
ОГЛАВЛЕНИЕ

Предисловие	14
Что такое собеседование по проектированию систем МО (ML System Design interview).....	14
Почему это важно	14
Для кого эта книга	15
Чего нет в книге	15
Дополнительные ресурсы	15
Благодарности.....	16
От издательства	18
Глава 1. Введение и общие сведения	19
Прояснение требований.....	20
Формулировка проблемы в виде задачи МО.....	21
Определение цели МО	22
Определение входных и выходных данных системы	22
Выбор подходящей категории МО	24
Темы для обсуждения	25
Подготовка данных	25
Инженерия данных	26
Типы данных.....	27
Конструирование признаков.....	29
Операции конструирования признаков	30
Темы для обсуждения	34
Разработка модели	34
Выбор модели.....	34
Обучение модели	36
Темы для обсуждения	40
Оценка.....	41
Автономная (offline) оценка	41
Оперативная (online) оценка	42
Темы для обсуждения	43

Развертывание и эксплуатация.....	43
Развертывание в облаке или на устройстве	43
Сжатие модели.....	44
Тестирование при эксплуатации	44
Пайплайн предсказаний	46
Темы для обсуждения	48
Мониторинг	48
Почему в действующей системе происходят сбои	49
Какие показатели нужно отслеживать.....	50
Инфраструктура	50
Итоги	51
Ссылки	51
Глава 2. Система визуального поиска	54
Прояснение требований.....	54
Формулировка проблемы в виде задачи МО	56
Определение цели МО	56
Определение входных и выходных данных системы	56
Выбор подходящей категории МО	56
Подготовка данных	58
Инженерия данных	58
Конструирование признаков.....	60
Разработка модели	60
Выбор модели.....	60
Обучение модели	61
Построение датасета.....	62
Выбор функции потерь	64
Оценка	66
Автономные метрики	66
Оперативные метрики	72
Эксплуатация.....	72
Предсказательный пайплайн	73
Пайплайн индексации.....	74
Эффективность алгоритмов поиска ближайшего соседа (NN).....	75
Какой алгоритм использовать?.....	77
Другие темы для обсуждения.....	78
Итоги	79
Ссылки	80

Глава 3. Система размытия в Google Street View	81
Прояснение требований.....	81
Формулировка проблемы в виде задачи МО.....	82
Определение цели МО	82
Определение входных и выходных данных системы.....	83
Выбор подходящей категории МО	83
Двухступенчатые сети.....	83
Одноступенчатые сети	84
Сравнение одноступенчатых и двухступенчатых сетей	84
Подготовка данных	85
Инженерия данных	85
Конструирование признаков.....	86
Разработка модели	88
Выбор модели.....	88
Обучение модели	90
Оценка.....	91
Автономные метрики	93
Оперативные метрики	95
Эксплуатация.....	95
Перекрытие ограничительных прямоугольников.....	95
Проектирование системы МО	96
Пайплайн пакетных предсказаний	96
Другие темы для обсуждения.....	98
Итоги	99
Ссылки	100
Глава 4. Поиск видео на YouTube	101
Прояснение требований.....	101
Формулировка проблемы в виде задачи МО.....	102
Определение цели МО	102
Определение входных и выходных данных системы.....	102
Выбор категории МО	103
Подготовка данных	105
Инженерия данных	105
Конструирование признаков.....	106
Разработка модели	110
Выбор модели.....	110
Обучение модели	116

