



ИСП РАН



ДОВЕРЕННЫЙ
ИСКУССТВЕННЫЙ
ИНТЕЛЛЕКТ

Современные методы машинного обучения

Лекция 1

Турдаков Денис Юрьевич

План

- Вводная часть о курсе
- Задачи машинного обучения
 - Классификация (наивный байесовский классификатор)
 - Кластеризация (k-means)
- Проблемы доверенного искусственного интеллекта

Часть 1. О курсе

О курсе

- Официальное название спецкурса:
 - бакалавриат — «Современные методы машинного обучения»
 - ~~магистратура~~
- Лекции по средам в 18.00 (ауд. 613)
 - предполагаются базовые знания
 - линейной алгебры,
 - теории вероятности и математической статистики,
 - программирования
 - не все имеют одинаковые знания
 - предполагается, что студенты могут быстро учиться

О курсе

- Курс ВМК МГУ, ИСП РАН
 - «Живые» лекции очно/в Zoom от сотрудников Исследовательского центра **доверенного искусственного интеллекта** ИСП РАН, в т.ч. с кафедр ВМК МГУ (СП, АЯ, ИИТ)
 - Теоретические основы
 - Опыт применения на практике
 - Наиболее интересные современные направления
- Итоговая отчетность: устный экзамен

Часть 2. Машинное обучение

Технологии искусственного интеллекта



Компьютер выполняет «интеллектуальные» действия, которые до этого были доступны только человеку:

- Играет в шахматы
- Ведет простой диалог

Решения основанные на правилах

Искусственный интеллект

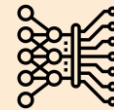


Машинное обучение

Решения основанные на данных:

- **Этап обучения:**
алгоритм получает примеры вопросов и правильных ответов
- **Этап применения:**
алгоритм получает вопрос и выдает ответ

Глубокое обучение



Иерархия алгоритмов машинного обучения: ответы одних попадают в вход другим

Модель нейронной сети

LLM

1950

1980

2010

2022

Переход от построения модели объекта автоматизации к решению проблемы по аналогии

Основные понятия

- X – множество объектов
- Y – множество меток
- $y : X \rightarrow Y$ – неизвестная зависимость (целевая функция), значения которой известны на конечном подмножестве объектов $\{x_1, \dots, x_l\} \subset X$
- $X^l = (x_i, y_i)_{i=1}^l$ – обучающая выборка

Постановка задачи

- По X^l восстановить зависимость y

Признаки объектов

- $f_j : X \rightarrow D_j, j = 1 \dots n$ – признаки объектов (features)
- Множество объектов задается матрицей

$$F = ||f_j(x_i)||_{l \times n} = \begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_l) & \dots & f_n(x_l) \end{pmatrix}$$

Алгоритм и модель

- В общем случае зависимость y узнать невозможно, поэтому будем ее приближать некоторой функцией $a : X \rightarrow Y$
- Функция $a : X \rightarrow Y$ должна допускать эффективную компьютерную реализацию; по этой причине будем называть её **алгоритмом**
- Поиск оптимального алгоритма осуществляют из предположения, что $a \in A = \{g(x, \theta) | \theta \in \Theta\}$ – принадлежит семейству параметрических функций (**модель**), где $g : X \times \Theta \rightarrow Y$ – фиксированная функция, Θ – пространство поиска

Задачи машинного обучения

Задача	На какой вопрос про входные данные пытаются ответить
Классификация (Classification)	Это A или B ?
Кластеризация (Clustering)	Как эти данные могут быть сгруппированы?

Задача	На какой вопрос про входные данные пытаются ответить
Восстановление регрессии (Regression)	Сколько этого?
Генерация (Generative models)	Как это выглядит? Что похоже на это?
Обучение ранжированию (Learning to Rank)	Упорядочить A и B
Обучение с подкреплением (Reinforcement learning)	Что мне делать сейчас?
Поиск аномалий (Anomaly detection)	Это аномалия?

Классификация: вероятностная постановка

- $X \times Y$ – вероятностное пространство с распределением $p(x, y) = p(y)p(x|y)$ из которого случайно и независимо выбирается l наблюдений $X^l = (x_i, y_i)_{i=1}^l$
- Будем аппроксимировать $p(x, y)$ через модель совместной плотности распределения объектов и ответов $\phi(x, y, \theta)$
- Определим значение параметров θ , при которых обучающая выборка данных максимально правдоподобна, то есть наилучшим образом согласуется с моделью плотности (метод максимума правдоподобия)

$$L(\theta, X^l) = \prod_{i=1}^l \phi(x_i, y_i, \theta) \rightarrow \max$$

Наивный байесовский классификатор

- Выбор наиболее вероятного значения

$$\hat{y} = \arg \max_{y \in Y} P(y|x) = \arg \max_{y \in Y} P(y|f_1, \dots, f_n)$$

- По правилу Байеса

$$\hat{y} = \arg \max_{y \in Y} \frac{P(f_1, \dots, f_n|y)P(y)}{P(f_1, \dots, f_n)} = \arg \max_{y \in Y} P(f_1, \dots, f_n|y)P(y)$$

- «Наивное» предположение об условной независимости признаков

$$\hat{y} = \arg \max_{y \in Y} P(y) \prod_{i=1}^n P(f_i|y)$$

Обучение наивного байесовского классификатора

- Сделаем предположение об априорном распределении

$$p(f_i|y, \theta) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\pi\sigma_y^2}\right)$$

- Воспользуемся методом максимального правдоподобия для оценки среднего и дисперсии распределения
- Оценка для $p(y)$ – частоты классов в выборке D
- Получим оценку $p(x,y)$
 - Можем предсказывать y для произвольного x
 - Можно даже генерировать (x,y) . Это генеративная модель

