Занятие 1. Введение в машинное обучение

Гирдюк Дмитрий Викторович

6 сентября 2024

СПбГУ, ПМ–ПУ, ДФС

Машинное обучение [1]

- Универсальное определение Тома Митчелла: «A computer program is said to learn from experience E with respect to some class of tasks T, and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E».
- Машинное обучение подразделяется на то, что за задачи T оно решает, какова природа функции оценки качества P, и какие данные E мы подаем ему на вход.

Основные понятия и обозначения

- Предположим, что существует некоторая среда или генеральная совокупность и связанное с ней вероятностное распределение $f.(\cdot)$.
- Дискретный набор данных из среды, выборку, обозначим через $\mathcal{D}$.
- Методы машинного обучения преднозначены для построения модели $h(\cdot;\theta)$, аппроксимирующей $f.(\cdot)$ в том или ином смысле. Через θ обозначены параметры модели.
- Конкретный смысл $f, \mathcal{D}, h$ напрямую зависит от типа задачи, с которой мы имеем дело.

Обучение с учителем

- Постановка задачи обучения с учителем (supervised learning): необходимо предсказать значение целевой переменной/отклик/таргет $y \in Y$ объекта по набору его признаков/фичей (features) $x \in X$.
- Тогда среда описывается совместным распределением $f_{X,Y}(x, y)$, а выборкой из нее является набор пар $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$.
- Результирующая модель возвращает значение y по признакам x: $y = h(x; \theta)$.

Обучение с учителем: подтипы [2]

- Классификация:
  - $Y = \{0, 1\}$ – бинарная (binary)
  - $Y = \{1, 2, \ldots, K\}$ – многоклассовая (multiclass)
  - $Y = \{0, 1\}^K$ – многозначная/многометочная? (multi-label)
- Регрессия:
  - $Y = R$ – одномерная (ordinal)
  - $Y = R^K$ – многомерная (multiple)
- Ранжирование:
  - Y – конечное упорядоченное множество

Примеры задач обучения с учителем

- Классификация: кредитный скоринг, выбор типа лечения пациента.
- Регрессия: предсказание цены товара, стоимость недвижимости.
- Ранжирование: поисковая выдача, рекомендация похожих товаров в интернет-магазине.

Обучение без учителя

- В противовес задачам обучения с учителем, в задачах обучения без учителя (unsupervised learning) значение целевой переменной отсутствует в выборке.
- Среда описывается распределением $f_X(x)$, из которого формируется выборка $\mathcal{D} = \{(x^{(i)})\}_{i=1}^{N}$.
- В зависимости от подтипа задачи, результирующая модель $h(x;\theta)$ может возвращать как некоторое целевое значение y, которое может отражать как номер кластера в задаче кластеризации, так и исходный объект в новом признаковом пространстве, например, в случае снижения размерности.

Примеры задач обучения без учителя

- Кластеризация: выделение групп пользователей соцсети, соцопросы.
- Поиск аномалий: обнаружение неполадок приборов по данным датчиков, подозрительные банковские операции.
- Ассоциативные правила: анализ рыночных корзин, выделение терминов в тексте.

Обучение с частичным привлечением учителя

- На стыке двух рассмотренных типов задач лежит обучение с частичным привлечением учителя (semi-supervised learning).
- Постановка задачи тут аналогична постановке задачи обучения с учителем, однако выборка разбивается на две части. Первая часть выборки содержит отклик $\mathcal{D}_1 = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$, а вторая нет $\mathcal{D}_2 = \{x_i\}_{i=N+1}^{N+M}$.
- Примеры задач аналогичны таковым в задачах обучения с учителем. Встречаются там, где объемы данных велики (среди миллионов транзакций проверена лишь малая доля), и/или стоимость разметки экспертами слишком высока.