8 Оглавление

Оценка.....	117
Автономные метрики.....	117
Оперативные метрики.....	118
Эксплуатация.....	119
Предсказательный пайплайн.....	120
Пайплайн индексации видео.....	121
Пайплайн индексации текста.....	121
Другие темы для обсуждения.....	121
Итоги.....	123
Ссылки.....	124
Глава 5. Обнаружение вредоносного контента	125
Прояснение требований.....	126
Формулировка проблемы в виде задачи МО.....	127
Определение цели МО.....	127
Определение входных и выходных данных системы.....	127
Выбор категории МО.....	131
Подготовка данных.....	137
Инженерия данных.....	137
Конструирование признаков.....	138
Разработка модели.....	143
Выбор модели.....	143
Обучение модели.....	144
Оценка.....	146
Автономные метрики.....	146
Оперативные метрики.....	147
Эксплуатация.....	148
Сервис обнаружения вредоносного контента.....	148
Сервис обработки нарушений.....	148
Сервис понижения приоритета.....	148
Другие темы для обсуждения.....	149
Итоги.....	150
Ссылки.....	151
Глава 6. Система рекомендации видео.....	153
Прояснение требований.....	154
Формулировка проблемы в виде задачи МО.....	155
Определение цели МО.....	155
Определение входных и выходных данных системы.....	156
Выбор категории МО.....	156

Подготовка данных	161
Инженерия данных	161
Конструирование признаков	163
Разработка модели	168
Матричная факторизация	168
Двухбашенная нейронная сеть	175
Работа двухбашенной нейронной сети	177
Матричная факторизация или двухбашенная нейронная сеть?	179
Оценка	179
Автономные метрики	179
Оперативные метрики	180
Эксплуатация	181
Генерация кандидатов	182
Скоринг	184
Повторное ранжирование	184
Трудности при разработке систем рекомендации видео	185
Другие темы для обсуждения	186
Итоги	187
Ссылки	188
Глава 7. Система рекомендации событий	189
Прояснение требований	190
Формулировка проблемы в виде задачи MO	191
Определение цели MO	191
Определение входных и выходных данных системы	191
Выбор подходящей категории MO	192
Списочные методы LTR	193
Подготовка данных	194
Инженерия данных	194
Конструирование признаков	196
Разработка модели	204
Выбор модели	204
Обучение модели	212
Оценка	214
Автономные метрики	214
Оперативные метрики	214
Эксплуатация	215
Пайплайн оперативного обучения	215
Предсказательный пайплайн	216

Другие темы для обсуждения.....	218
Итоги.....	219
Ссылки.....	220
Глава 8. Предсказание кликов по рекламе на социальных платформах	221
Введение	221
Прояснение требований.....	222
Формулировка проблемы в виде задачи МО.....	223
Определение цели МО	223
Определение входных и выходных данных системы.....	223
Выбор категории МО.....	224
Подготовка данных.....	225
Инженерия данных	225
Конструирование признаков.....	226
Признаки пользователя.....	228
Разработка модели.....	230
Выбор модели.....	230
Обучение модели	241
Оценка.....	242
Автономные метрики	242
Оперативные метрики	244
Эксплуатация.....	245
Пайплайн подготовки данных	246
Пайплайн непрерывного обучения	247
Предсказательный пайплайн	247
Другие темы для обсуждения.....	248
Итоги	249
Ссылки	250
Глава 9. Похожие объекты на платформах краткосрочной аренды жилья.....	251
Прояснение требований.....	252
Формулировка проблемы в виде задачи МО.....	253
Определение цели МО	253
Определение входных и выходных данных системы.....	253
Выбор категории МО.....	254
Подготовка данных.....	255
Инженерия данных	255
Конструирование признаков.....	256

Разработка модели	257
Выбор модели.....	257
Обучение модели	257
Построение датасета.....	258
Оценка.....	262
Автономные метрики	262
Оперативные метрики	264
Эксплуатация.....	264
Обучающий пайплайн	264
Индексирующий пайплайн.....	265
Предсказательный пайплайн	266
Другие темы для обсуждения.....	267
Итоги.....	268
Ссылки	269
Глава 10. Персонализированная лента новостей	270
Введение	270
Прояснение требований.....	271
Формулировка проблемы в виде задачи MO.....	272
Определение цели MO	272
Определение входных и выходных данных системы.....	274
Выбор категории MO.....	274
Подготовка данных	275
Инженерия данных	275
Конструирование признаков.....	277
Признаки пользователей	280
Разработка модели	282
Выбор модели.....	282
Обучение модели	285
Оценка.....	287
Автономные метрики	287
Оперативные метрики	287
Эксплуатация.....	289
Предсказательный пайплайн	289
Другие темы для обсуждения.....	290
Итоги.....	291
Ссылки	292

