Вводная лекция

Методы оптимизации

Александр Безносиков

Московский физико-технический институт

5 сентября 2024

ФПМИ

Команда курса: лектор

- Безносиков Александр Николаевич
- почта: beznosikov.an@phystech.edu, anbeznosikov@gmail.com
- tg: @abeznosikov

Команда курса: семинаристы

- Андреев Артем Викторович
  tg: @artyomandreyev
- Богданов Александр Иванович
  tg: @d0dos
- Былинкин Дмитрий Андреевич
  tg: @lxstsvund
- Кормаков Георгий Владимирович
  tg: @gkormakov
- Корнилов Никита Максимович
  tg: @Tugnir

Команда курса: лектор и семинаристы

- Моложавенко Александр Александрович
  tg: @MetaMelon
- Ткаченко Светлана
  tg: @Aikiseito
- Чежегов Савелий Андреевич
  tg: @Savochak
- Юдин Никита Евгеньевич
  tg: @nikitayudin

Команда курса: ассистенты

- Кузнецова Алина
- Лисов Петр
- Пичугин Александр
- Ребриков Алексей
- Торопин Иван
- Хафизов Фанис
- Янко Иван

Команда курса: помощь с материалами

- Баширов Наиль
- Подлипнова Ирина
- Прохоров Борис
- Шестаков Александр

Telegram беседа

Диск с материалами

Форма обратной связи

Таблица с оценками

Правила игры: система оценивания

Вид активности	**Баллы**
тесты на 10 минут в начале каждого семинара по теме прошлой лекции и прошлого семинара	3
домашнее задание (выдается каждую неделю)	3 + 3
контрольная работа в середине семестра по темам лекций и семинаров	3
коллоквиум в конце семестра по темам лекций и семинаров	3
разбор статьи	2
Итого:	17

Правила игры: система оценивания

- Для получения оценки удовлетворительно и выше необходимо, чтобы было выполнено хотя бы одно из следующих условий: оценка за коллоквиум ≥ 1, оценка за КР ≥ 1.
- Для получения оценки хорошо и выше необходимо, чтобы было выполнено оба следующих условия: оценка за коллоквиум ≥ 1, оценка за КР ≥ 1.

Правила игры: комментарии

- Ни один из видов активности не является 100% обязательным, но смотри дополнительные правила выше.
- Тесты будут проводиться на каждом семинаре.
- ДЗ будут появляться в четверг. Время на выполнение: 2 недели
- ДЗ состоит из двух частей: основной и дополнительной. Основная часть легче и предполагается, что ее достаточно для хорошего погружения в курс, дополнительная часть – для более глубого погружения в заинтересовавшие темы.
- При подозрении в списывании ДЗ баллы за конкретное домашнее задание обнуляются у всех авторов, подозреваемых в списывании (в том числе и у тех, кто дал списать).
- КР проводится в середине семестра в отдельный день (суббота/воскресенье).

Правила игры: комментарии

- Коллоквиум проходит в конце семестра во время последнего семинара и на зачетной неделе (на выбор). Программа коллоквиума соответствует всей программе курса, изученной в рамках лекций и семинаров. Принимают коллоквиум семинарист и несколько приглашенных преподавателей. Процедура коллоквиума соответствует процедуре проведения обычного устного экзамена на Физтехе с билетами, дополнительными вопросами/задачами и беседой в рамках курса.
- Разбор статьи предполагает полный разбор популярной статьи в области оптимизации. Необходимо разобраться в идее, в доказательствах, а также воспроизвести численные эксперименты, добавив к ним свои. Отчетность: 15 минутное выступление с презентацией на семинарском занятии. Заявки на разбор статьи принимаются до 31 октября.

Задача оптимизации

- **Вопрос:** какие приложения задач оптимизации знаете/уже встречали?