Пример

```
from sklearn.nayve_bayes import *

corpus = [['list of texts'], ['classes']]

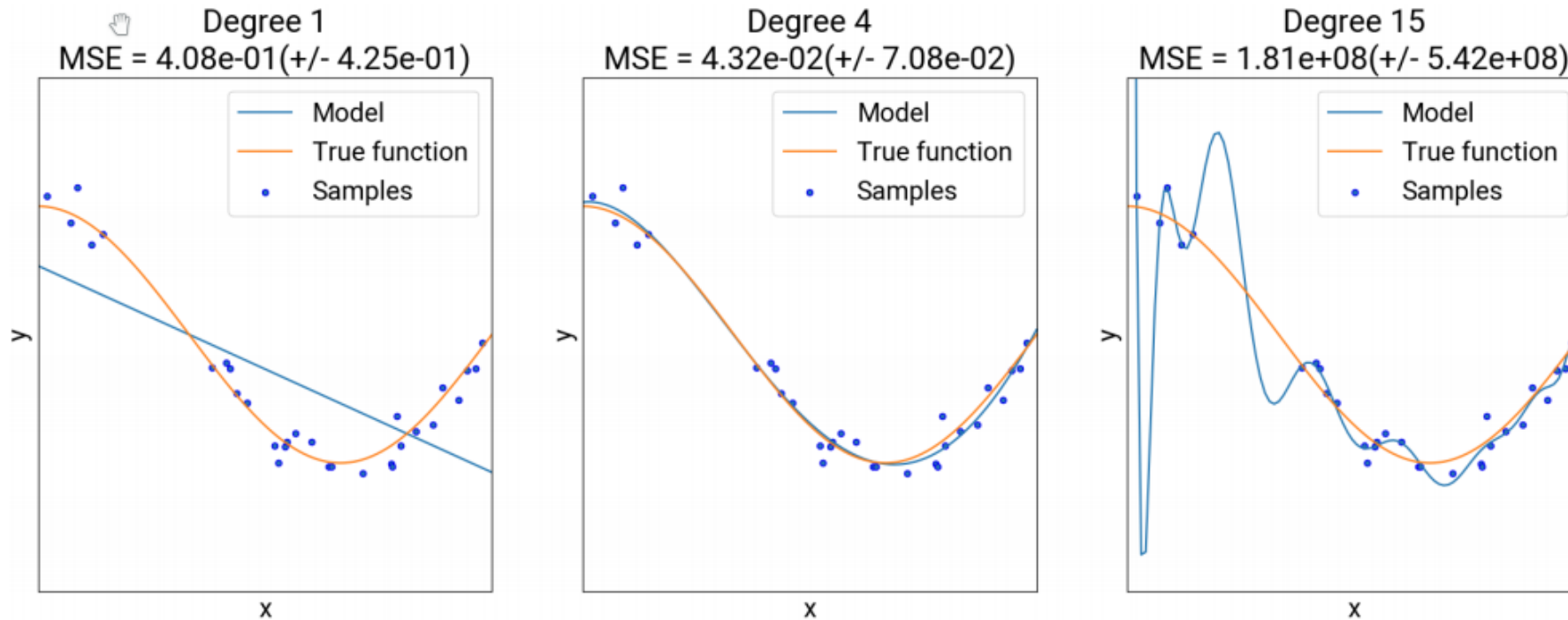
# initialize classifier
classifier = GaussianNB()

# use unigrams and bigrams as features
vectorizer = CountVectorizer(ngram_range=(1,2))
y = corpus[1]
X = vectorizer.fit_transform(corpus[0])
classifier.fit(X,y) # train classifier

#transform new texts into feature vectors
unseen_texts = ["list of unseen texts"]
feature_vectors = vectorizer.transform(unseen_texts)
answers = classifier.predict(feature_vectors)
```

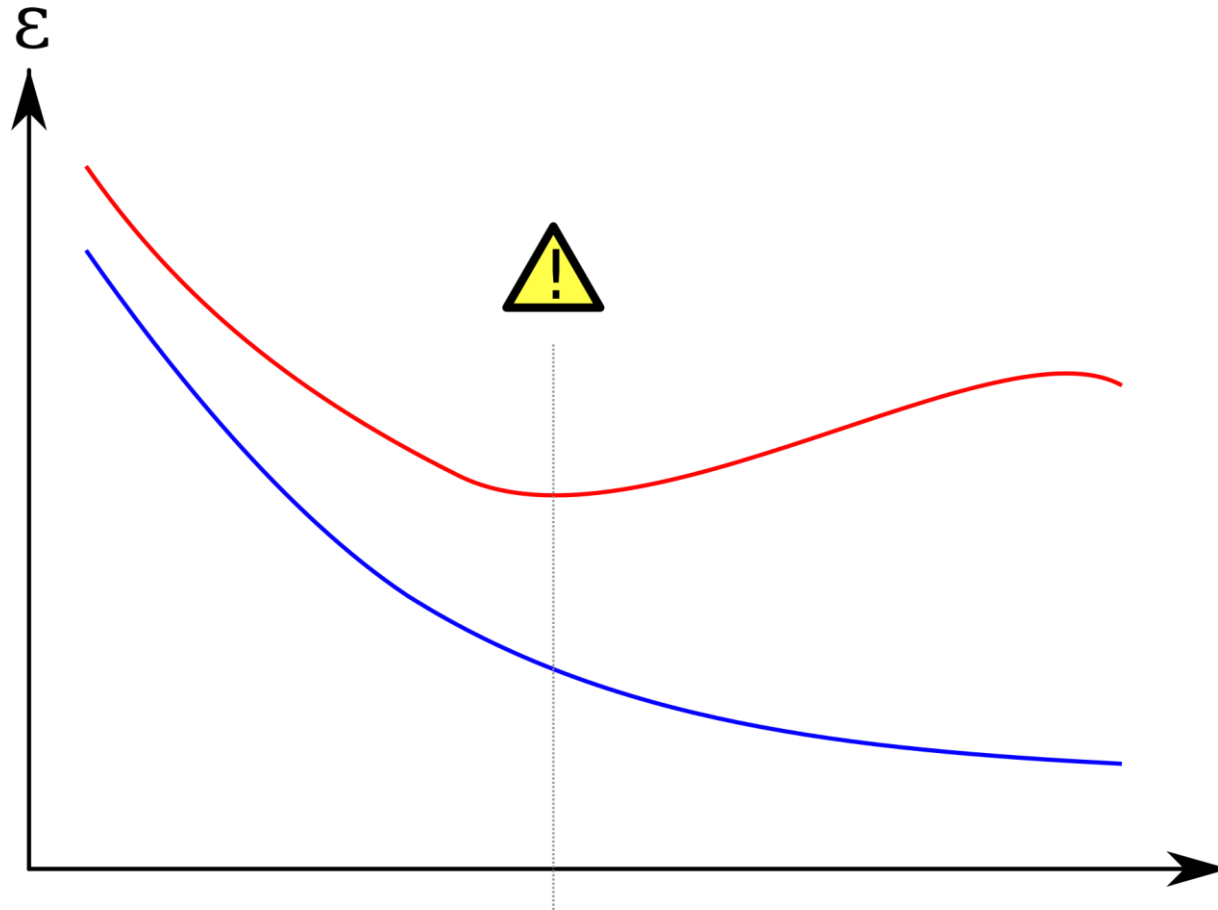
Переобучение

- А что если аппроксимировать данные полиномом с высокой степенью (в качестве признаков брать функции высоких порядков)?



