Тема 6. Методы анализа и интеллектуальной обработки больших данных

Лекция 6.1. Data Mining: задачи и методология

Data Mining — процесс автоматического извлечения неочевидных, ранее неизвестных, потенциально полезных знаний из больших массивов данных. Является центральным этапом более широкого процесса KDD (Knowledge Discovery in Databases): отбор данных — предобработка — трансформация — Data Mining — интерпретация результатов.

— Классификация: построение модели, относящей новый объект к одному из заранее определённых классов. Обучение с учителем (supervised). Примеры: диагностика заболевания по симптомам, оценка кредитоспособности клиента, фильтрация спама.

— Регрессия: предсказание непрерывного числового значения. Примеры: прогноз цены квартиры по характеристикам, прогноз спроса на следующий месяц.

— Кластеризация: разбиение объектов на однородные группы без заранее известных меток (обучение без учителя). Примеры: сегментация клиентов, группировка новостей по темам.

— Обнаружение аномалий: выявление объектов, статистически значимо отличающихся от общей массы. Примеры: мошеннические транзакции, производственные дефекты, кибератаки.

— Ассоциативные правила: поиск устойчивых закономерностей совместного появления элементов. Пример: анализ рыночной корзины (market basket analysis).

— Уменьшение размерности: сжатие пространства признаков с минимальной потерей информации. Позволяет визуализировать высокоразмерные данные и ускоряет последующее обучение.

Лекция 6.2. Статистический анализ в контексте больших данных

Описательная статистика

Первый шаг в анализе любого набора данных — вычисление базовых описательных статистик для понимания его структуры. Среднее значение (mean) — сумма делённая на количество, чувствительно к выбросам. Медиана — серединное значение, устойчива к выбросам. Мода — наиболее часто встречающееся значение, актуальна для категориальных данных. Стандартное отклонение характеризует разброс данных. Асимметрия (skewness) показывает, насколько распределение отличается от симметричного. Эксцесс (kurtosis) — «тяжесть» хвостов распределения по сравнению с нормальным.

Особенности статистики на больших данных

При объёме выборки в миллиарды записей любое, даже ничтожное, различие между группами даёт p-value < 0,001. Традиционная статистическая значимость теряет смысл: статистически значимое различие может быть практически незначимым. Поэтому для больших данных на первый план выходит практическая значимость — Cohen's d, eta-squared, размер эффекта. Вопрос не «есть ли различие?» (почти всегда есть), а «насколько велико различие и имеет ли оно практический смысл?».

A/B тестирование

A/B тест — рандомизированный эксперимент для сравнения двух вариантов продукта или решения. Пользователи случайно распределяются в контрольную (A) и экспериментальную (B) группы. Измеряется целевая метрика (CTR, конверсия, выручка на пользователя, время на сайте). Для корректности важны: случайность распределения, одновременность проведения, отсутствие взаимовлияния групп (network effects), достаточный размер выборки (статистическая мощность). Google, Netflix и Amazon одновременно проводят сотни A/B тестов.

Лекция 6.3. Кластерный анализ

K-Means: алгоритм и практика

K-Means — итерационный алгоритм разбиения на K кластеров путём поочерёдной оптимизации назначений и центроидов. Шаги: 1) инициализировать K центроидов (K-Means++ выбирает их наиболее разнесёнными); 2) назначить каждую точку ближайшему центроиду; 3) пересчитать центроиды как среднее по кластеру; повторять до изменения назначений менее чем на порог. Сходится быстро, масштабируется на миллиарды точек в Spark MLlib.

— Выбор числа кластеров K: метод локтя — строить график суммарного внутрикластерного расстояния от K, искать «локоть» кривой. Силуэтный коэффициент: от -1 до 1, оценивает насколько хорошо точка находится в своём кластере относительно соседних.

— Ограничения K-Means: предполагает сферические кластеры равного размера и плотности, чувствителен к выбросам (они смещают центроид), плохо работает с кластерами неправильной формы.

— Bisecting K-Means (доступен в Spark MLlib): иерархический вариант — начинает с одного кластера и рекурсивно делит наибольший, пока не достигнет K кластеров.

DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) не требует задавать K заранее. Кластер определяется как плотная область — множество точек, в чьей

eps-окрестности не менее minPts соседей. Точки, не входящие ни в один кластер, считаются шумом. Находит кластеры произвольной формы, автоматически определяет количество кластеров, устойчив к выбросам. Недостатки: сложно подобрать eps и minPts, плохо масштабируется на очень большие данные без модификаций.