Задача оптимизации

- **Вопрос:** какие приложения задач оптимизации знаете/уже встречали? Приложений у оптимизации масса: машинное обучение, анализ данных, статистика, финансы, логистика, планирование, управление и многие другие.
- Оптимизация часто выступает инструментом во многих прикладных задачах, люди пользуются готовыми решениями/пакетами/"черными ящиками", который способны решать задачи оптимизации. Цель курса – как познакомиться с такими "черными ящиками", так и заглянуть внутрь и понять, что лежит внутри (в том числе с токи зрения теории).

Машинное обучение

Вопрос: Как формулируется задача оптимизации в машинном обучении?

- На практике обычно имеется только конечная выборка/конечный набор данных $(a_1, b_1), \ldots, (a_n, b_n)$, на котором формулируется задача минимизации эмпирической функции потерь:

$$\min_{x \in \mathbb{R}^d} \left\{ \hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \ell(g(x, a_i), b_i) \right\}. \tag{1}$$

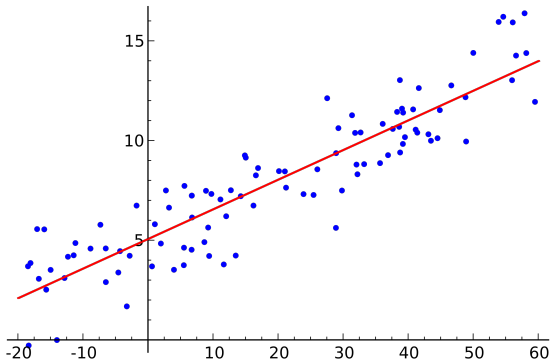
- Однако предполагается, что выборка взята из некоторого распределения $\mathcal{D}$, и хочется получить хорошее приближенное решение задачи минимизации ожидаемой функции потерь:

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) = \mathbb{E}_{(a,b) \sim \mathcal{D}}[\ell(g(x, a), b)] \right\}. \tag{2}$$

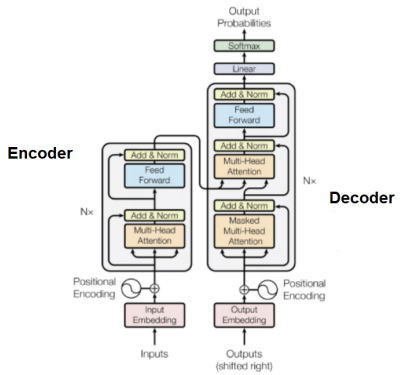
Машинное обучение

Вопрос: Какие примеры ℓ и g уже встречались?

- Линейная регрессия: $g(x, a) = \langle x, a \rangle$, $\ell(b_1, b_2) = (b_1 - b_2)^2$.



- NLP: g – NLP transformer (LLM), $l(b_1, b_2) = -\int b_1(x) \log b_2(x) dx$ (cross-entropy на прдсказание следующего слова)



Машинное обучение

- (1) называется оффлайн постановкой, (2) обычно называется онлайн постановкой. **Вопрос:** почему?
- В (1), у нас есть выборки данных, которые можно обрабатывать как угодно, забыв об исходном распределении $\mathcal{D}$. В (2) может не быть выборки. Данные поступают в сеть по одной точке за раз. Нам нужно обрабатывать эти данные в момент их поступления, времени на сбор большого набора данных нет.

Машинное обучение

- **Вопрос:** Возникает естественный вопрос - как связаны эти задачи? (1) является Монте-Карло аппроксимацией интеграла (2).

Теорема

Если $\ell(g(x, a_i), b_i)$ выпукла по x для любых a_i, b_i, M-Липшицева по x, и мы ограничим область поиска минимума множеством Q с диаметром D, тогда для $\hat{x}^* = \operatorname{argmin}_{x \in Q} \hat{f}(x)$ с вероятностью хотя бы $1 - \delta$

$$\left| f(\hat{x}^*) - \min_{x \in Q} f(x) \right| = O\left(\sqrt{\frac{M^2 D^2 d \ln(n) \ln(d/\delta)}{n}} \right).$$

Машинное обучение

- Популярность оптимизации в машинном обучении

Adam: A method for stochastic optimization 193709 2014
DP Kingma, J Ba
arXiv preprint arXiv:1412.6980

- Одна из ключевых секций на ведущих ML конференциях.

