

Перечень и содержание тем

Тема 1. Трассировка и анализ алгоритмов

Пошаговое выполнение алгоритмов и псевдокода. Значения переменных на каждой итерации. Условные операторы и циклы. Целочисленное деление div . Проверка чётности. Накопление результата. Алгоритмы последовательного уменьшения и деления диапазона пополам. Округление вверх и вниз. Связь числа шагов с двоичным делением и логарифмическим ростом. Проверка граничных случаев.

Тема 2. Формализация задач и дискретные модели

Перевод текстового условия в математическую модель. Равномерное распределение целых объектов по группам. Частное и остаток от деления. Оператор mod . Минимизация разницы между размерами групп. Интерпретация остатка как числа групп, получающих дополнительный объект. Работа с целочисленными ограничениями и построение простых дискретных моделей по условию задачи.

Тема 3. Идентификация модели по наблюдениям

Набор возможных правил как пространство гипотез. Проверка согласованности правила с наблюдаемыми ответами. Классификация точек линейными неравенствами. Выбор дополнительных запросов, различающих оставшиеся правила. Двоичный результат запроса и информационная оценка числа необходимых наблюдений. Сигнатуры гипотез по ответам. Минимальное число экспериментов, достаточное для однозначного различения кандидатов.

Тема 4. Описательные статистики и анализ данных

Чтение небольших таблиц наблюдений. Вычисление производных величин из исходных данных, в том числе интервалов времени. Среднее арифметическое и интерпретация среднего значения. Перевод временных величин в единые единицы. Контроль размерности, сумм и количества наблюдений. Анализ простых журналов событий и табличных данных.

Тема 5. Математическое ожидание и независимость

Случайные события и независимость. Вероятность совместного наступления независимых событий. Математическое ожидание числа совпадений и других счётных величин. Индикаторы событий и линейность математического ожидания. Работа с фиксированным числом успехов в серии испытаний. Разложение события на взаимоисключающие случаи и получение ожидаемого результата.

Тема 6. Условная вероятность и формула Байеса

Условная вероятность. Априорная и апостериорная вероятность. Чувствительность системы и вероятность ложной тревоги. Матрица исходов для редкого и массового классов. Эффект низкой базовой частоты: большое число ложных срабатываний при редком событии. Формула Байеса через подсчёт случаев. Условные вероятности в процессах с остановкой при выполнении заданного условия.

Тема 7. Метрики расстояния и близость объектов

Числовое представление объектов признаками. Точки в пространстве признаков. Евклидово расстояние на плоскости. Сравнение расстояний до нескольких объектов и поиск ближайшего. Квадраты расстояний как удобный способ сравнения без лишнего извлечения корня. Округление численного результата и геометрическая интерпретация близости объектов.

Тема 8. Нормализация признаков

Признаки с различными единицами измерения и масштабами. Влияние крупномасштабного признака на евклидово расстояние. Min-max нормализация по формуле $(x - \min)/(\max - \min)$. Приведение признаков к диапазону [0; 1]. Пересчёт расстояний после преобразования. Интерпретация того, как масштабирование меняет вклад отдельных признаков.

Тема 9. Линейные модели и разделяющая граница

Линейная функция вида $z = w_1x_1 + w_2x_2 + b$. Классификация по знаку или пороговому значению функции. Разделяющая граница как уравнение $z = 0$. Определение координаты точки на границе. Геометрический смысл линейного классификатора и влияние коэффициентов модели на положение разделяющей прямой.

Тема 10. Нейронные сети: прямой проход

Простейшая нейронная сеть с входными, скрытыми и выходными значениями. Последовательное вычисление значений нейронов. Линейные комбинации входов. Функция активации ReLU: $\max(0; t)$. Обнуление отрицательных значений. Вычисление выхода сети по значениям скрытого слоя. Прямой проход без обучения и обратного распространения ошибки.

Тема 11. Функции ошибки и градиентный спуск

Линейный прогноз и параметр модели. Квадратичная функция ошибки как сумма квадратов отклонений. Сравнение качества нескольких значений параметра. Производная функции ошибки. Градиент и направление уменьшения ошибки. Правило обновления параметра $w' = w - \alpha E'(w)$. Шаг обучения, движение к минимуму и проверка результата после одной итерации.

Тема 12. Метод ближайших соседей

Классификация методом k ближайших соседей. Вычисление расстояний от нового объекта до обучающих примеров. Сортировка соседей по близости. Выбор k ближайших объектов и решение по большинству классов. Использование нечётного k для уменьшения числа ничьих. Анализ того, как изменение k может изменить итоговый класс объекта.