- При $d=15$ модель получилась слишком сложной и обучилась на шуме

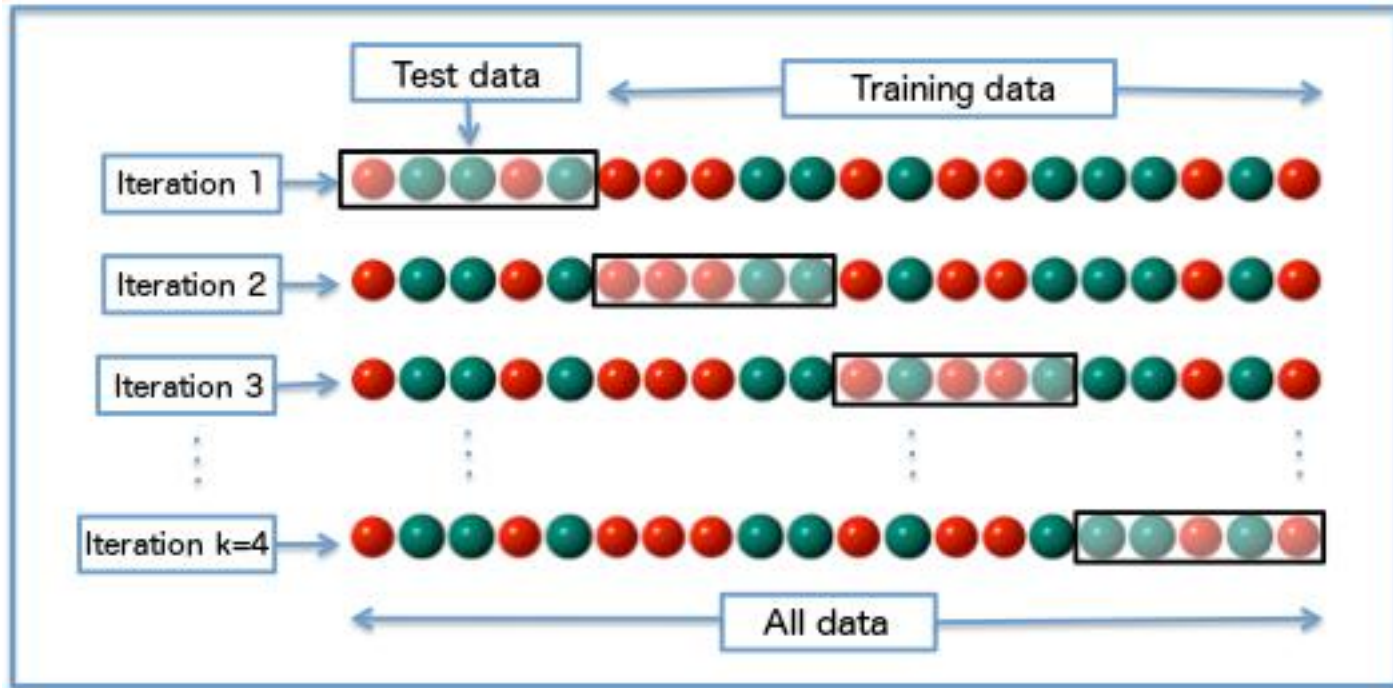
Переобучение



- (Синий) Ошибка на тренировочных данных
- (Красный) Ошибка на валидационных данных

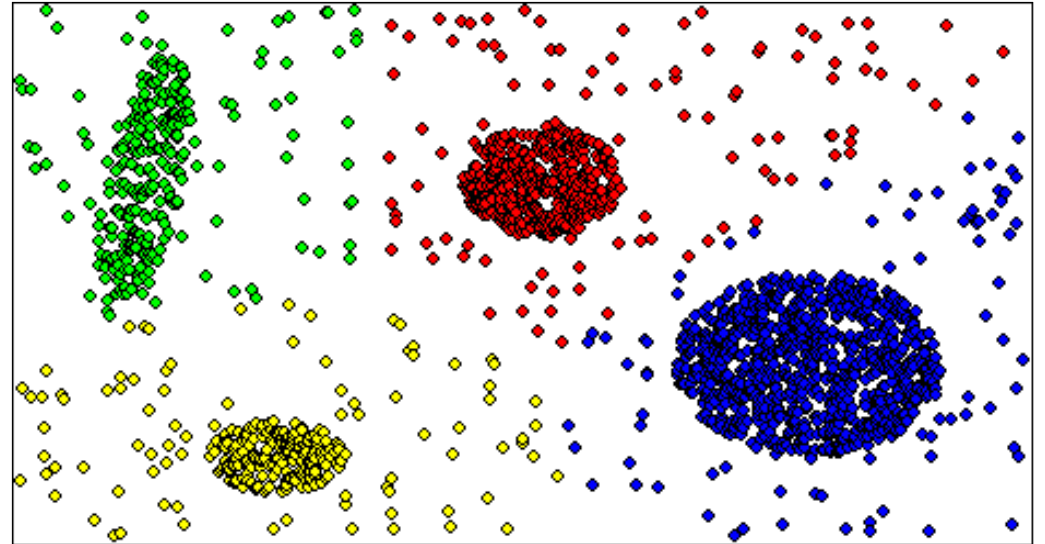
Проведение экспериментов

- Данные делятся на несколько частей
 - Тренировочная
 - Тестовая
 - Валидационная
- Перекрестная проверка (cross-validation)



Кластеризация

- Входные элементы можно разбить на несколько групп, по принципу схожести



Вход для алгоритмов

- Пусть каждый объект $\{x_1, x_2, \dots, x_k\}$ представлен вектором $x_i = (f_{i_1}, \dots, f_{i_n})$ в пространстве $X \subseteq R^n$
- Задается расстояние между векторами
 - Евклидово $d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_n - q_n)^2} = \sqrt{\sum_{k=1}^n (p_k - q_k)^2}$
 - Чебышева $l_\infty(\vec{x}, \vec{y}) = \max_{i=1, \dots, n} |x_i - y_i|$
 - Хэмминга $d_{ij} = \sum_{k=1}^p |x_{ik} - x_{jk}|$
 - Минковского $\rho(x, y) = \left(\sum_{i=1}^n |x_i - y_i|^p \right)^{1/p}$
 - ...