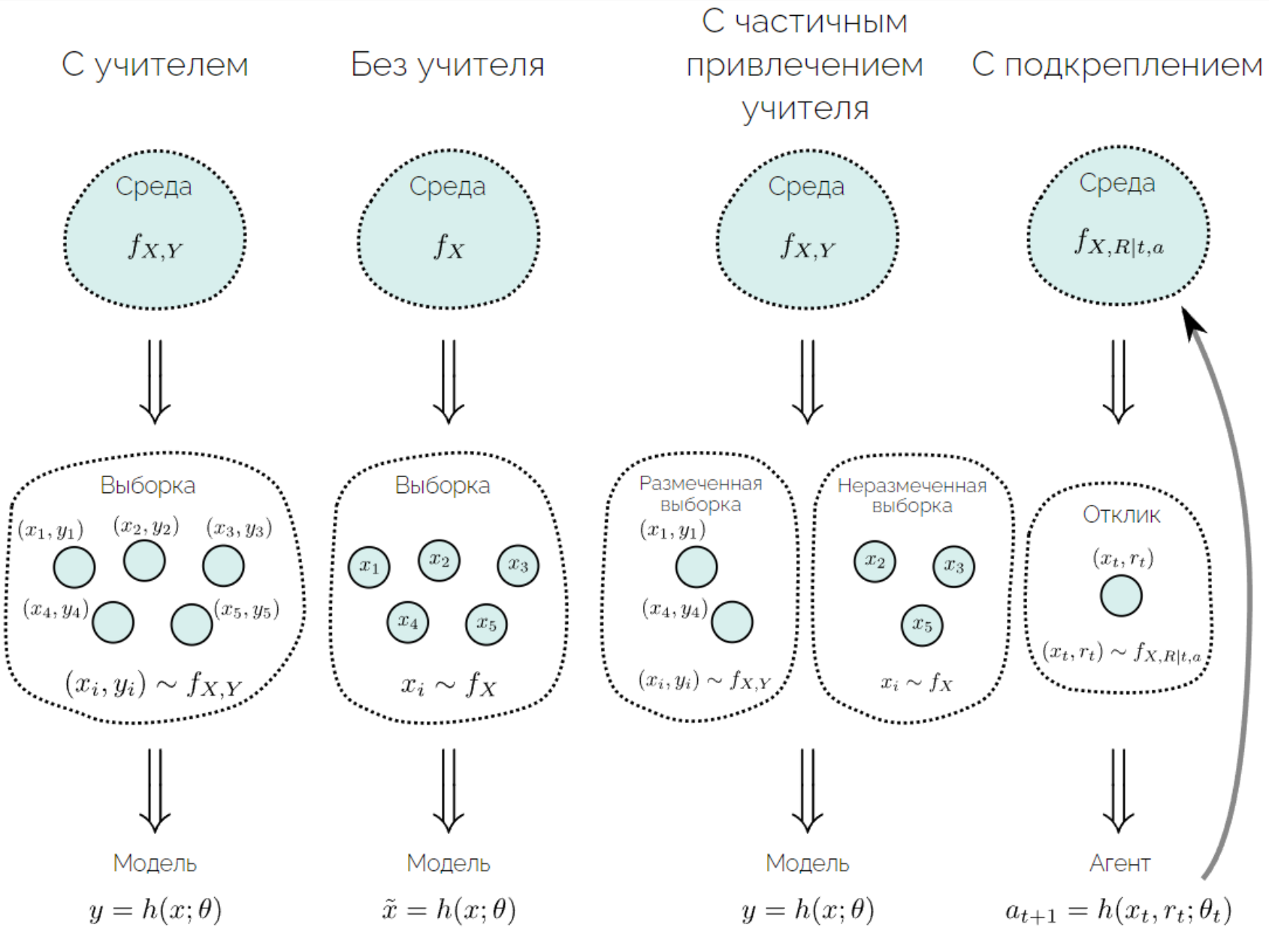
Обучение с подкреплением

- Фундаментальным отличием обучения с подкреплением (reinforcement learning) от рассмотренных ранее состоит в активном взаимодействием с, в общем случае, изменяющейся средой.
- Модель в таком сценарии принято называть агентом, который взаимодействует со средой в дискретные моменты времени t, принимая на вход не только состояние среды x_t, но и некоторую абстрактную награду r_t за свои предыдущие действия. На выходе модель выдает новое действие $a_{t+1} = h(x_t, r_t; \theta_t)$, которое оказывает влияние на среду, чем закольцовывает процесс.
- Изменение среды провоцирует изменение модели, а именно ее параметров θ_t.

Примеры задач обучения с подкреплением

- Рекомендательные системы: система рекомендаций подстраивается под изменяющиеся вкусы пользователей.
- Управление беспилотными системами: автопилоты, роботы-помошники.
- Игры как прокси-цель: шахматы, го, и даже дота со старкрафтом.
- В обработке естественного языка: RLHF в LLM.

Типы задач машинного обучения [3]



Типы задач машинного обучения

- Обучение с учителем (supervised learning)
  - Классификация (classification)
  - Регрессия (regression)
  - Ранжирование (learning to rank)
  - Частичное обучение (semi-supervised learning)
- Обучение без учителя (unsupervised learning)
  - Кластеризация (clustering)
  - Преобразование пространства признаков: снижение размерности (dimensionality reduction), матричные разложения (matrix factorization), etc.
  - Поиск аномалий (anomaly detection)
  - Самообучение (self-supervised learning)
  - Поиск ассоциативных правил (association rule learning)
- Обучение с подкреплением (reinforcement learning)

Типы моделей машинного обучения

- Модели машинного обучения принято подразделять по двум классификациям.
- Параметрические/непараметрические.
- Генеративные/дискриминативные.

