Семинар по визуализации и генерации признаков

При подготовке презентации использованы материалы:

- А. Г. Дьяконова (https://goo.gl/cRWSwU)
- Виктора Кантора (https://vk.com/data_mining_in_action)
- Евгения Соколова (https://github.com/esokolov/ml-minor-hse)
- Статьи Хабрахабра (https://habrahabr.ru/company/mlclass/blog/248129/)

Работа над задачей

- Визуализация данных (Exploratory analysis)
- Предобработка данных (+ разделение выборки)
- Построение бейзлайна
- Доработка модели:
  - выбор алгоритма, подбор гиперпараметров
  - генерация новых признаков (feature engineering)
  - настройка композиции алгоритмов

Работа над задачей

- Визуализация данных (Exploratory analysis)
- Предобработка данных (+ разделение выборки)
- Построение бейзлайна
- Доработка модели:
  - выбор алгоритма, подбор гиперпараметров
  - генерация новых признаков (feature engineering) + отбор признаков
  - настройка композиции алгоритмов

Работа над задачей

- Визуализация данных (Exploratory analysis)
- Предобработка данных (+ разделение выборки)
- Построение бейзлайна
- Доработка модели:
  - выбор алгоритма, подбор гиперпараметров
  - генерация новых признаков (feature engineering)
  - настройка композиции алгоритмов

в каком порядке?

Задачи визуализации

- какую предобработку данных нужно провести
- какие признаки/объекты могут быть полезными/вредными для решения
- какие методы лучше использовать для предсказания
- какие признаки добавить
- каковы особенности задачи

Типы графиков

- линия
- scatter (точки) - рекомендуется ставить параметр прозрачности!
- boxplot (показывает среднее и разброс вещественного признака [в разрезе категориального])
- гистограмма (разбивает вещественную ось на интервалы и считает второй параметр в каждом интервале)
- круговая гистограмма (доли категориальных признаков)

Визуализация: базовые вещи

1. Число признаков, типы данных, наличие dummy-признаков

data.describe() # pd.DataFrame method

Какие признаки? (тип: вещественные, натуральные, бинарные, категориальные, порядковые, текстовые...)

Какая задача? (в привычных случаях регрессия/классификаця)

Много ли пропусков?

Одинаковый ли масштаб признаков?

Визуализация: базовые вещи

2. Гистограммы признаков

data.hist # pd.DataFrame method

Нет ли перекоса в сторону каких-то значений признака (к ним нужно будет внимательно относиться при обучении)
Если представлена только одна категория - такой признак нужно удалить на этапе очистки данных! (Иначе он будет портить модель)
Прикинуть, какому распределению может соответствовать гистограмма
Постараться интерпретировать полученные распределения признаков, нет ли чего-нибудь странного в данных?
Обратите особое внимание на целевой признак!

Отдельно распределения для обучающей и контрольной выборок!

Визуализация: базовые вещи

3. Графики признак - целевой признак

Выводы делать опасно, но прикинуть, что может оказаться полезным, можно.

4. Понижение размерности

Специальные методы (MDS, tSNE) для более наглядных картинок

Общие советы

- Визуализация dummy-признаков (Id)

Распространенные типы данных

- Вещественные признаки
- Категориальные признаки
- Даты и время
- Координаты (“географические признаки”)
- Тексты
- Изображения

Вещественные признаки

- Нелинейные трансформации
- Дискретизация
  - по порогам (по визуализации)
  - округление
  - кластеризация значений
- Агрегация нескольких признаков в один (сумма, максимум - по семантике)

Даты и время

- дата → день недели, месяц, год
- время → время суток, час, минута
- праздник / выходной (бинарный признак)
- циклическое кодирование
- признаки на основе разности
  - два временных признака
  - между объектами
  - время до важного события (праздника)
  - количество прошедших секунд с какого-то момента

Даты и время как правило нужно учитывать при разделении выборки!