Алгоритм К-средних (k-means)

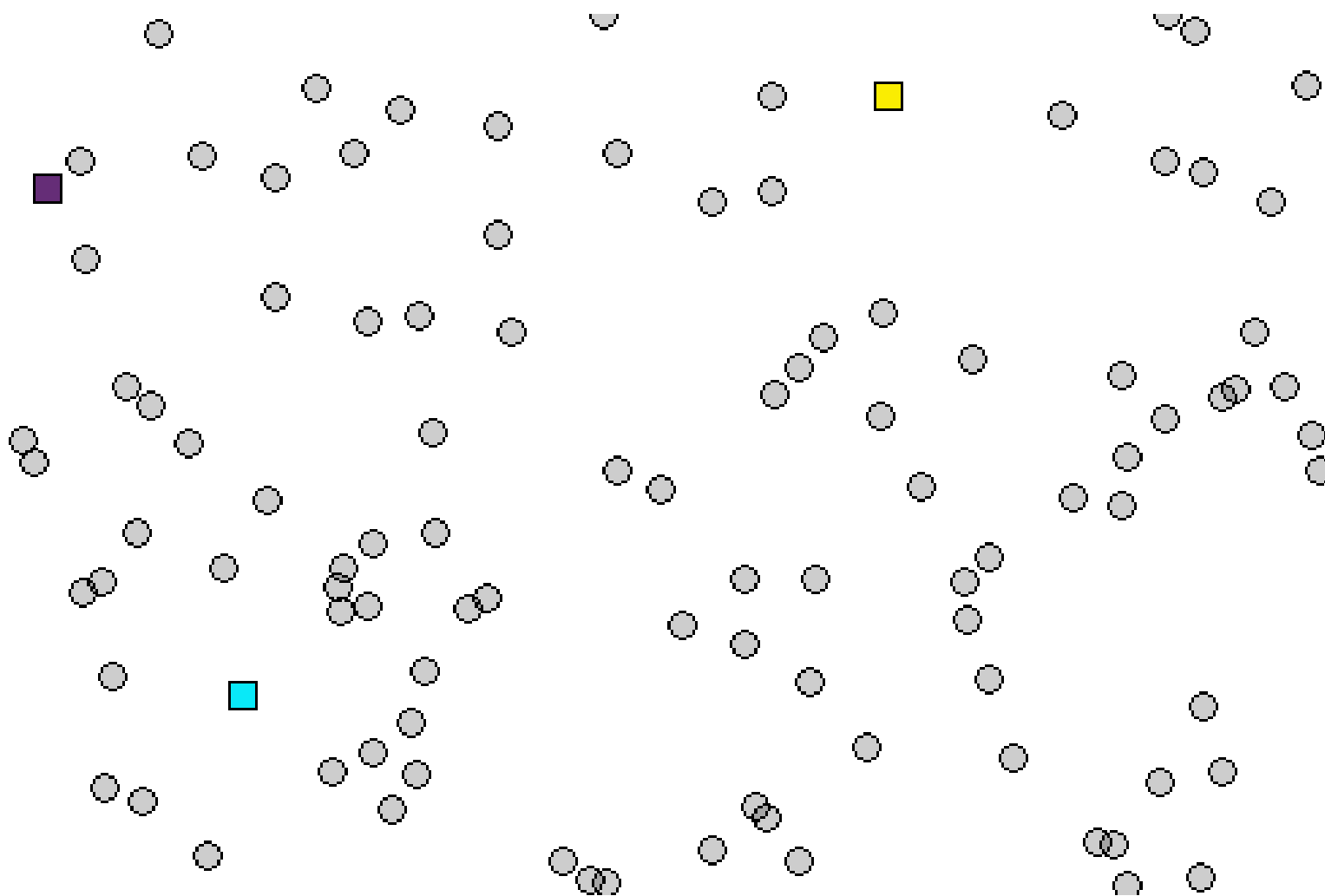
- Алгоритм k-means разбивает данные на k кластеров
 - Каждый кластер имеет центр - центроид
 - Параметр k - задается вручную
- Алгоритм
 1. Выбираются k точек в качестве начальных центроидов
 2. Сопоставить каждой точке ближайший центроид
 3. Пересчитать центроиды
 4. Если алгоритм не сошелся перейти на шаг 2

Критерий останова

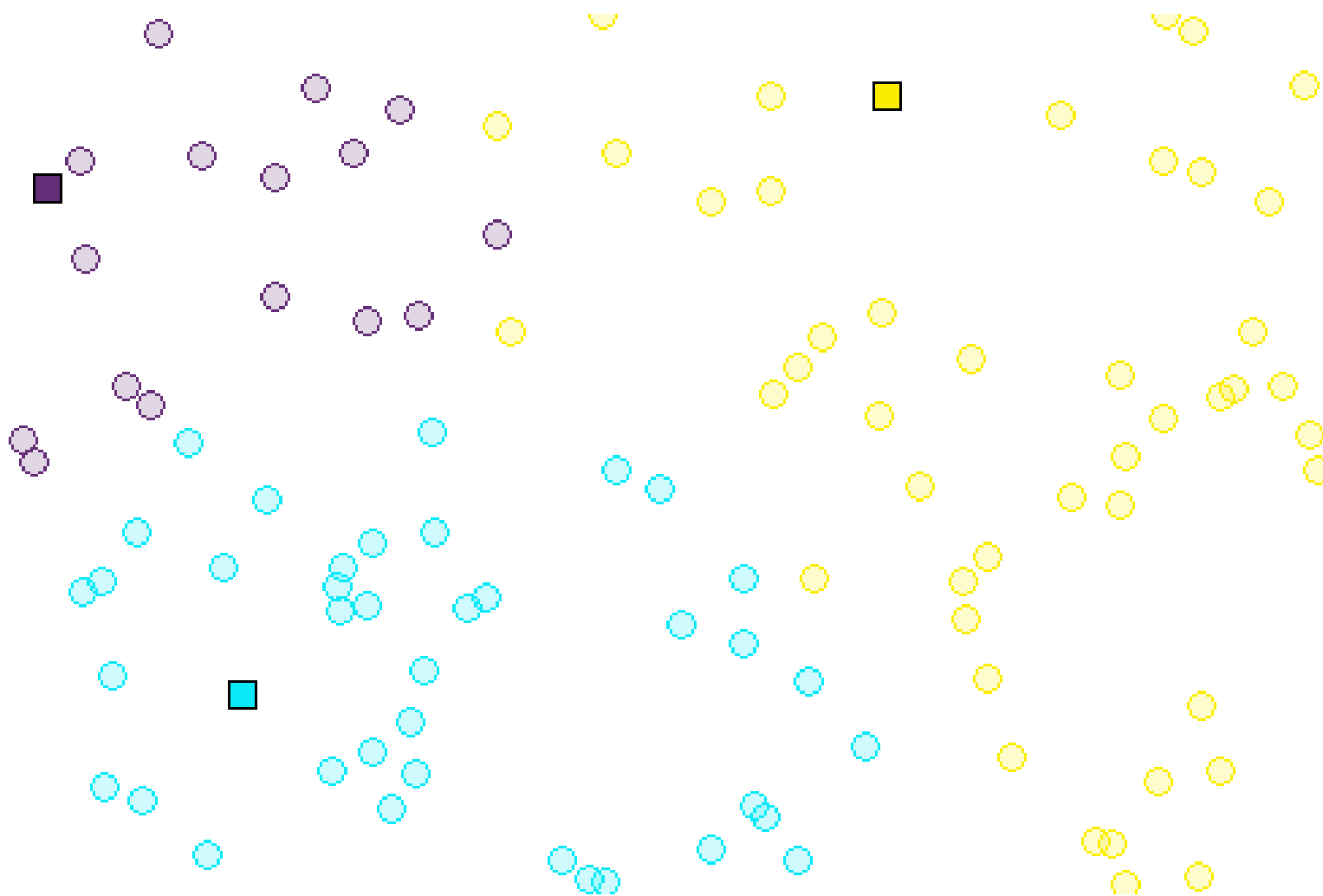
- Нет перехода точек в другой кластер
- Нет (незначительно) изменение центроидов
- Мало убывает погрешность (sum of squared error)

