$ относительно $\widehat{f(x)}$.

e) Проиллюстрируйте $\widehat{\mathrm{Var}(\widehat{f(x)})}$, при условии, что σ^2 НЕ известно.

В своем отчете представьте исходный код и несколько (не менее 3) иллюстраций, показывающих оценки для “малого”, “среднего” и “большого” числа наблюдений.

Пример и формулы

$$y_i = f_a(x_i) + \varepsilon_i = a_1 f_1(x_i) + \cdots + a_m f_m(x_i) + \varepsilon_i = F_{x_i} a + \varepsilon_i$$

Например, $f_a(x)$ - полином: $y_i = 1 + 1 \cdot x_i - 1 \cdot x_i^2 + 0.2 \cdot x_i^3 + \varepsilon_i$

$$F_x = \begin{bmatrix} 1 & x & x^2 & x^3 \end{bmatrix}, \qquad m = 4, \qquad a = \begin{bmatrix} 1 \\ 1 \\ -1 \\ 0.2 \end{bmatrix}$$

Исходные данные: $(x_i, y_i), \qquad i = 1, \ldots, n$

Каноническая информация: (T, v, V, n)

Элементарная информация: (T_i, v_i, V_i, n_i)

$$n_i = 1, \qquad V_i = y_i^2, \qquad v_i = F_{x_i}^T \cdot y_i = \begin{bmatrix} f_1(x_i)\, y_i \\ \vdots \\ f_4(x_i)\, y_i \end{bmatrix},$$

$$T_i = F_{x_i}^T \cdot F_{x_i} = \begin{bmatrix} f_1(x_i)^2 & f_1(x_i)\, f_2(x_i) & \cdots & f_1(x_i)\, f_4(x_i) \\ \vdots & \vdots & \ddots & \vdots \\ f_4(x_i)\, f_1(x_i) & f_4(x_i)\, f_2(x_i) & \cdots & f_4(x_i)^2 \end{bmatrix}$$

Обновление:

$$(T, v, V, n) + (T_i, v_i, V_i, n_i) = (T + T_i, v + v_i, V + V_i, n + n_i)$$

Оценивание $f(x)$: $\qquad (T, v, V, n) * x \mapsto$

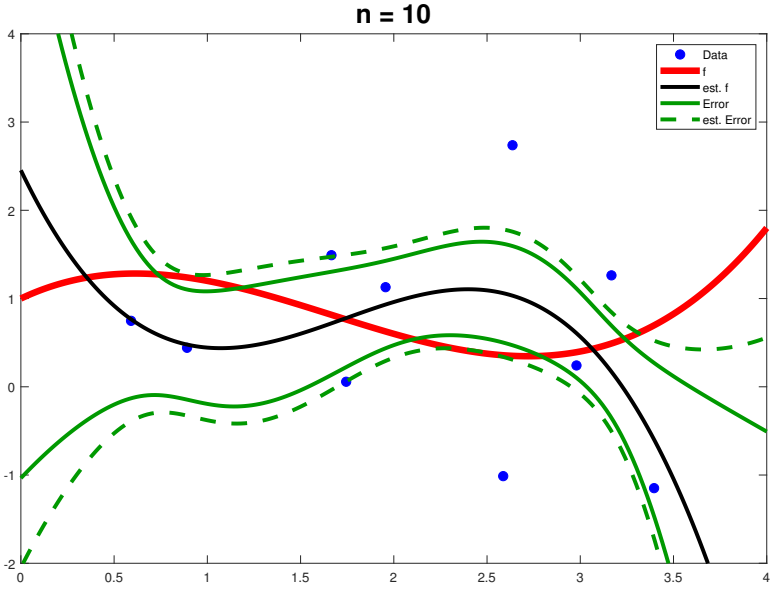
$$\widehat{f(x)} = F_x T^{-1} v, \qquad \mathrm{Var}(\widehat{f(x)}) = \sigma^2 F_x T^{-1} F_x^T$$

$$\widehat{\mathrm{Var}(\widehat{f(x)})} = \frac{V - v^T T^{-1} v}{n - m} F_x T^{-1} F_x^T$$

Пример части кода в MatLab

```
in = in + Info([x,y]); % Update: Elem. Info & Combine
est = in * xv;         % Apply Info
```

