Конструирование признаков

Лекция 6

Что такое признак и зачем он нужен

- Признак — числовое или категориальное описание свойства объекта
- Модель оперирует не объектами, а их признаковыми представлениями
- Качество признаков определяет эффективность и интерпретируемость модели
- Хорошая система признаков позволяет выявлять закономерности в данных

Конструирование признаков определяет качество модели

- Признаки задают пространство, в котором обучается модель
- Ошибки в признаках сильнее влияют, чем ошибки в алгоритме
- Богатое и устойчивое описание данных повышает обобщаемость
- Сложные модели не компенсируют слабое признаковое представление

Пропуски и почему они появляются

- Пропуски — отсутствие значений в данных
- Источники: ошибки измерений, передачи, фильтрации, человеческий фактор
- По происхождению:
  - Случайные — вызваны внешними, неповторяющимися причинами
  - Систематические — закономерно связаны с объектом или значением
- Типы пропусков:·
  - MCAR (Missing Completely at Random) — полностью случайные·
  - MAR (Missing at Random) — случайные по условию·
  - MNAR (Missing Not at Random) — систематические, зависят от скрытого значения
- Игнорирование пропусков искажает распределения и закономерности

Стратегии обработки пропусков (вменение данных)

Цель вменения (imputation) — восстановить структуру данных без искажения статистики

- Методы зависят от природы пропусков (случайные / систематические)
- Основные подходы:
  - Удаление наблюдений — при малой доле пропусков и случайном характере
  - Статистическая подстановка — среднее, медиана, мода
  - Стохастическое заполнение — добавление случайного шума в диапазоне значений
  - Моделирование пропусков — прогнозирование значений по другим признакам
  - Добавление индикатора пропуска — фиксация факта отсутствия данных

Когда не стоит заполнять пропуски

- Вменение уместно не всегда — иногда пропуск несёт смысловую информацию
- Заполнение может исказить распределения и нарушить зависимости
- Не стоит заполнять:
  - если пропуск отражает реальное отсутствие признака
  - если доля пропусков слишком велика (>30–40%)
  - если механизм пропусков неслучаен (MNAR)
- Альтернатива: оставить пропуск как категорию, добавить индикатор, исключить признак

Категориальное кодирование

- Категориальные признаки нельзя напрямую подать в модель
- Кодирование переводит категории в числовое представление
- Основные методы:
  - One-Hot (унитарное) — бинарные индикаторы категорий
  - Ordinal (порядковое) — числовое ранжирование значений
  - Frequency (частотное) — подмена частотой появления
  - Target (целевое) — усреднение целевой переменной по категории
  - Weight of Evidence (WoE) — логарифм отношения вероятностей классов

Унитарное (One-Hot) кодирование

- Каждая категория преобразуется в отдельный бинарный признак
- Значение 1 — категория активна, 0 — отсутствует
- Метод не предполагает порядка между категориями
- Преимущества: простота, отсутствие ложных рангов
- Недостатки: рост размерности, разреженность, потеря статистической связи

Порядковое кодирование (Ordinal encoding)

- Используется, когда категории имеют естественный порядок
- Каждой категории присваивается числовой ранг
- Сохраняет порядок, но не гарантирует равенство расстояний
- Преимущества: компактность, сохранение относительных связей
- Недостатки: вводит ложные интервалы, если порядок условен

Частотное кодирование (Frequency encoding)

- Каждая категория заменяется частотой её встречаемости в данных
- Кодировка отражает статистическую распространённость признака
- Компактно при большом числе уникальных значений
- Преимущества: простота, стабильность, масштабируемость
- Недостатки: теряет различия между категориями с близкой частотой

Целевое кодирование (Target encoding)

- Каждая категория заменяется статистикой целевой переменной
- Отражает связь категории с результатом
- Преимущества: информативность, уменьшение размерности
- Недостатки: риск утечки информации, переобучение при малых выборках

Метод весомости признака (Weight of Evidence)

- Используется для отражения связи категории с целевой переменной через вероятности
- Основано на сравнении распределения «хороших» и «плохих» исходов по категориям
- Применяется в задачах кредитного скоринга, риска и бинарной классификации
- Преимущества: интерпретируемость, устойчивость, монотонная связь с целевой переменной
- Недостатки: требует бинарной цели и достаточного числа наблюдений

Преобразование значений признаков

- Преобразования выравнивают распределения и стабилизируют дисперсии
- Повышают линейность связей между признаками и целевой переменной
- Основные типы преобразований:
  - Логарифмическое — сглаживает асимметрию и уменьшает влияние выбросов
  - Обратное — усиливает различия при малых значениях
  - Степенное — регулирует степень нелинейности зависимости
  - Экспоненциальное — расширяет масштаб малых изменений
  - Преобразование Бокса-Кокса(Box–Cox) — параметрический вариант стабилизации распределений

Дискретизация признаков

- Преобразует непрерывные признаки в интервальные категории
- Упрощает структуру данных и повышает интерпретируемость
- Виды дискретизации:
  - Равная ширина — деление диапазона значений на интервалы одинакового размера
  - Равная частота — интервалы содержат одинаковое число наблюдений
  - Основанная на предметных знаниях — интервалы определяются экспертно или по логике процесса

Обработка выбросов

- Цель — уменьшить влияние аномальных наблюдений без искажения данных
- Выбросы могут быть ошибками измерений или редкими, но реальными случаями
- Основные методы:
  - Удаление — исключение ошибочных наблюдений
  - Интерпретация выбросов как отсутствующих значений — при сомнительной достоверности значений
  - Дискретизация — объединение крайних значений в отдельную категорию·
  - Кодирование выбросов — введение признака, отражающего факт аномальности

Итоги

Конструирование признаков - система методов подготовки данных для анализа и моделирования

- Цель — получение информативных, согласованных и устойчивых признаков
- Основные направления:
  - обработка пропусков и аномалий
  - преобразование и масштабирование значений
  - кодирование и дискретизация признаков

- Результат — данные, отражающие закономерности системы и пригодные для обучения модели