$$SSE = \sum_{j=1}^k \sum_{x \in C_j} dist(x, m_j)^2$$

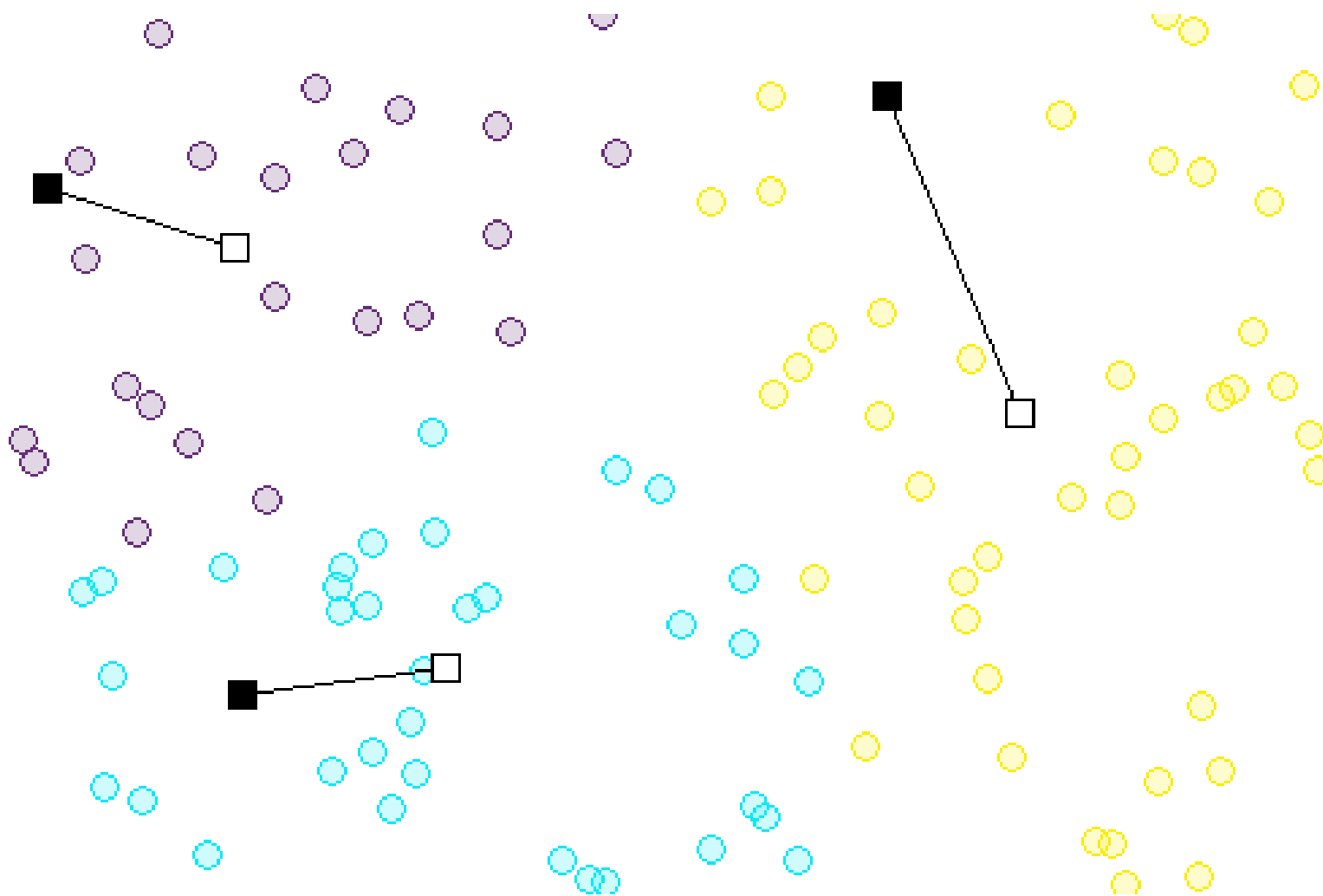
Алгоритм К-средних. Пример.