Пример иллюстрации



4 Простая Задача Линейного Оценивания и Каноническая Информация

Рассмотрим следующую серию измерений неизвестной величины x:

$$y_i = a_i x + \varepsilon_i, \qquad i = 1, \ldots, n,$$

где y_i - результаты измерений, a_i - известные коэффициенты и случайные величины ε_i, представляющие ошибки измерения, независимы и одинаково распределены с нулевым средним и дисперсией σ^2:

$$\mathsf{E}\varepsilon_i = 0, \qquad \mathsf{D}\varepsilon_i = \mathsf{E}\varepsilon_i^2 = \sigma^2, \qquad i = 1, \ldots, n,$$

a) Постройте оптимальную линейную по $y_1, \ldots, y_n$ оценку $\widehat{x}$ для x? $\widehat{x} = ?$

b) Является ли она несмещенной оценкой?

c) Какова ее дисперсия (выраженная через σ^2)? $\mathsf{D}\widehat{x} = ?$

d) Как бы Вы оценили σ^2 если она неизвестна? $\widehat{\sigma^2} = ?$

e) Что можно использовать в качестве оценки для $\mathsf{D}\widehat{x}$ если σ^2 неизвестна? $\widehat{\mathsf{D}\widehat{x}} = ?$

f) Предположим, что дисперсия σ^2 известна. Какую "каноническую информацию" было бы достаточно извлечь из серии данных

$$(y_1, a_1), \ldots, (y_n, a_n), \qquad i = 1, \ldots, n$$

для того, чтобы вычислить оценку $\widehat{x}$ и ее дисперсию $\mathsf{D}\widehat{x}$?

g) Предположим, что дисперсия σ^2 НЕ известна. Какую "каноническую информацию" было бы достаточно извлечь из серии данных для того, чтобы вычислить $\widehat{x}$, $\widehat{\sigma^2}$, and $\widehat{\mathsf{D}x}$?

h) Как следует обновлять такую "информацию" когда поступает новое "наблюдение" (y_{n+1}, a_{n+1})?

i) Как следует "объединять" информацию в канонической форме?

Пожалуйста, не пытайтесь использовать общие формулы, а получите все, насколько это возможно, с нуля.

5 Оптимальное Линейное Оценивание - Примеры

1. (Эта задача является частным случаем задачи 3. Поэтому если Вы достаточно уверены, можете пропустить ее, а затем просто извлечь ответы из задачи 3).

Рассмотрим следующий набор измерений неизвестных переменных x_1 и x_2:

$$y_1 = x_1 + x_2 + \nu_1,$$

$$y_2 = x_1 - x_2 + \nu_2,$$

$$y_3 = -x_1 + x_2 + \nu_3,$$

де y_i - результаты измерений, а ν_i - независимые одинаково распределенные случайные ошибку измерений с нулевым средним и дисперсией σ^2:

$$\mathsf{E}\nu_i = 0, \qquad \mathsf{E}\varepsilon_i^2 = \sigma^2, \qquad i = 1, 2, 3.$$

(a) Записать этот набор измерений в матричной форме

$$y = Ax + \nu$$

и укажите матрицы A и $S = \mathsf{D}\nu$.

(b) Найти матрицы вариаций $\mathsf{D}\widehat{x}$ для оптимальной линейной оценки x и дисперсии $\widehat{x_1}$ и $\widehat{x_2}$.

2. Рассмотрим два измерения одной неизвестной переменной x с коррелированными шумом:

$$y_1 = x + \nu_1,$$

$$y_2 = x + \nu_2,$$

где

$$\nu_1 = \varepsilon_1 + \varepsilon_0,$$

$$\nu_2 = \varepsilon_2 + \varepsilon_0,$$

$$\varepsilon_1, \varepsilon_2 \sim (0, \sigma_1^2), \qquad \varepsilon_0 \sim (0, \sigma_0^2),$$

$$\sigma_0^2 + \sigma_1^2 = \sigma^2, \qquad r = \frac{\sigma_0^2}{\sigma_0^2 + \sigma_1^2}.$$

(a) Как и в задаче 1.

(b) Как и в задаче 1.

(c) Проанализируйте, как дисперсия $\mathsf{D}\widehat{x}$ зависит от параметра корреляции r при $0 \leq r \leq 1$ и $\sigma^2 = const$. Хороша или плоха высокая корреляция для точности оценки в данном примере? Можно нарисовать график. Как можно объяснить такое поведение?

3. Рассмотрим ту же схему измерения, как и в задаче 1, но с коррелированным шумом:

$$\nu_1 = \varepsilon_1 + \varepsilon_0,$$

$$\nu_2 = \varepsilon_2 + \varepsilon_0,$$

$$\nu_3 = \varepsilon_3 + \varepsilon_0,$$

$$\varepsilon_1, \varepsilon_2, \varepsilon_3 \sim (0, \sigma_1^2), \qquad \varepsilon_0 \sim (0, \sigma_0^2),$$

$$\sigma_0^2 + \sigma_1^2 = \sigma^2, \qquad r = \frac{\sigma_0^2}{\sigma_0^2 + \sigma_1^2}.$$

(a) Как и в задаче 1.