Статистика

Предположим, что некоторая переменная b зависит от переменных $a_1, a_2, a_3, \ldots, a_{d-1}$ линейным образом:

$$b(a_1, \ldots, a_{d-1}) = x_0 + x_1 a_1 + x_2 a_2 + \ldots + x_{d-1} a_{d-1},$$

где коэффициенты $x_0, \ldots, x_{d-1}$ нам неизвестны. Предположим, что мы хотим найти эти коэффициенты, измеряя переменную b при различных значениях $a_1, \ldots, a_{d-1}$. Казалось бы, в этом нет ничего сложного, ведь для решения системы достаточно провести d измерений. **Вопрос:** какая проблема?В действительности измерения производятся с некоторой погрешностью: для заданного набора $a_1^i, a_2^i, \ldots, a_{d-1}^i$ мы измеряем

$$b_i = x_0 + x_1 a_1 + \ldots + x_{d-1} a_{d-1} + \xi_i,$$

где $\xi_i \sim \mathcal{N}(0, \sigma^2)$.

Статистика

- Другими словами, мы предполагаем, что $b_i \sim \mathcal{N}(x_0 + x_1 a_1 + \ldots + x_{d-1} a_{d-1}, \sigma^2)$, где параметры $x = (x_0, \ldots, x_{d-1})^\top$ должны быть найдены по выборке $\{b_i\}_{i=1}^n$ (мы будем считать, что $b_1, \ldots, b_n$ – независимые случайные величины). **Вопрос:** Из каких соображений выбрать параметры $x_0, \ldots, x_{d-1}$?
- Можно, например, рассмотреть оценку максимального правдоподобия (для краткости введем вектор $x^i = (1, x_1^i, \ldots, x_{d-1}^i)^\top$):

$$\begin{aligned} x^* &= \arg\max_{x \in \mathbb{R}^d} \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(b_i - \langle a^i, x\rangle)^2\right) \\ &= \arg\max_{x \in \mathbb{R}^d} \ln\left(\prod_{i=1}^d \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(b_i - \langle a^i, x\rangle)^2\right)\right). \end{aligned}$$

Статистика

- Поскольку логарифм произведения равен сумме логарифмов, а аддитивные и мультипликативные константы не меняют точку оптимума, получаем:

$$\begin{aligned} x^* &= \arg\max_{x\in\mathbb{R}^d} \left\{ \text{Const} + \sum_{i=1}^{n} -\frac{1}{2\sigma^2}(b_i - \langle a^i, x\rangle)^2 \right\} \\ &= \arg\min_{x\in\mathbb{R}^d} \sum_{i=1}^{n} (b_i - \langle a^i, x\rangle)^2 \\ &= \arg\min_{x\in\mathbb{R}^d} \|Ax - b\|_2^2, \end{aligned}$$

где A составлена из строк $(a^i)^\top$.

Рюкзак

- На производстве могут выпускать товары d типов, нужно оптимально организовать производство (разработать план $\{x_i\}_{i=1}^d$, сколько товаров каждого типа производить). Чтобы
  1. выполнялись ограничения на рабочую силу (для производства 1 у.е. товара iго типа в месяц нужно a_i работников, всего работников A): $\sum_{i=1}^d a_i x_i \le A$,
  2. была возможность закупить сырье (для производства 1 у.е. товара iго типа в месяц нужно b_j сырья j при этом цена формируется в зависимости от размера заказа по функции $c_j(\cdot)$, всего средств на закупку C): $\sum_{i=1}^d \sum_{j=1}^n c_j(b_j x_i) \le C$,
  3. а главное максимизировалась прибыль от продажи (цена за 1 у.е. товара iго типа d_i): $\max_{\{x_i\}\ge 0} d_i x_i$.