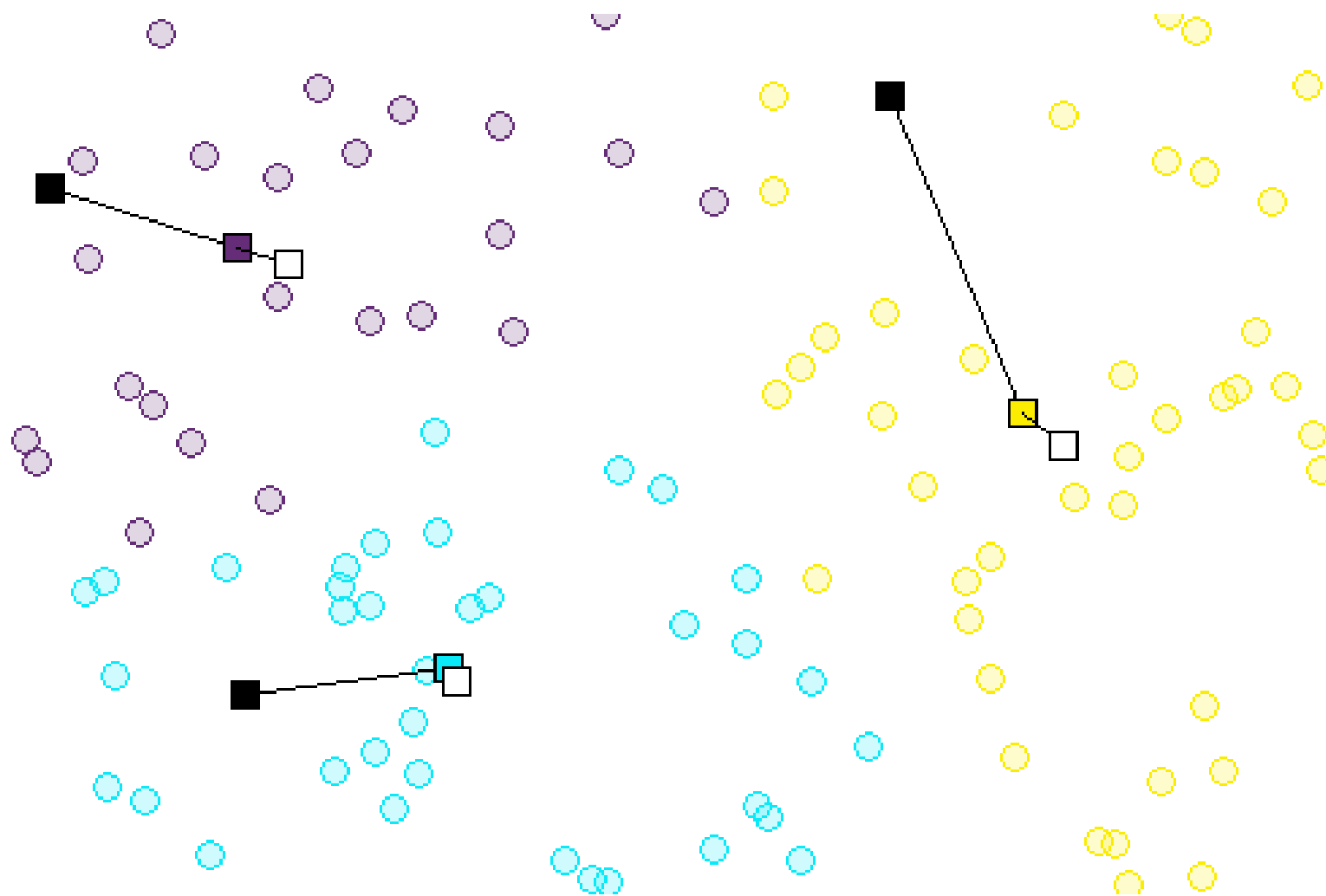
Алгоритм К-средних. Пример.



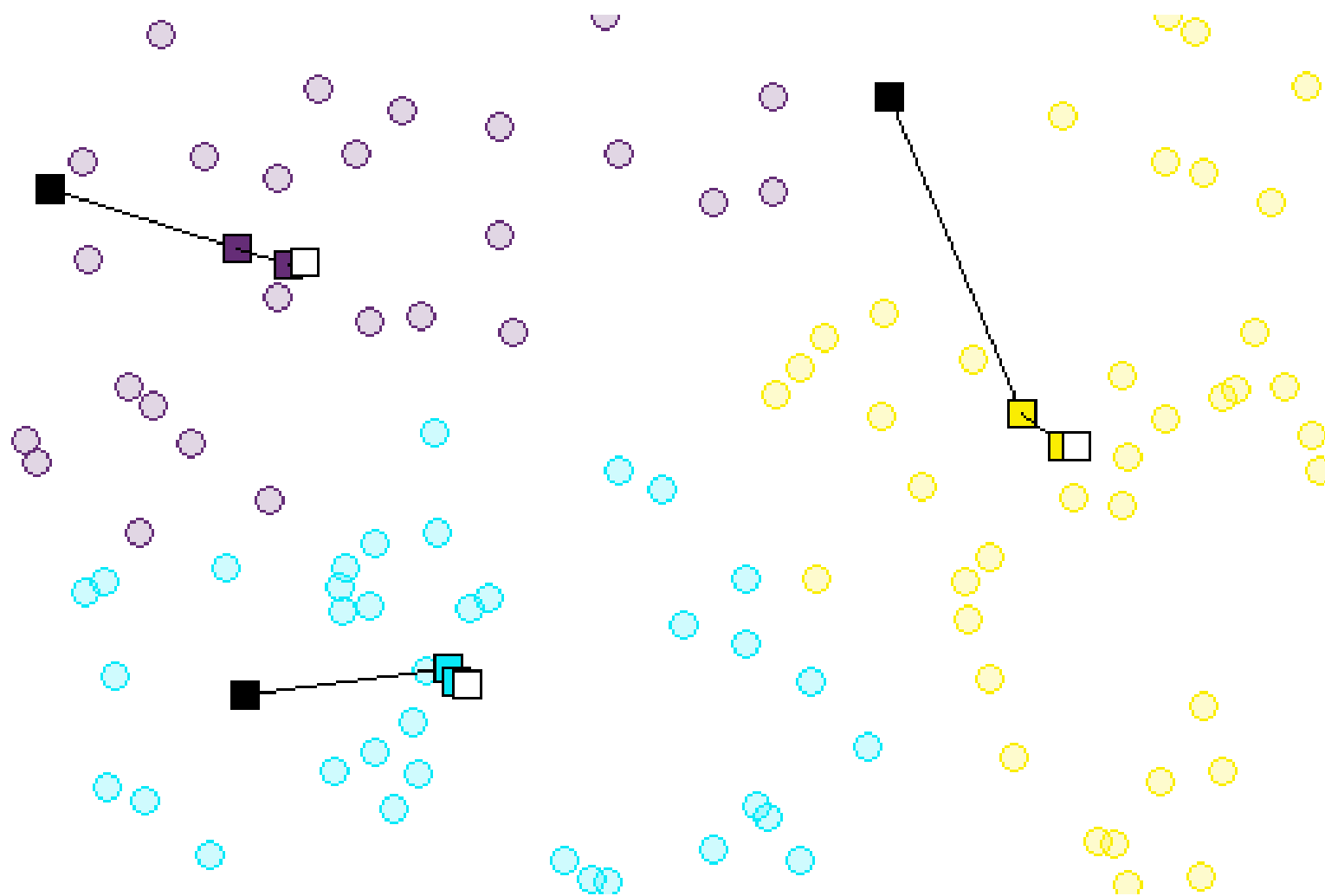
Алгоритм К-средних. Пример.



Алгоритм К-средних. Пример.