Глава 11. Списки возможных знакомых	293
Введение	293
Прояснение требований.....	293
Формулировка проблемы в виде задачи MO.....	294
Определение цели MO	294
Выбор категории MO	295
Подготовка данных	299
Инженерия данных	299
Конструирование признаков.....	301
Разработка модели	303
Выбор модели.....	303
Графовые нейронные сети (GNN)	303
Обучение модели	304
Оценка.....	308
Автономные метрики	308
Оперативные метрики	308
Эксплуатация.....	309
Эффективность.....	309
Проектирование системы MO	311
Другие темы для обсуждения.....	314
Итоги	315
Ссылки	316
Послесловие.....	317

Посвящается Нилуфар, которая была и остается
моим лучшим другом.

Али Аминнган

Посвящается Джулии.

Алекс Сюй

ПРЕДИСЛОВИЕ

Мы рады, что вы читаете эту книгу, чтобы лучше подготовиться к собеседованию по проектированию систем машинного обучения (МО). **Проектирование систем (system design)** — одна из самых сложных тем на любом собеседовании из области МО, так что без хорошей подготовки не обойтись.

Что такое собеседование по проектированию систем МО (ML System Design interview)

Собеседование по проектированию систем МО, как правило, обязательно для претендентов на вакансии, связанные с проектированием и реализацией систем МО: инженер данных, дата-сайентист, инженер машинного обучения и т. д.

На таком собеседовании оценивается умение кандидата проектировать комплексные системы МО — визуальный поиск, рекомендации видео, предсказание кликов по рекламе и т. д. Вопросы на собеседовании бывают непростыми, потому что у них обычно нет четкой структуры. Часто однозначных ответов не существует, потому что темы вопросов оказываются широкими и затрагивают разные области, что открывает возможность для различных интерпретаций и решений.

Чтобы успешно пройти собеседование по проектированию систем МО, надо хорошо понимать фундаментальные концепции и методы МО, а также уметь их применять, чтобы решать практические задачи. На собеседовании обычно необходимо продемонстрировать, что вы разбираетесь в пайплайнах данных и конструировании признаков, а также умеете проектировать эффективные системы МО. Возможно, вам еще придется проявить умение выбирать подходящие модели для конкретных задач, настраивать их параметры и оценивать производительность. В принципе, цель собеседования состоит в том, чтобы оценить, насколько хорошо соискатель применяет теоретические знания МО, чтобы проектировать и реализовывать эффективные системы.

Почему это важно

Большинство соискателей, которые проходят собеседования по МО, владеют теоретическими основами, но затрудняются в области проектирования систем

МО, где нет единых руководящих принципов. Тем не менее умение проектировать такие системы — это важнейший навык для инженеров, особенно в контексте карьерного роста. Неправильно выбранная архитектура системы МО может привести к значительным потерям времени и других ресурсов.

Очевидно, собеседование по проектированию систем МО — это важнейшая часть процедуры приема на работу. Чем лучше вы себя покажете, тем вернее можете рассчитывать на более привлекательную позицию и более высокий заработок.

Для кого эта книга

Книга будет ценным источником информации для всех, кто интересуется проектированием систем МО, будь то новички или опытные инженеры. А если вам нужно подготовиться к собеседованию по МО, то эта книга написана специально для вас.

Чего нет в книге

Эта книга — не пособие по основам машинного обучения. Она написана для дата-сайентистов, инженеров данных и инженеров МО, которым нужна помощь, чтобы подготовиться к собеседованию по проектированию систем МО. Книга предназначена в первую очередь для инженеров МО в бизнесе и в меньшей степени для ученых в области МО в образовательных учреждениях или НИИ.