Географические признаки

- Проекции на разные оси
- Кластеризация
- Плотность точек в данной области
- Расстояния до центров кластеров, важных географических объектов

Категориальные признаки

- Label encoding (не для линейных моделей)
- One-hot encoding
- Count encoding (число / доля объектов с таким значением)
- Хеширование
- Усреднение значений вещественного признака по категориям (совсем не обязательно целевого)
- По значению целевого признака (счетчики и усреднение) CV-оценки!

Можно учитывать взаимодействия признаков! (с малым числом категорий)

Текстовые признаки

Текст - последовательность символов

Разные уровни:

- в данных есть строковый признак: без пробела / относительно мало вариантов → в категориальные признаки
- в данных есть строковый признак - длинное предложение / цельный текст → свои методы генерации признаков (фиксированная размерность)
- объект - текст → свои методы обработки последовательностей переменной длины (RNN, HMM...)

Тексты

Текст - последовательность символов

- в данных есть строковый признак: без пробела / относительно мало вариантов → в категориальные признаки
- в данных есть строковый признак - длинное предложение / цельный текст → свои методы генерации признаков (фиксированная размерность)
- объект - текст → свои методы обработки последовательностей переменной длины (RNN, HMM...)

Два уровня: character-based vs word-based

Предобработка текстовых данных

1. Удаление служебных символов: тегов, знаков, [пунктуации]
   …
2. [Приведение к нижнему регистру]
3. Токенизация — разделение на слова (дефисы? пунктуация?)
4. Нормализация
   1. Стемминг — удаление окончаний
   2. Лемматизация — приведение слова к нормальной форме
5. Удаление стоп-слов

Признаки из текстов

- Мешок слов
  - Для каждого слова: сколько раз встретилось / встретилось или нет

$$d \to (n_1, \ldots, n_W)$$

71 the	31 garden
22 and	19 of
19 to	19 house
18 in	18 white
12 trees	12 first
11 a	11 president
8 for	7 as
7 gardens	7 rose
7 tour	6 on
6 was	6 east
6 tours	5 planting
5 he	5 is
5 grounds	5 that
5 gardener	4 history
4 text-decoration	4 john
4 kennedy	4 april
4 been	4 today
4 with	4 none
4 adams	4 spring
4 at	4 had
3 mrs	3 lawn
…	…