(b) Как и в задаче 1.

(c) Проанализируйте, как дисперсии $\mathsf{D}\widehat{x_1}$ and $\mathsf{D}\widehat{x_2}$ зависит от параметра корреляции r при $0 \leq r \leq 1$ при $\sigma^2 = \text{const}$. Хороша или плоха высокая корреляция для точности оценки в данном примере? Можно нарисовать график. Как можно объяснить такое поведение?

Рекомендуется использовать символические пакеты (например, Maple, Symbolic Toolbox в MatLab, ...).

6 Каноническая Информация для Многочисленных Измерений и Априорной Информации

Напишите программу, демонстрирующую различные аспекты оптимального линейного оценивания. Модель измерения имитирует регистрацию сильно "сглаженного" и зашумленного сигнала.

В пунктах 1 и 2 ниже более подробно описаны некоторые предложения для вашей программы. Сама проблема (на самом деле, четыре тесно связанных варианта) сформулирована в пункте 3.

1. Моделирование измерения.

(a) Сгенерируйте случайный сигнал x, обладающий заданными корреляционными свойствами:

i. Сгенерируйте "белый шум" $\mu \sim (0, I)$, компоненты которого μ_i независимы и одинаково распределены $\mathsf{E}\mu_i = 0$ и $\mathsf{D}\mu_i = 1$.

ii. "Сгладьте" его с помощью некоторой матрицы B.

iii. Тогда $x = B\mu \sim (0, F)$, where $F = \mathsf{D}x = BB^T$.

(b) Постройте матрицу A.

(c) Сформируйте результат измерения $y = Ax + \nu$.

2. Оценивание: Постройте оптимальную линейную оценку $\widehat{x}$ и матрицу вариаций $\mathsf{D}(\widehat{x} - x)$. Покажите на графике:

(a) Исходный сигнал x (кривая с компонентами x_i),

(b) Его оценка $\widehat{x}$ (кривая с компонентами $\widehat{x}_i$),

(c) Стандартные отклонения для оценок $\widehat{x}_i$ $(= \sqrt{\mathsf{D}(\widehat{x}_i - x_i)} = \sqrt{\mathsf{D}(\widehat{x} - x)_{ii}}$. Их можно проиллюстрировать, изобразив соответствующий "коридор" около $\widehat{x}_i$).

3. Проиллюстрируйте оценки (пункт 2) для различных ситуаций:

(a) Однократное измерение (y, A, S).

i. Преобразуйте измерение (y, A, S) в каноническую форму (T, v).

ii. Построить оценку, используя каноническую информацию.

(b) Однократное измерение (y, A, S) с априорной информацией $x \sim (0, F)$:

i. Преобразуйте априорную информацию в каноническую форму.

ii. Преобразуйте измерение в каноническую форму.

iii. Объедините эти части канонической информации.

iv. Построить оценку, основанную на комбинированной канонической информации.

(c) Много измерений (y_j, A_j, S_j) без априорной информации.

i. Смоделируйте последовательность измерений одного и того же сигнала x, с различными матрицами A_j (и, возможно, S_j).

ii. Извлеките каноническую информацию из каждого измерения.

iii. Объедините части канонической информации.

iv. Постройте оценку, основанную на комбинированной канонической информации.

(d) Много измерений (y_j, A_j, S_j), но теперь с априорной информацией $x \sim (0, F)$.

i. Преобразуйте априорную информацию в каноническую форму.

ii. Смоделируйте последовательность измерений одного и того же сигнала x, с различными матрицами A_j (и, возможно, S_j).

iii. Извлеките каноническую информацию из каждого измерения.

iv. Объедините части канонической информации.

v. Постройте оценку, основанную на комбинированной канонической информации.

7 Задача Калибровки

В этом задании предлагается написать программу, реализующую и демонстрирующую оптимальную калибровку для линейного оценивания. В качестве основы используется процесс имитирующий простой эксперимент измерения сигнала, рассмотренный в домашнем задании 6. Здесь предлагаются этапы проекта.

1. Моделирование измерений:

   (a) Случайно генерируем некоторый «неизвестный» профиль $x \sim (0, F)$.

   (b) Строим матрицу A.

   (c) Проводим измерение

$$y = Ax + \nu, \quad \nu \sim (0, \sigma^2 I).$$

2. Калибровка:

   (a) Случайно генерируем калибровочные сигналы $\varphi_k,\ k = 1, \ldots, K$ так же, как и сигнал x.