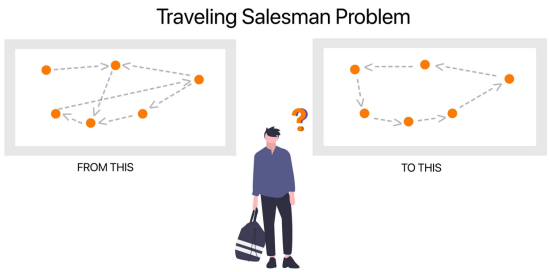
Вопрос: по-хорошему x_i дискретны, а мы решаем непрерывную задачу это сильно страшно? Зависит от масштабов производства: округление может дать хороший результат, но существуют и более хитрые трюки, как решать дискретные задачи через непрерывные.

Задача коммивояжера

- На плоскости заданы N точек с координатами $\{x_i, y_i\}_{i=1}^{N}$, нужно построить путь кратчайшей длины, проходящий через все точки. Формально:

$$\min_{\text{path}} \left[\sum_{i=1}^{N-1} \left(x_{\text{path}(i+1)} - x_{\text{path}(i)}\right)^2 + \left(y_{\text{path}(i+1)} - y_{\text{path}(i)}\right)^2 \right].$$

Существует масса способов ее решения, один из самых популярных – линейное программирование.



Немного истории

- 1847: Коши и градиентный спуск для линейных систем
- 1950ые: линейное программирование (быстро перешло в нелинейное программирование), появление стохастических методов
- 1980ые: появление теории для общих задач.
- 2010ые: задачи оптимизации большого размера, теория стохастических методов

Задача оптимизации

$$\begin{array}{c} \min \quad f(x) \\ g_i(x) \& 0, \\ i=1,\ldots,m, \\ x \in Q \end{array}$$

- $Q \subseteq \mathbb{R}^d$ — подмножество d-мерного пространства
- $f : Q \to \mathbb{R}$ — некоторая функция, заданная на множестве Q
- В качестве & берётся $\leq$ либо $=$
- $g_i(x) : Q \to \mathbb{R}$, $i = 1, \ldots, m$ — функции, задающие ограничения

Вопрос: что можно сказать про эту задач? сложная ли эта задача?

Задачи оптимизации. Первые наблюдения.

1. В общем случае задачи оптимизации могут не иметь решения. Например, задача $\min_{x \in \mathbb{R}} x$ не имеет решения.
2. Задачи оптимизации часто нельзя решить аналитически.
3. Их сложность зависит от вида целевой функции f, множества Q и может зависеть от размерности x.

Если же задача оптимизации имеет решение, то на практике её обычно решают, вообще говоря, приближённо. Для этого применяются специальные алгоритмы, которые и называют методами оптимизации.

Методы оптимизации

- Нет смысла искать лучший метод для решения конкретной задачи. Например, лучший метод для решений задачи $\min_{x \in \mathbb{R}^d} \|x\|^2$ сходится за 1 итерацию: этот метод просто всегда выдаёт ответ $x^* = 0$. Очевидно, что для других задач такой метод не пригоден.
- Эффективность метода определяется для класса задач, т.к. обычно численные методы разрабатываются для *приближённого* решения множества однотипных задач.
- Метод разрабатывается для класса задач $\Longrightarrow$ метод не может иметь с самого начала полной информации о задаче. Вместо этого метод использует модель задачи, например, формулировку задачи, описание функциональных компонент, множества, на котором происходит оптимизация и т.д.

- Предполагается, что численный метод может накапливать специфическую информацию о задаче при помощи некоторого *оракула*. Под оракулом можно понимать некоторое устройство (программу, процедуру), которое отвечает на последовательные вопросы численного метода.