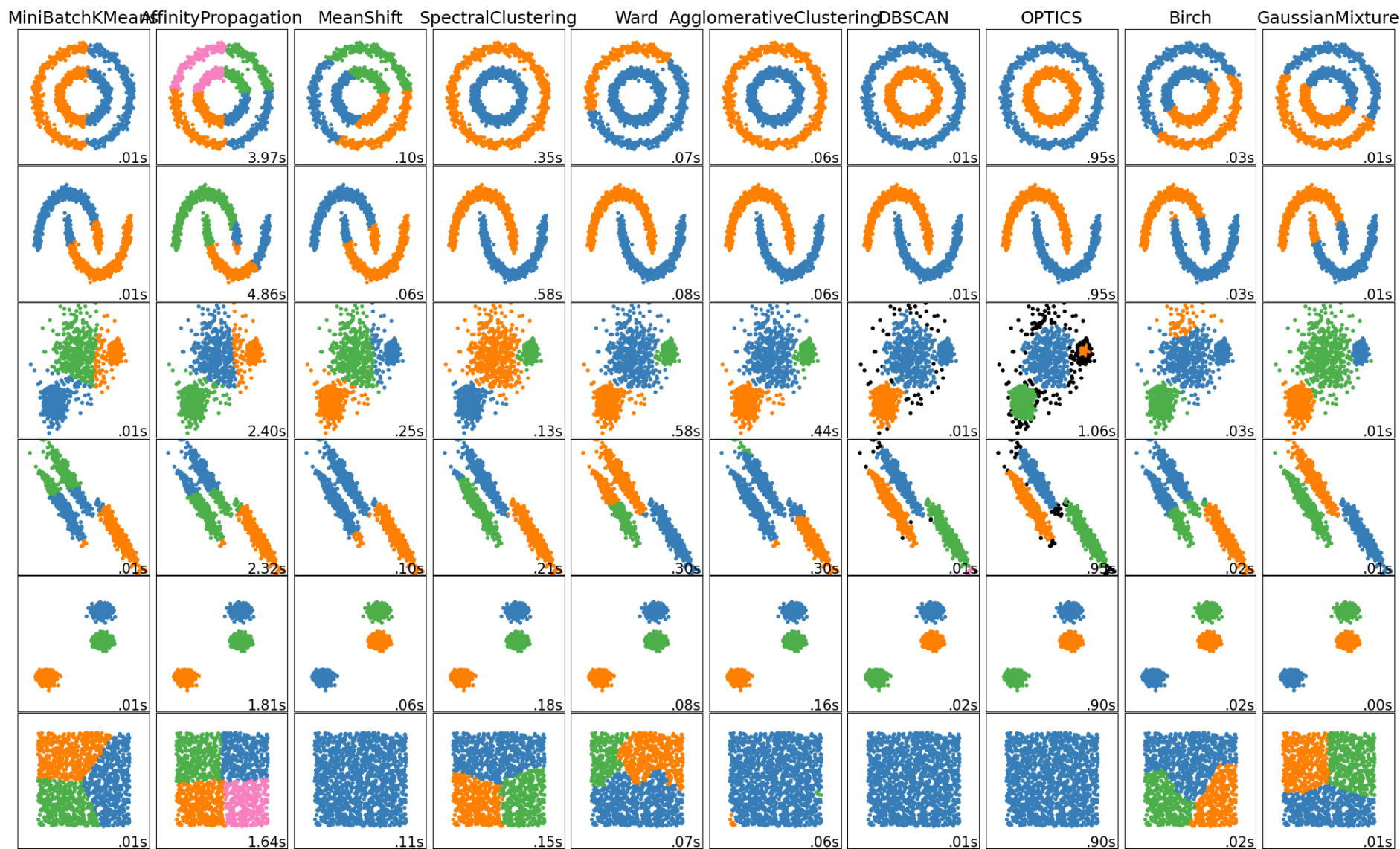
Алгоритм К-средних. Пример.



Проблемы K-means

- Алгоритм чувствителен к начальному выбору центроидов
 - запуск с различной начальной инициализацией и выбор варианта с наиболее плотными кластерами
- Чувствителен к выбросам
 - можно фильтровать выбросы
- Не подходит для нахождения кластеров, не являющихся эллипсоидами
 - преобразование пространства

Какой алгоритм кластеризации выбрать



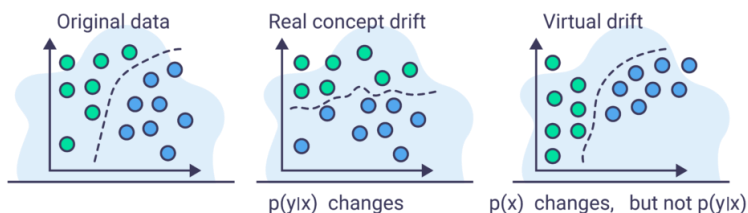
* <https://scikit-learn.org/stable/modules/clustering.html>

Часть 3. Современные проблемы доверенного искусственного интеллекта

Проблемы с доверием к машинному обучению

Проблемы разработки и эксплуатации

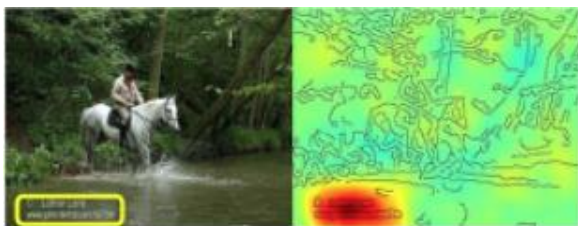
- Переобучение
- Дрейф данных



Предвзятость моделей

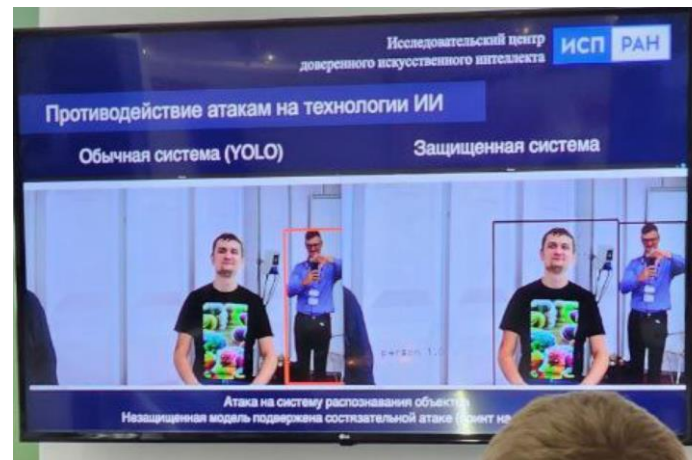


Интерпретация результатов



Безопасность

- Состязательные атаки для манипуляции поведением системы



- Встраивание закладок на этапе обучения



- Извлечение конфиденциальных данных из обученных моделей
- «Кража» самих моделей

Дефекты машинного обучения

	Информационная безопасность	Функциональная надежность	Социогуманитарные аспекты
Датасеты	- Утечка датасетов - Закладки и отравление данных	- Несогласованность данных - Зашумленность данных - Неполнота данных (приводят к ухудшению качества и некорректному функционированию моделей)	- Датасеты противоречат культурным, этическим и правовым нормам - Незаконное или неэтичное использование авторских данных
Модели	- Угроза утечки чувствительных данных из модели - Угроза несанкционированного доступа через дефекты машинного обучения - Угроза внедрения кода или данных	- Дрейф данных - Некорректная обработка выбросов - Некорректная интерпретация выводов модели - Галлюцинации больших моделей ИИ	- DeepFake - Генерация неэтичного или деструктивного контента