Тема 13. Деревья решений и качество разбиения

Дерево решений и бинарные вопросы к признакам. Пороговые условия и проверка категориального признака. Разделение объектов на группы «да» и «нет». Оценка чистоты разбиения через число объектов класса-меньшинства. Сравнение нескольких кандидатов на первый вопрос дерева и выбор разбиения с наименьшей ошибкой. Подсчёт состава классов в дочерних узлах.

Тема 14. Переобучение и обобщающая способность

Обучающая и проверочная выборки. Ошибка на обучающих данных и ошибка на новых данных. Сложность модели и выбор варианта по проверочному качеству. Разрыв между обучающей и проверочной ошибкой. Переобучение как запоминание случайных особенностей и шума. Обобщающая способность модели и признаки чрезмерной сложности.

Тема 15. Метрики классификации, порог и цена ошибки

Вероятностный выход классификатора и порог принятия решения. Истинно положительные результаты, ложные тревоги и пропуски. Доля найденных объектов целевого класса. Отбор top-k объектов с наибольшей оценкой. Изменение качества при разных порогах. Разная стоимость ошибок первого и второго типа. Выбор порога по минимальным ожидаемым или фактическим потерям.

Тема 16. Пространство гипотез и активное обучение

Множество гипотез, согласующихся с наблюдаемыми данными. Исключение моделей, противоречащих хотя бы одному примеру. Выбор нового объекта для запроса метки. Информативные и неинформативные запросы. Сокращение пространства версий после получения ответа. Активное обучение как выбор наиболее полезного следующего наблюдения. Ограниченность вывода: единственная оставшаяся гипотеза среди кандидатов не доказывает истинность механизма вне выбранного класса моделей.

Тема 17. Валидация и утечки данных

Разделение данных на обучение и тест с учётом структуры наблюдений. Групповое разбиение, когда связанные примеры одного объекта не должны попадать одновременно в обучение и тест. Утечка данных и завышенная оценка качества. Сравнение случайного и честного разбиения. Базовый уровень качества для несбалансированных классов. Оценка модели на действительно новых объектах.

Тема 18. Надёжность и ограничения моделей

Проверка поведения модели на данных, отличающихся от типовых обучающих примеров. Устойчивость к новым формулировкам и изменению распределения. Поверхностные признаки как источник хрупкости модели. Ошибки разметки и систематическое смещение решений. Риски полностью автоматического принятия решений. Настройка более строгого порога для критичных ошибок и передача сомнительных случаев человеку. Интерпретация качества с учётом реального сценария применения.

Рекомендуемая литература

Основная литература

1. Deisenroth M. P., Faisal A. A., Ong C. S. Mathematics for Machine Learning. Свободно доступная электронная версия: <https://mml-book.github.io/book/mml-book.pdf>
2. James G., Witten D., Hastie T., Tibshirani R., Taylor J. An Introduction to Statistical Learning. Версии с R и Python и открытые PDF: <https://www.statlearning.com/>
3. Çetinkaya-Rundel M., Hardin J. Introduction to Modern Statistics. Открытый учебник и PDF: <https://www.openintro.org/book/ims/>
4. OpenIntro. Advanced High School Statistics. Учебник по вероятности и статистике для старших классов. Открытый PDF: <https://www.openintro.org/book/ahss/>
5. Ng A., Ma T. CS229 Lecture Notes: Machine Learning. Открытый PDF Stanford: https://cs229.stanford.edu/main_notes.pdf

Дополнительная литература

1. Goodfellow I., Bengio Y., Courville A. Deep Learning. Свободная онлайн-версия книги: <https://www.deeplearningbook.org/>
2. Google. Machine Learning Crash Course. Интерактивный курс по основам машинного обучения: <https://developers.google.com/machine-learning/crash-course/>
3. Scikit-learn. User Guide. Материалы по линейным моделям, ближайшим соседям, деревьям, предобработке, валидации и метрикам: https://scikit-learn.org/stable/user_guide.html

Электронные ресурсы

1. Яндекс Образование — Учебник по машинному обучению: <https://education.yandex.ru/handbook/ml>
2. OpenIntro — открытые учебники по вероятности и статистике: <https://www.openintro.org/>
3. Электронный архив УрФУ: <https://elar.urfu.ru/>