Параметрические и непараметрические модели

- Параметрические модели содержат набор параметров, который явно не зависит от размера выборки.
- Непараметрические модели – модели, у которых нет параметров? Нет! Модели, количество параметров которых явно зависит от размера выборки.

Примеры моделей машинного обучения

- Многомерная линейная регрессия

$$y = h(x; \theta) = \theta_0 + \sum_{j=1}^{p} \theta_j x_j = \theta^T \tilde{x}$$

где вектор $x = [1, x_1, \ldots, x_p]^T$ дополнен единицей.

- Метод k ближайших соседей (k-nearest neighbors, kNN):

$$y = h(x; k) = \frac{1}{k} \sum_{x^{(i)} \in N_k(x)} x^{(i)}$$

где через $N_k(x)$ обозначено множество из k ближайших соседей объекта x с некоторой заранее заданной метрикой ρ (например, евклидовой).

Пример: линейная регрессия

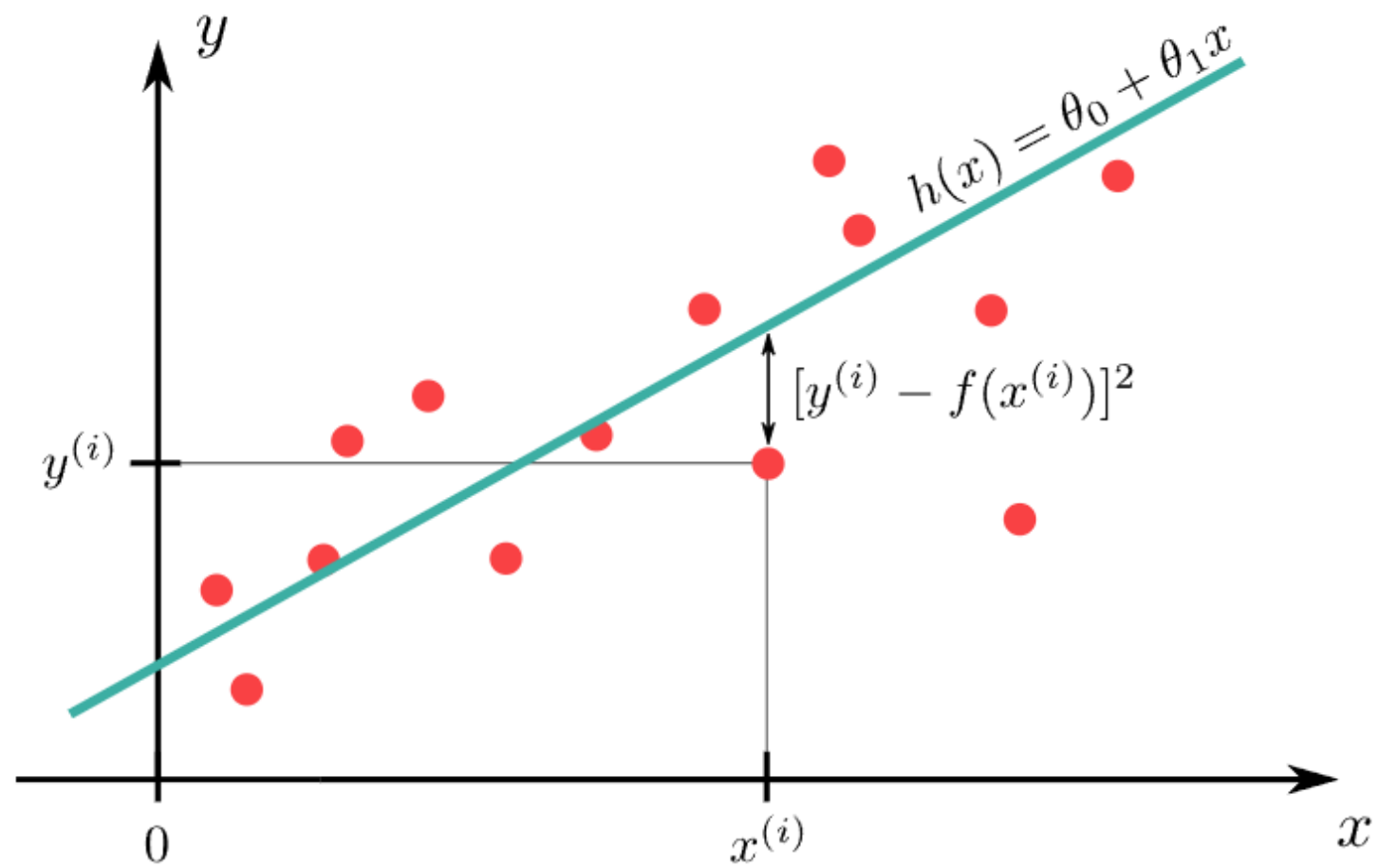


Иллюстрация одномерной линейной регрессии.

- Вопрос, как будет выглядеть метод ближайшего соседа для данного набора точек?

Генеративные и дискриминативные модели

- Генеративные модели стремятся восстановить полное распределение среды $f_{X,Y}$, тогда как дискриминативные лишь условное распределение $f_{Y|X}$. Вопрос: к каким типам моделей относятся линейная регрессия и kNN?
- Из генеративных моделей всегда можно получить условное распределение $f_{Y|X}$. Понятно, что они куда более трудозатратны с вычислительной точки зрения.
- Генеративные модели чаще всего используются при обучении без учителя, дискриминативные – при обучении с учителем.

Обучение моделей

- Рассматривая типы задач машинного обучения, мы опускали важный вопрос: как именно происходит обучение?