Дополнительные ресурсы

В конце каждой главы приводится длинный список ссылок на дополнительные материалы. Все активные ссылки доступны в репозитории GitHub:

<https://bit.ly/ml-bytebytego>



БЛАГОДАРНОСТИ

Как бы нам ни хотелось похвастаться, что все примеры проектирования в этой книге полностью оригинальны, все же придется признаться, что большинство идей, которые здесь рассматриваются, можно найти и в других местах: технических блогах, исследовательских статьях, коде, презентациях на YouTube и прочих источниках. Мы собрали эти блестящие идеи, исследовали их и дополнили собственными наблюдениями и опытом, чтобы представить в простой и понятной форме. Авторы хотели бы сказать огромное спасибо многим техническим специалистам и руководителям, которые внесли вклад в работу над этой книгой и рецензировали ее:

- Винит Ахлувалия (Vineet Ahluwalia) (Стэнфордский университет)
- Топоджой Бисвас (Topojoy Biswas) (Walmart)
- Да Чэн (Da Cheng) (Tiktok)
- Рохит Джайн (Rohit Jain) (Twitter)
- Кальян Дипак (Kalyan Deepak) (Flipkart)
- Димитрис Котсакос (Dimitris Kotsakos) (Elastic)
- Субхам Кумар (Subham Kumar) (Amazon)
- Джастин Ли (Justin Li) (Discord)
- Ли Сюй (Li Xu) (TikTok)
- Рави Мандлия (Ravi Mandliya) (Discord)
- Саранг Меткар (Sarang Metkar) (Meta)
- Шабаз Патель (Shabaz Patel) (One Concern)
- Каустубх Пхаднис (Kaustubh Phadnis) (Walmart)
- Рави Рамчандран (Ravi Ramchandran) (Walmart Labs)
- Дэван Султания (Dewang Sultania) (Adobe)
- Сяо Чжу (Xiao Zhu) (Databricks)
- Цзяин Ши (Jiaying Shi) (Amazon)
- Цзяньцян Ван (Jianqiang Wang) (Snapchat)
- Чжэхуэй Ван (Zhehui Wang) (Amazon)

- Шо Сян (Shuo Xiang) (Parafin)
- Бонту Шридеви (Bonthu Sridevi) (Технологический институт Вишну)
- Диала Эззеддин (Diala Ezzeddine) (Tao Media)
- Юаньцзюнь Ян (Yuanjun Yang) (Twitter)

Наконец, мы хотим особо поблагодарить Элвиса Жэня (Elvis Ren), Хуа Ли (Hua Li) и Сана Лама (Sahn Lam) за неоценимый вклад в работу над книгой.

ОТ ИЗДАТЕЛЬСТВА

Ваши замечания, предложения, вопросы отправляйте по адресу comp@piter.com (издательство «Питер», компьютерная редакция).

Мы будем рады узнать ваше мнение!

На веб-сайте издательства www.piter.com вы найдете подробную информацию о наших книгах.

1 ВВЕДЕНИЕ И ОБЩИЕ СВЕДЕНИЯ

Мы написали эту книгу, чтобы помочь инженерам по машинному обучению и дата-сайентистам успешно пройти собеседование по проектированию систем МО. Книга также может пригодиться всем, кто хочет получить общее представление о том, как МО применяется в реальном мире.

Многие технические специалисты полагают, что системы МО исчерпываются такими алгоритмами МО, как логистическая регрессия или нейронные сети. Тем не менее реальные системы МО далеко не ограничиваются разработкой моделей. Эти системы обычно весьма сложны; они состоят из множества компонентов, включая стеки данных, служебную инфраструктуру (благодаря которой система становится доступной миллионам пользователей), пайплайн для оценки ее эффективности, а также средства мониторинга, которые следят за тем, чтобы качество модели не ухудшалось со временем.