   (b) Проводим калибровочные измерения

$$\psi_k = A\varphi_k + \nu_k.$$

   (c) Собираем каноническую калибровочную информацию (G, H).

3. Используя наблюдение y строим оптимальную линейную оценку $\widehat{x}$ и матрицу, описывающую погрешность $Q = \mathrm{Var}(\widehat{x} - x)$. Показываем на том же графике:

(a) Исходный сигнал x (кривая с компонентами x_i),

(b) Его оценка $\widehat{x}$ (кривая с компонентами $\widehat{x}_i$),

(c) Стандартные отклонения для оценок $\widehat{x}_i$

$$\sqrt{\mathsf{E}(\widehat{x}_i - x_i)^2} = \sqrt{\mathrm{Var}(\widehat{x} - x)_{ii}} = \sqrt{Q_{ii}}$$

можно проиллюстрировать, показывая соответствующий «коридор» вокруг $\widehat{x}_i$).

4. Иллюстрируем оценку (Фаза 3), когда не только увеличивается количество калибровочных измерений K, но также исходный сигнал x измеряется N раз:

(a) Проводим N измерений $y_n = Ax + \nu_n$ для $n = 1, \ldots, N$ и собираем соответствующую информацию (N, S), где $S = \sum_{n=1}^{N} y_n$.

(b) Проводим K калибровочных измерений, собирать калибровочную информацию (G, H) и вычисляем A_0 и J.

(c) Используя эти два типа информации строим и иллюстрируем (как в Фазе 3) оптимальную линейную оценку $\widehat{x}$ и ее точность.

(d) Делаем это для нескольких комбинаций чисел N и K “малые”, “средние” и “большие”.

(e) Можно более детально исследовать как точность оценки зависит от N и K. Для этого можно показать полную погрешность оценки

$$\mathsf{E}||\widehat{x} - x||^2 = \mathrm{tr}Q$$

как функцию N и K. Чтобы проиллюстрировать функцию двух переменных, вы можете показать ее,

например, как поверхность или как псевдоцветное изображение. Можно также изобразить контурные линии.

8 Многокритериальная Оптимизация: Компромисс Между Систематическими и Случайными Погрешностями

В этом домашнем задании вам предлагается написать программу, которая бы реализовала и продемонстрировала раздельный анализ систематических и случайных ошибок в линейной оценке. Базовый процесс имитирует простой эксперимент по измерению сигнала и в значительной степени основан на домашнем задании №6. Ниже приведены предлагаемые фазы проекта.

1. Постройте множества Парето (кривые) для различных информационных сценариев:

   (a) Очень мало информации (маленький и даже единичный T). Одно измерение с короткой (усеченной) матрицей A (количество наблюдений меньше количества неизвестных).

   (b) Умеренная информация. Около 4 измерений с нормальной A.

   (c) Больше информации. Около 16 измерений с нормальной A.

   и покажите их на одном графике.

2. Проиллюстрируйте результаты оценки для нескольких (6) интересных случаев:

   (a) Нет случайного подавления шума ($\lambda \approx 0$).

   (b) Умеренное или сильное подавление шума.

для различных информационных сценариев.

3. Для каждой иллюстрации оценки покажите на одном графике:

(a) Исходный сигнал x (кривая с компонентами x_i),

(b) Его оценка $\widehat{x}$ (кривая с компонентами $\widehat{x}_i$),

(c) Стандартные отклонения для случайной (аддитивной) ошибки $\sqrt{H_{ii}}$. Ее можно проиллюстрировать, показав соответствующий "коридор" вокруг $\widehat{x}_i$).

(d) Иллюстрация систематической (мультипликативной) ошибки $\sqrt{G_{ii}}$. Возможно, на отдельном графике.

Instead of the original parametrization by $\lambda \in [0, +\infty]$ it might be convenient to use $\lambda = \frac{t}{1-t}$ where $t \in [0, 1]$.

9 Наименьшее Информационное Пространство

Пусть $\{x_1, ..., x_n\} \in \mathbb{R}^+$ будет мультимножеством действительных чисел. Для обработки $p : \mathbb{R}^+ \to \mathbb{R} \cup \{\text{☹}\}$, которая вычисляет разброс точек:

$$p(\{x_1, ..., x_n\}) = \begin{cases} \max\limits_{0 \le i \le n} x_i - \min\limits_{0 \le i \le n} x_i & n > 0 \\ \text{☹} & n = 0 \end{cases}$$

найдите кандидата на наименьшее информационное пространство и докажите, что оно действительно наименьшее.