- Введем понятие функции потерь (лосс-функции) $L(y, h(x; \theta))$ – численное выражение ошибки пресказания y моделью h на по набору признаков x объекта.
- Пример функции потерь для задачи регрессии – квадрат отклонения $L(y, h(x; \theta)) = (y - h(x, \theta))^2$. Для задачи классификации – $L(y, h(x; \theta)) = I[y \neq h(x; \theta)]$.

Теория статистического обучения

- Теория статистического обучения: рассматривая случайные величины X и Y, ищем такую модель, которая минимизирует совместное матожидание функции потерь

$$E[L(Y, h(X; \theta)] = \int L(y, h(x; \theta)) f_{X,Y}(x, y) dxdy$$

- Но так как совместное распределение $f_{X,Y}$ неизвестно, подменяем матожидание выборочным средним по обучающей выборке

$$Q(h) = \frac{1}{N} \sum_{i=1}^{N} L(y^{(i)}, h(x^{(i)}; \theta))$$

- Функционал выше называется эмпирическим риском.

Обобщающая способность модели

- Допустим, мы знаем как нам минимизировать эмпирический риск (обучать модель). Гарантирует ли это, что на других данных из того же распределения $f_{X,Y}$ наша модель будет показывать себя так же хорошо?
- Вообще говоря, совершенно необязательно.
- Возникает вопрос, как нам оценивать способность модели обобщаться на данные, которые она раньше явно не использовала при обучении?

Тренировочное и тестовое множества

- В машинном обучении принято делить обучение модели на 2 этапа: на собственно обучение/тренировку и применение/тестирование.
- Для этого мы разделяем исходную выборку $\mathcal{D}$ на две части: тренировочную и тестовую/отложенную выборки.
- На тренировочной, очевидно, обучаем модель. Тестовую выборку используем только для проверки (в идеале, единожды): ошибка на ней характеризует то, насколько ваша модель хорошо будет справляться "в реальной жизни".