Тактики атак на дефекты ИИ

Reconnaissance&	Resource Development&	Initial Access&	ML Model Access	Execution&	Persistence&	Privilege Escalation&	Defense Evasion&	Credential Access&	Discovery&	Collection&	ML Attack Staging	Exfiltration&	Impact&
5 techniques	9 techniques	6 techniques	4 techniques	3 techniques	4 techniques	3 techniques	3 techniques	1 technique	6 techniques	3 techniques	4 techniques	4 techniques	7 techniques
Search for Victim's Publicly Available Research Materials	Acquire Public ML Artifacts	ML Supply Chain Compromise	AI Model Inference API Access	User Execution &	Poison Training Data	LLM Prompt Injection	Evade ML Model	Unsecured Credentials &	Discover ML Model Ontology	ML Artifact Collection	Create Proxy ML Model	Exfiltration via ML Inference API	Evade ML Model
Search for Publicly Available Adversarial Vulnerability Analysis	Obtain Capabilities &	Valid Accounts &	ML-Enabled Product or Service	Command and Scripting Interpreter &	Backdoor ML Model	LLM Plugin Compromise	LLM Prompt Injection		Discover ML Model Family	Data from Information Repositories &	Backdoor ML Model	Exfiltration via Cyber Means	Denial of ML Service
Search Victim-Owned Websites	Develop Capabilities &	Evade ML Model	Physical Environment Access	LLM Plugin Compromise	LLM Prompt Injection	LLM Jailbreak	LLM Jailbreak		Discover ML Artifacts	Data from Local System &	Verify Attack	LLM Meta Prompt Extraction	Spamming ML System with Chaff Data
Search Application Repositories	Acquire Infrastructure	Exploit Public-Facing Application &	Full ML Model Access		LLM Prompt Self-Replication				LLM Meta Prompt Extraction		Craft Adversarial Data	LLM Data Leakage	Erode ML Model Integrity
Active Scanning &	Publish Poisoned Datasets	LLM Prompt Injection							Discover LLM Hallucinations				Cost Harvesting
	Poison Training Data	Phishing &							Discover AI Model Outputs				External Harms
	Establish Accounts &												Erode Dataset Integrity
	Publish Poisoned Models												
	Publish Hallucinated Entities												

MITRE ATLAS™

Тактики атак появились в 2023 году

Тактики атак появились в 2024 году

Комплексное понимание доверия к ИИ

	Информационная безопасность	Функциональная надежность	Социогуманитарные аспекты
Датасеты	Методы обеспечения информационной безопасности обучающих данных - Федеративное и распределенное обучение - Дифференциальная приватность - Защита от закладок и отравления данных	Методы создания высококачественных наборов данных - Подходы к созданию согласованных датасетов - Методы очистки датасетов - Методы заполнения пропущенных данных - Аугментация данных	Методы противодействия угрозам, возникающим при использовании фундаментальных и генеративных моделей в специфичных предметных областях - Инструменты создания обучающих данных и бенчмарков для генеративного ИИ на основе доверенных источников - Бенчмарки соответствия больших моделей ИИ культурным и правовым нормам - Watermarking контента - Инструменты выявления деструктивного контента, сгенерированного ИИ
Модели (обучение)	Методы реализации и противодействия реализации угроз, специфичных для генеративного ИИ - Удаление незначимых частей (сжатие) моделей - Методы обучения моделей, повышающие устойчивость к атакам - Разучивание примеров - Гомоморфное шифрование при обучении	Методы обеспечения функциональной надежности моделей машинного обучения в условиях неопределенности - Развитие принципов обучения и оптимизации - Интерпретируемость моделей - Методы оценки неопределенности - Физически-информированные нейронные сети (science-informed AI)	
Модели (эксплуатация)	Методы обеспечения доверия к интеллектуальным системам на этапе эксплуатации - Встраивание водяных знаков в модели и данные для защиты от кражи - Повышение устойчивости мультимодальных моделей к атакам через инъекции промптов - Бенчмарки безопасности - Сертифицированная робастность	Методы мониторинга и непрерывной адаптации моделей, функционирующих в нестационарной среде - Обнаружение и обработка дрейфа данных - Обнаружение и фильтрация выбросов (out-of-distribution) - Построение мультиагентных систем - Методы переноса знаний (в т.ч. для малоресурсных задач) - Гибридный искусственный интеллект	

План лекций (предварительный)

№	Дата	Название	Лектор
1	17.09	Задачи машинного обучения	Турдаков
2	24.09	Введение в нейронные сети	Перминов
3	01.10	Математика полносвязной сети: геометрия, интерпретируемость, SLAP атака	Перминов
4	08.10	Атаки на модели машинного обучения и методы защиты	Чистякова
5	15.10	Задачи обработки изображений. Атаки на методы обработки изображений	Анциферова
6	22.10	Генерация изображений. Диффузионные модели: атаки и защита	Матюшин
7	29.10	Доказательства в машинном обучении. Новые результаты и открытые проблемы	Паутов
8	05.11	Анализ временных рядов. Медицинские приложения, ЭКГ. Устойчивость.	Аветисян
9	12.11	GAN. Создание и детекция дипфейков	Анциферова
10	19.11	Маркировка данных. Борьба с дипфейками	Маркин и КО
11	26.11	Дрейф данных. OOD. Инструменты MLOps.	Рындин
12	03.12	Обучение LLM. RuAdapt	Тихомиров
13	10.12	Генерация кода	Сорокин
14	17.12	RAG	Турдаков

Полезные ссылки

- <https://education.at.ispras.ru> – как попасть на СП / в ИСП РАН
- MOOC Курс от SRR на Stepik
 - <https://stepik.org/course/50352> - введение в нейронные сети
 - <https://stepik.org/course/54098>
- <http://www.machinelearning.ru>
 - [http://www.machinelearning.ru/wiki/index.php?title=Машинное_обучение_\(курс_лекций,_К.В.Воронцов\)](http://www.machinelearning.ru/wiki/index.php?title=Машинное_обучение_(курс_лекций,_К.В.Воронцов))
- <https://github.com/dformoso/machine-learning-mindmap/blob/master/Machine%20Learning.pdf>
- Лекции 2021
 - https://www.youtube.com/watch?v=8sRF59OX7_I&list=PL5cBzMoPJgCUr-UCFNIWh5jWQ6ewn35rP

Следующая лекция

- Введение в нейронные сети