Признаки из текстов

- Мешок слов: $d \to (n_1, \ldots, n_W)$
- Tf-Idf представление

$$d \to (TF(1)IDF(1) \ldots TF(W)IDF(W))$$

$$TF(i) = \frac{n_i}{n},\ n = \sum_{i=1}^{W} n_i$$

$$IDF(i) = \frac{N}{N_i}$$

N ← всего документов

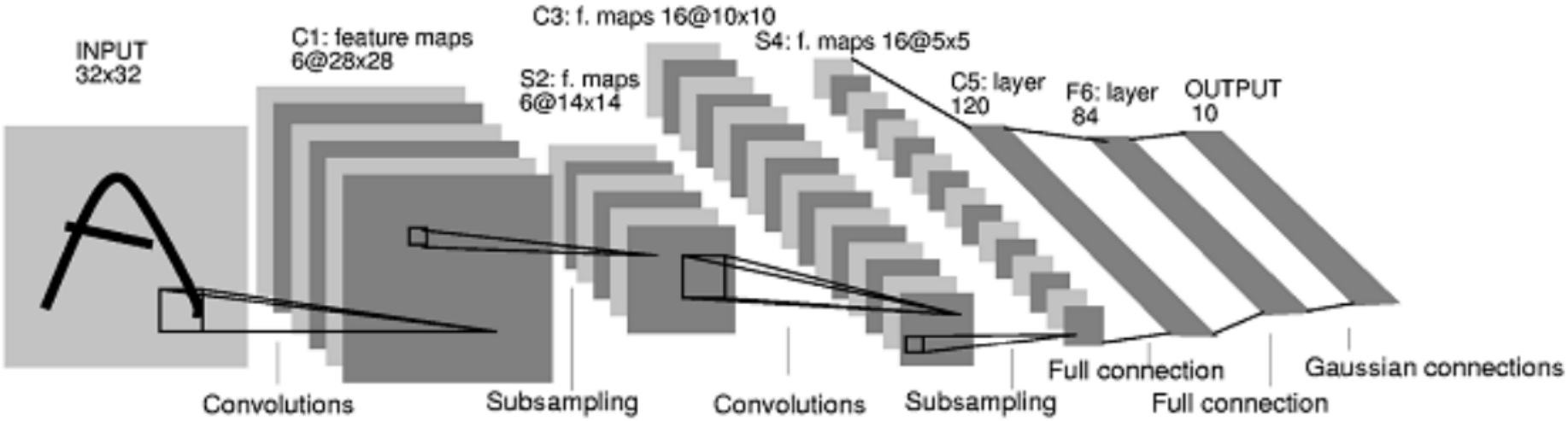
N_i ← документов со словом i

Признаки из текстов

- Мешок слов и TF-IDF для n-грамм
- Мешок буквосочетаний (устойчивость к опечаткам)
- Скрытые представления
  - усреднение word2vec
  - doc2vec
  - из RNN
  - матричные разложения (например, тематическое моделирование)

Признаки для изображений

- Выходы слоев нейросети



Классические признаки для изображений

- гистограммы яркости - не учитывают порядок пикселей
- гистограммы градиентов (фильтр Собеля) - уже как-то учитывают
- и другие штуки из Computer Vision

“Глобальные” методы

- Кластеризация объектов
- Расстояния до эталонных объектов
- Выходы других алгоритмов как признаки
- Понижение размерности (PCA, автокодировщики) - особенно для разреженных данных

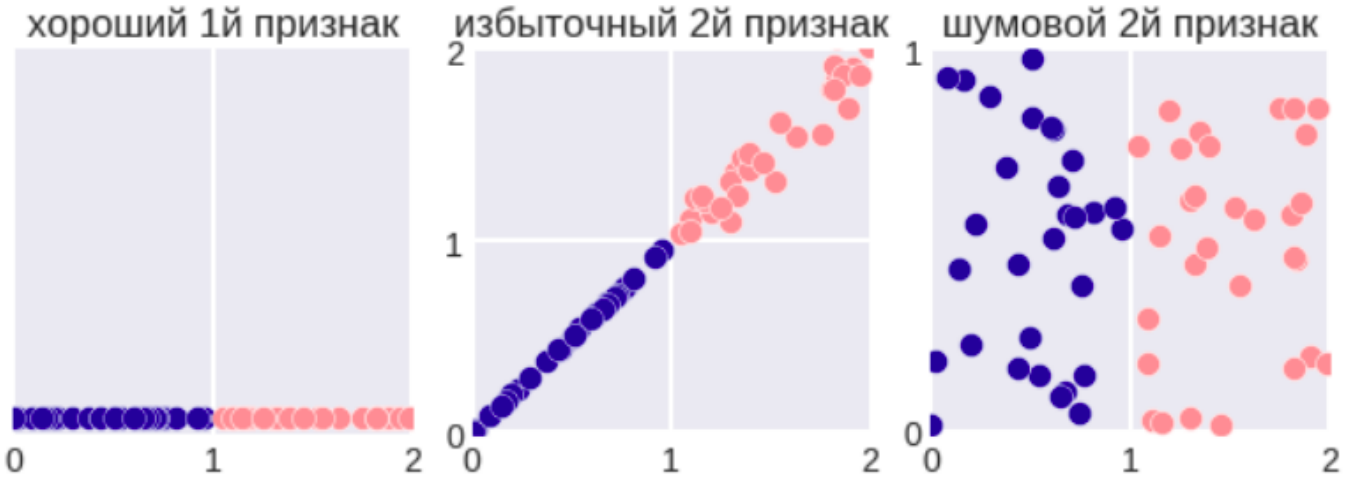
Итого: популярные приемы генерации

- кластеризация → новый признак кластер
- отношения с другими объектами и признаками (расстояния, разности значений признаков)
- понижение размерности