Вопрос: Какого рода вопросы хочется задавать оракулу?

Примеры оракулов

- **Оракул нулевого порядка** в запрашиваемой точке x возвращает значение целевой функции $f(x)$.
- **Оракул первого порядка** в запрашиваемой точке возвращает значение функции $f(x)$ и её градиент в данной точке $\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n}\right)$.
- **Оракул второго порядка** в запрашиваемой точке возвращает значение и градиент функции $f(x), \nabla f(x)$, а также её гессиан в данной точке $\left(\nabla f^2(x)\right)_{ij} = \frac{\partial^2 f(x)}{\partial x_j \partial x_j}$.

Общая итеративная схема метода оптимизации $\mathcal{M}$

Входные данные: начальная точка x^0 (0 – верхний индекс), требуемая точность решения задачи $\varepsilon > 0$.
Настройка. Задать $k = 0$ (счётчик итераций) и $I_{-1} = \varnothing$ (накапливаемая информационная модель решаемой задачи).
Основной цикл

1. Задать вопрос к оракулу $\mathcal{O}$ в точке x^k.
2. Пересчитать информационную модель: $I_k = I_{k-1} \cup (x^k, \mathcal{O}(x^k))$.
3. Применить правило метода $\mathcal{M}$ для получения новой точки x^{k+1} по модели I_k.
4. Проверить критерий остановки $\mathcal{T}_\varepsilon$. Если критерий выполнен, то выдать ответ $\bar{x}$, иначе положить $k := k + 1$ и вернуться на шаг 1.

Примеры итерационных методов. Градиентный спуск

Рассмотрим задачу оптимизации

$$\min_{x \in \mathbb{R}^d} f(x),$$

где функция $f(x)$ дифференцируема. Предположим, что в любой точке мы можем посчитать её градиент.

Алгоритм 1 Градиентный спуск с постоянным размером шага

Вход: размер шага $\gamma > 0$, стартовая точка $x^0 \in \mathbb{R}^d$, количество итераций K

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: Вычислить $\nabla f(x^k)$
3: $x^{k+1} = x^k - \gamma \nabla f(x^k)$
4: **end for**

Выход: x^K

Вопрос: в чем Алгоритм 1 отличается от определения общей итеративной схемы? В итеративной схеме использовался $\mathcal{T}_\varepsilon$.

Критерии останова

- По аргументу: $\|x^k - x^*\| \leq \varepsilon.$

  Вопрос: какие проблемы тут видим?

  - x^* – неизвестно, но можно так:
    $$\|x^{k+1} - x^k\| \leq \|x^{k+1} - x^*\| + \|x^k - x^*\|.$$
    Тогда если $\|x^{k+1} - x^*\| \leq \|x^k - x^*\| \leq \varepsilon$, следует $\|x^{k+1} - x^k\| \leq 2\varepsilon$ (в обратную сторону, очевидно, неверно). $\|x^{k+1} - x^k\| \leq \varepsilon$ – это скорее практический вариант критерия, который работает, если есть понимание (интуиция), что $\|x^k - x^*\| \to 0$.
  - x^* – не уникально. Тогда можно поменять следующий критерий
- По функции: $f(x^k) - f^* \leq \varepsilon.$

  Часто f^* известно, например, для $f(x) = \|Ax - b\|^2$. На практике можно использовать $|f(x^k) - f(x^{k+1})|$.
- По норме градиента: $\|\nabla f(x^k)\| \leq \varepsilon.$

  Вопрос: когда такой критерий можно использовать? В безусловной оптимизации

Сложность методов оптимизации

- **Аналитическая/Оракульная сложность** — число обращений к оракулу, необходимое для решения задачи с точностью ε.
- **Арифметическая/Временна'я сложность** — общее число вычислений (включая работу оракула), необходимых для решения задачи с точностью ε.