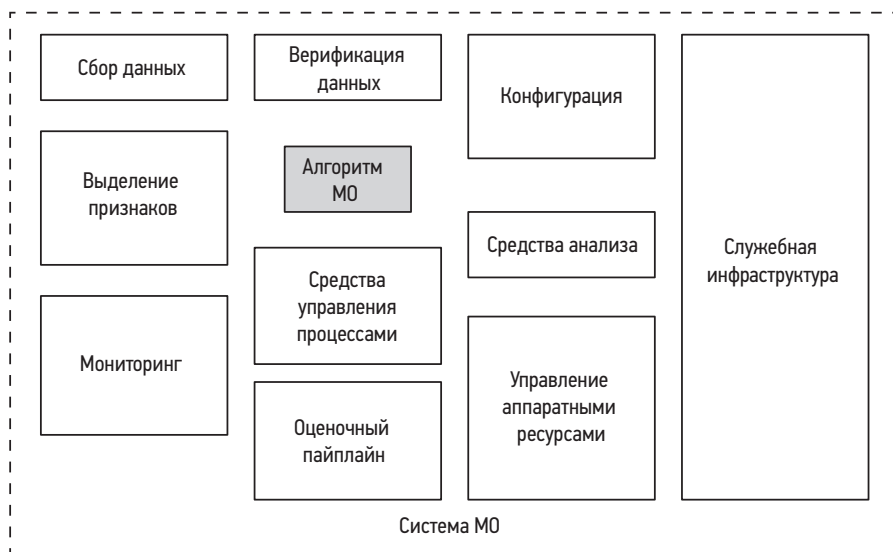


Рис. 1.1. Компоненты системы МО, готовой к эксплуатации

Скорее всего, на собеседовании по проектированию систем МО вам предстоит отвечать на вопросы открытого типа. Например, вам могут предложить спроектировать систему для рекомендаций фильмов или службу поиска видео. У таких задач нет единственно правильного решения. Эксперт, проводящий собеседование,

хочет посмотреть, как вы размышляете, глубоко ли понимаете различные темы из области МО, умеете ли проектировать комплексные системы и находить компромиссы между конфликтующими факторами, которые влияют на проектирование.

Чтобы успешно проектировать сложные системы МО, очень важно придерживаться определенной логики. Без структуры сложно разобраться в проектировочных решениях. Здесь мы предлагаем схему, на которую в этой книге будут опираться примеры проектирования систем МО. Схема состоит из семи основных шагов.

1. Прояснение требований
2. Формулировка проблемы как задачи МО
3. Подготовка данных
4. Разработка модели
5. Оценка
6. Развертывание и эксплуатация
7. Мониторинг и инфраструктура



Рис. 1.2. Основные шаги проектирования систем МО

Каждое собеседование по проектированию систем МО отличается от других, потому что вопросы носят открытый характер и не существует универсальных удачных решений. Эта схема помогает упорядочить мысли, но строго следовать ей необязательно. Сохраняйте гибкость. Если эксперта, проводящего собеседование, в первую очередь интересует разработка модели, почти всегда стоит подстраиваться под его запросы.

Давайте подробнее рассмотрим каждый шаг этой схемы.

Прояснение требований

Вопросы на собеседовании по проектированию систем МО обычно намеренно ставятся нечетко, с минимумом информации. Например, вопрос может звучать так: «Спроектируйте систему для рекомендации событий». Прежде всего стоит задать уточняющие вопросы. Но какие именно? Нужны такие вопросы, которые помогут понять конкретные требования. Этот систематизированный список вопросов можно взять за основу.

- **Бизнес-цель.** Если система должна рекомендовать отпускное жилье для бронирования, то цели могут заключаться в том, чтобы увеличить количество бронирований и выручку.

- **Функции, которые должна поддерживать система.** Какие из требуемых функциональных возможностей могут повлиять на проектирование системы МО? Допустим, вам предложено спроектировать систему для рекомендации видео. Вероятно, стоит уточнить, могут ли пользователи ставить рекомендуемому контенту лайки или дизлайки, потому что этими оценками можно размечать обучающие данные.
- **Данные.** Откуда поступают данные? Каков размер датасета? Размечены ли данные?
- **Ограничения.** Какая вычислительная мощность доступна? Будет ли система работать в облаке или на локальном устройстве? Планируется ли, что со временем модель будет автоматически совершенствоваться?
- **Масштаб системы.** Сколько пользователей будет у системы? С каким количеством объектов (например, видеороликов) придется иметь дело? С какой скоростью растут эти показатели?
- **Производительность.** Насколько быстрыми должны быть предсказания? Должна ли система работать в реальном времени? Что важнее — точность или низкая задержка?