Диллема смещение-дисперсия

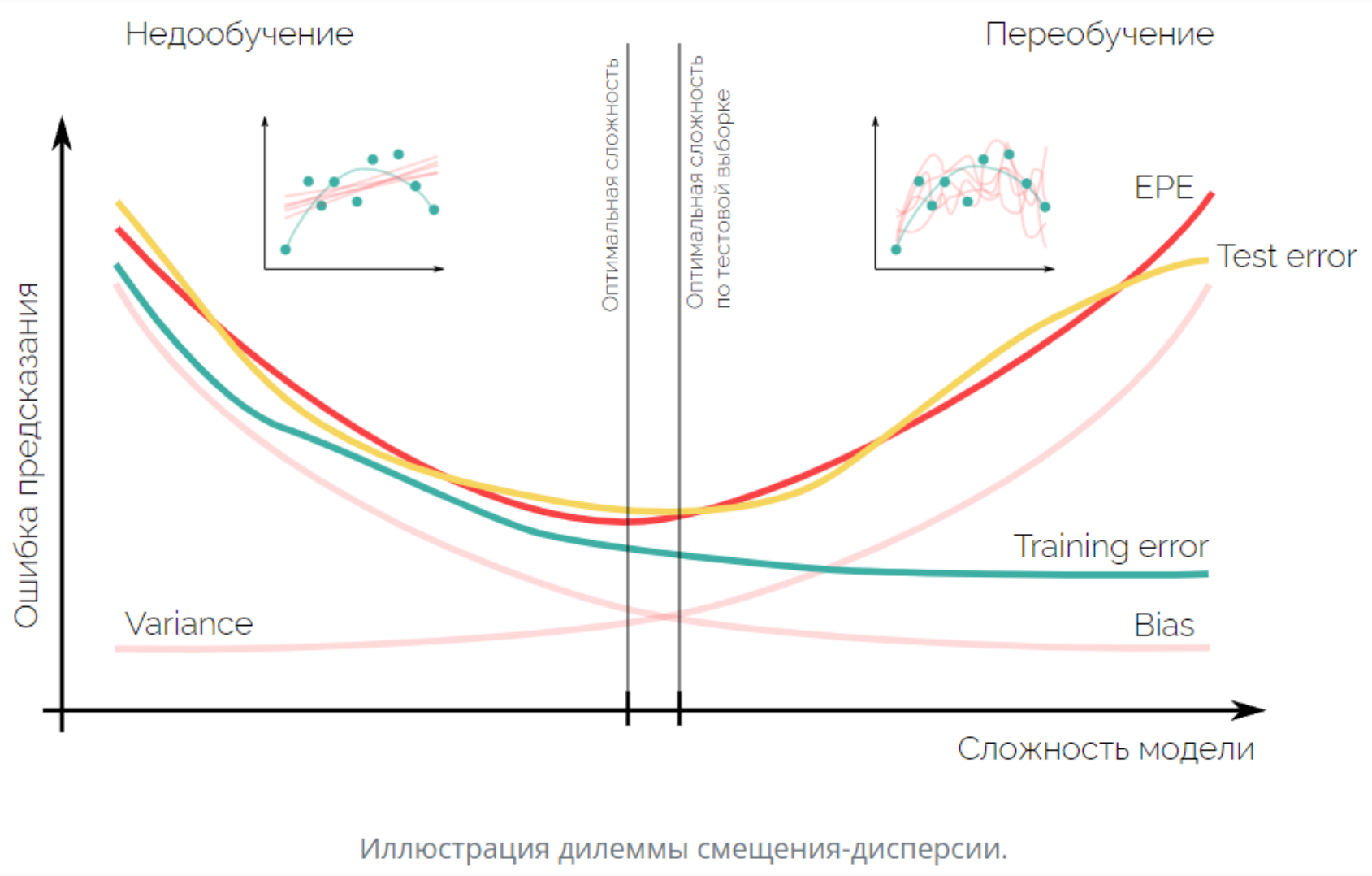


Иллюстрация дилеммы смещения-дисперсии.

(*) EPE – expected prediction error

buzzwords?

- Statistical learning
- Machine learning (ML)
- Data mining
- Pattern recognition
- Predictive analytics
- Data science (DS)
- Artificial intelligence (AI)

Методология CRISP-DM

- CRISP-DM (Cross-Industry Standard Process for Data Mining) — наиболее распространённая методология по исследованию данных.
- Фазы цикла исследования данных:
  - Понимание бизнес-целей (Business Understanding).
  - Начальное изучение данных (Data Understanding).
  - Подготовка данных (Data Preparation).
  - Моделирование (Modeling).
  - Оценка (Evaluation).
  - Внедрение (Deployment).

Итого

- Познакомились с базовыми понятиями машинного обучения, его классификацией и примерами прикладных задач.
- Получили представление о классификации типов моделей машинного обучения.
- Разобрались с названиями областей знаний, покрывающих, пересекающих и являющихся частью машинного обучения.
- Узнали об общей методологии работы над проектами по исследованию данных.
- Репозиторий курса: [4].

Использованные источники i

1. *Murphy K. P.* **Probabilistic Machine Learning: An introduction.** MIT Press, 2022. URL: probml.ai.
2. *Воронцов К.* **Презентация по основным задачам машинного обучения из курса лекций Воронцова К.В.** URL: http://www.machinelearning.ru/wiki/images/f/fc/Voron-ML-Intro-slides.pdf.
3. *Першин А.* **Введение в машинное обучение.** URL: http://getsomemath.ru/subtopic/machine_learning/basics_of_ml/intro_to_ml.
4. *Гирдюк Д.* **Репозиторий с материалами курса.** URL: https://github.com/dmgirdyuk/2023-spbu-dfs-ml101.