Итого по генерации признаков

- Генерация признаков тесно связана с data exploration
- Что подойдет и заработает - очень сильно зависит от задачи
- Одним sklearn не обойтись, очень полезным оказывается pandas + numpy

Отбор признаков и отбор объектов



Картинка из слайдов Дьяконова А. Г.

В удалении объектов - третья ситуация

Отбор признаков

Зачем:

- скорость работы алгоритмов
- повышение качества
- борьба с переобучением

Методы отбора признаков

- Статистические методы: оценивается зависимость целевого признака от других (фильтрация) - обычно одномерные методы
- Отбор с помощью моделей (feature_importance_)
- Методы-обертки: выбор признаков, дающих лучшее качество для модели

Одномерные методы

Корреляция

$$R_j = \frac{\sum_{i=1}^{\ell}(x_{ij} - \bar{x_j})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{\ell}(x_{ij} - \bar{x_j})^2 \sum_{i=1}^{\ell}(y_i - \bar{y})^2}}$$

- Чем больше $|R_j|$, тем информативнее признак
- Учитывает только линейную связь

- x_{ij} — значение j-го признака на i-м объекте
- $\bar{x_j}$ — среднее значение j-го признака
- y_i — значение целевой переменной на i-м объекте
- $\bar{y}$ — среднее значение целевой переменной

T-score

$$R_j = \frac{|\mu_1 - \mu_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

- Для задач бинарной классификации
- Чем больше R_j, тем информативнее признак
- μ_1, μ_2 — средние значения признаков в первом и втором классах
- σ_1^2, σ_2^2 — дисперсии
- n_1, n_2 — число объектов в первом и втором классах

F-score

$$R_j = \frac{\sum_{k=1}^{K} \frac{n_j}{K-1} \left(\mu_j - \mu\right)^2}{\frac{1}{\ell - K} \sum_{k=1}^{K} \left(n_j - 1\right) \sigma_j^2}$$

- Для задач многоклассовой классификации
- Чем больше R_j, тем информативнее признак
- $\mu_1, \dots, \mu_K$ — средние значения признаков в классах
- μ — среднее значение признака по всей выборке
- $\sigma_1^2, \dots, \sigma_K^2$ — дисперсии
- $n_1, \dots, n_K$ — число объектов в первом и втором классах

Одномерные методы

Для категориальных признаков - взаимная информация

$$MI = \sum_{x_i \in X} \sum_{y_j \in Y} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i) p(y_j)}$$

Отбор с помощью моделей

- L1-регуляризация
- feature_importance_ в деревьях и лесах

Решающие деревья

- Чем сильнее уменьшили $H(X)$, тем лучше признак
- Уменьшение критерия:

$$H(X_m) - \frac{|X_l|}{|X_m|} H(X_l) - \frac{|X_r|}{|X_m|} H(X_r)$$

- Важность признака R_j: просуммируем уменьшения по всем вершинам, где разбиение делалось по признаку j

Случайный лес

- Сумма важностей R_j по всем деревьям
- Чем больше, тем важнее признак
- Учитывается важность признаков в совокупности

Методы-обертки

- Переставить значения признака и оценить падение в качестве
- Качество работы алгоритма для подмножества признаков

Итого по отбору признаков

- Разные методы выделят разные признаки, универсального рецепта нет
- Лучше выбирать одним методом, а предсказывать другим

Трансформация ответов

Иногда качество решения задачи уже после обучения можно повысить, трансформируя предсказания модели

Примеры:

- “Нормализация” целевого признака до обучения и возвращение в исходную шкалу после
- Подбор порогов при предсказании порядковых величин