Это не исчерпывающий список, но его можно принять за отправную точку. Не забывайте, что могут быть и другие важные аспекты — например, конфиденциальность и этика.

Предполагается, что к концу этого этапа вы согласуете с экспертом рамки системы и требования к ней. Обычно имеет смысл составить список требований и ограничений: это поможет убедиться, что все одинаково представляют себе задачу.

Формулировка проблемы в виде задачи МО

При решении задач МО крайне важно правильно сформулировать проблему. Допустим, эксперт предлагает вам увеличить степень вовлеченности пользователей на платформе видеостриминга. Безусловно, недостаточная вовлеченность — это проблема, но это не задача МО. Таким образом, чтобы решить проблему, ее следует переформулировать в виде задачи МО.

В реальности сначала надо выяснить, действительно ли для решения проблемы нужно МО. Однако на собеседовании по проектированию систем МО логично допустить, что от МО все-таки будет польза. Таким образом, чтобы сформулировать проблему как задачу МО, можно поступить так:

- Определить цель МО.
- Определить входные и выходные данные системы.
- Выбрать подходящую категорию МО.

Определение цели МО

Бизнес-цель может состоять в том, чтобы повысить продажи на 20 % или увеличить степень удержания пользователей. Однако цели не всегда определяются четко, а модель невозможно обучить, просто приказав ей увеличить продажи на 20 %. Чтобы система МО решила задачу, нужно преобразовать бизнес-цель в точно определенную цель МО. Хорошей целью МО будет такая, которую можно решить с помощью моделей МО. Некоторые примеры перечислены в табл. 1.1, а в дальнейших главах встретятся и другие примеры.

Таблица 1.1. Преобразование бизнес-целей в цели МО

Приложение	Бизнес-цель	Цель МО
Приложение для продажи билетов на мероприятия	Повысить продажи билетов	Максимизировать количество регистраций на мероприятия
Видеостриминговое приложение	Повысить степень вовлеченности пользователей	Максимизировать время, которое пользователи проводят за просмотром видео
Система прогнозирования кликов по рекламе	Увеличить количество кликов	Максимизировать CTR (кликабельность)
Выявление вредоносного контента в социальной сети	Сделать платформу безопаснее	Точно предсказывать, является ли контент вредоносным
Система рекомендации друзей	Увеличить темп расширения сети пользователя	Максимизировать количество установленных связей

Определение входных и выходных данных системы

Когда цель МО ясна, нужно определить входные и выходные данные системы. Например, для системы выявления вредоносного контента в социальной сети входными данными является пост, а выходными — решение о том, считать ли его вредоносным.

Иногда система может состоять более чем из одной модели МО. В таком случае требуется определить входные и выходные данные для каждой модели. Например, в примере с выявлением вредоносного контента одна модель может выявлять призыв к насилию, а другая — непристойные изображения. Система опирается на обе эти модели, чтобы решить, считать ли пост вредоносным.

Еще одно важное соображение состоит в том, что может существовать несколько способов определить входные и выходные данные модели — см. пример на рис. 1.4.