Иерархическая кластеризация

Строит дерево кластеров (дендрограмму). Агломеративный подход: каждая точка — отдельный кластер, затем два ближайших кластера объединяются на каждом шаге. Число кластеров выбирается постфактум — «срезом» дендрограммы на нужном уровне. Не требует задавать K, но сложность O(n^2 log n) делает метод неприменимым для миллионов точек.

Лекция 6.4. Ассоциативные правила

Ассоциативные правила — метод поиска устойчивых совместных появлений наборов элементов в транзакционных данных. Классическое применение — анализ рыночной корзины: «покупатели, взявшие хлеб, в 65% случаев берут масло». Используется в рекомендательных системах, анализе кликов, медицинской диагностике.

Метрики качества правил

— Support (поддержка) = freq(A,B) / N. Доля транзакций, содержащих и A, и B. Фильтрует редкие, статистически ненадёжные правила.

— Confidence (достоверность) = freq(A,B) / freq(A). Вероятность B при условии A. Правило A -> B с confidence 0,8 означает: при покупке A покупатель в 80% случаев берёт B.

— Lift = Confidence(A->B) / Support(B). Lift > 1: A и B встречаются вместе чаще случайного; = 1: независимы; < 1: реже случайного. Lift нивелирует влияние популярности B.

Алгоритмы

Apriori использует принцип антимонотонности: если набор редкий, все его надмножества тоже редкие — их можно отсечь. Требует многократных проходов по базе данных. FP-Growth строит компактное дерево FP-Tree за один проход, затем извлекает частые наборы без генерации кандидатов. Значительно быстрее Apriori, особенно при низком минимальном support. Реализован в Spark MLlib как класс FPGrowth.

Лекция 6.5. NLP на больших данных

Около 80% корпоративных данных — неструктурированный текст: документы, переписка, отзывы, публикации в социальных сетях. NLP (Natural Language Processing) открывает доступ к этому массиву для количественного анализа.

Стандартный NLP-пайплайн

— Токенизация: разбивка текста на токены. Уровни: символьный, субсловный (BPE, WordPiece), словный, фразовый.

— Нормализация: стемминг (отсечение суффиксов: «играл» -> «игр»; быстро, грубо) или лемматизация (словарная форма: «играл» -> «играть»; медленнее, точнее).

— Удаление стоп-слов: частые слова без смысловой нагрузки (предлоги, союзы). Сокращает размерность словаря.

— Векторизация: преобразование текста в числовой вектор для машинного обучения.

Методы векторизации

— TF-IDF: TF (частота слова в документе) умножается на IDF (обратная частота слова во всём корпусе). Снижает вес универсальных слов, повышает вес специфических. Стандарт для информационного поиска.

— Word2Vec: нейросеть обучает 100–300-мерные плотные векторы слов. Семантически близкие слова близки в векторном пространстве. «Король» - «Мужчина» + «Женщина» = «Королева». Реализован в Spark MLlib.

— GloVe: матричная факторизация глобальной матрицы совстречаемости слов. Захватывает как локальный, так и глобальный контекст.

— BERT и производные (RoBERTa, XLM-R): контекстные embeddings — вектор слова зависит от контекста. «Банк» в «речном банке» и «финансовом банке» имеют разные векторы. Требует GPU для обучения, возможно дообучение на специализированных корпусах.

Лекция 6.6. Анализ временных рядов

Временной ряд — упорядоченная по времени последовательность наблюдений одной или нескольких переменных. Огромная часть Big Data — временные ряды: показания промышленных датчиков, биржевые котировки, веб-трафик, транзакции, журналы мониторинга систем.

Декомпозиция временного ряда

— Тренд: долгосрочное направление изменения среднего уровня ряда.

— Сезонность: регулярные периодические колебания с фиксированным периодом (суточные, недельные, квартальные, годовые). В отличие от цикличности, период строго фиксирован.

— Циклические колебания: нерегулярные долгосрочные колебания, связанные с экономическими и другими циклами. Период нефиксирован.

— Остаток (шум): случайная компонента после удаления тренда и сезонности.

Методы прогнозирования

— ARIMA (AutoRegressive Integrated Moving Average): классический параметрический метод для стационарных рядов. Параметры p (AR), d (дифференцирование), q (MA) выбираются по ACF/PACF.

— SARIMA: расширение ARIMA с явным сезонным компонентом.

— Prophet (Meta): модель на основе суммы трендовой функции и сезонных компонент. Устойчива к пропускам данных и аномалиям, поддерживает учёт праздников и событий.

— LSTM: рекуррентная нейросеть с механизмом долгосрочной памяти. Хорошо моделирует сложные долгосрочные зависимости в нелинейных рядах.

— В Spark: применение моделей scikit-learn/statsmodels через Pandas UDF для параллельного прогнозирования тысяч рядов одновременно.