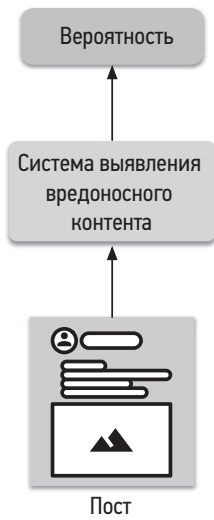


Рис. 1.3. Входные и выходные данные системы выявления вредоносного контента

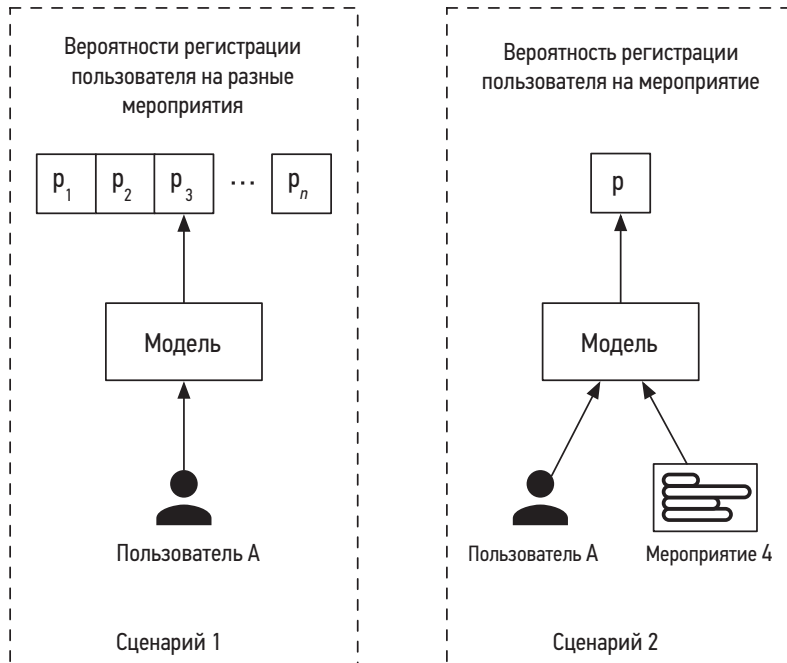


Рис. 1.4. Разные способы определения входных и выходных данных модели

Выбор подходящей категории МО

Существует много способов переформулировать проблему в виде задачи МО. Большинство проблем можно представить так, чтобы они относились к одной из категорий МО, изображенных на рис. 1.5. Поскольку эти категории, вероятно, уже знакомы большинству читателей, мы ограничимся краткой сводкой.

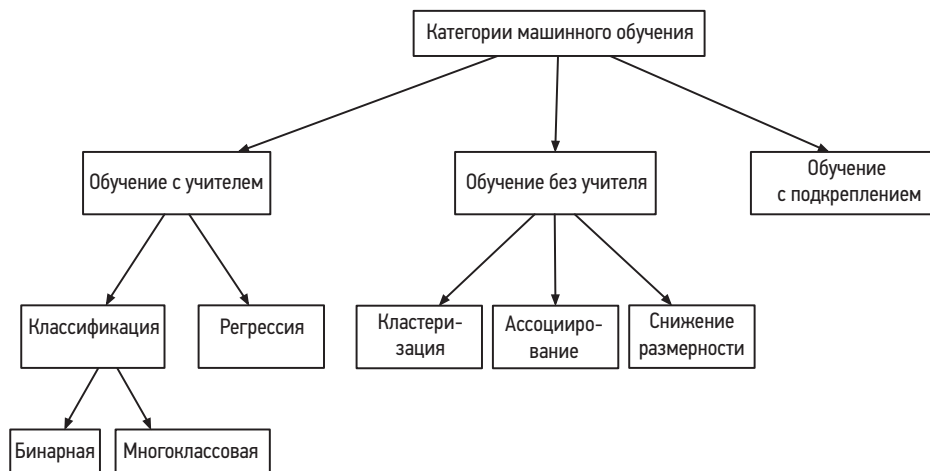


Рис. 1.5. Распространенные категории МО

Обучение с учителем. В этих моделях используется обучающий набор данных. На практике многие проблемы относятся к этой категории, потому что обучение на размеченном наборе данных обычно приводит к лучшим результатам.

Обучение без учителя. Чтобы делать предсказания, такие модели обрабатывают данные, которые не содержат правильных ответов. Цель обучения — выявить осмысленные закономерности в данных. Популярные алгоритмы обучения без учителя — кластеризация, ассоциирование и снижение размерности.

Обучение с подкреплением. Система учится решать задачу, многократно взаимодействуя со средой методом проб и ошибок. Например, таким способом можно научить робота ходить по комнате или натренировать такую программу, как AlphaGo, чтобы она успешно соревновалась с человеком в игре го.

По сравнению с обучением с учителем, обучение без учителя и обучение с подкреплением менее популярны в реальных системах, потому что модели МО обычно лучше обучаются, если есть обучающие данные. Поэтому для большинства проблем, которые рассматриваются в этой книге, применяется обучение с учителем. Давайте поближе познакомимся с разными его видами.