Методы машинного обучения. Преобразование данных, оценивание и выбор моделей

Воронцов Константин Вячеславович
www.MachineLearning.ru/wiki?title=User:Vokov
вопросы к лектору: voron@forecsys.ru

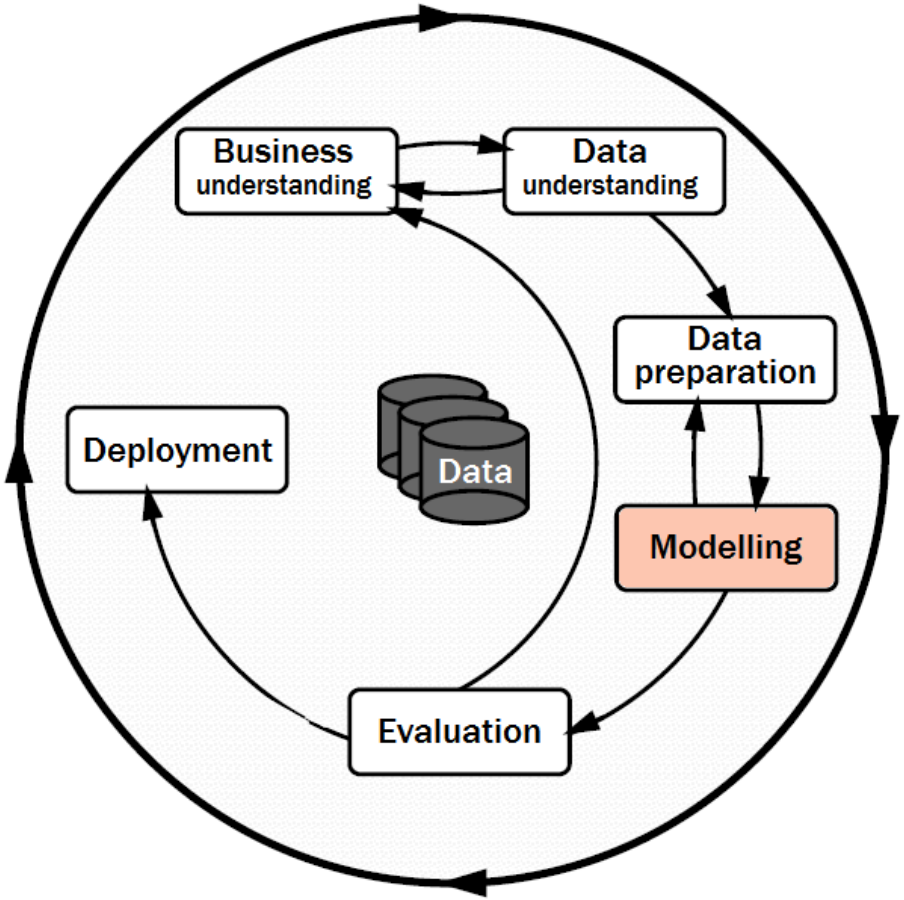
материалы курса:
github.com/MSU-ML-COURSE/ML-COURSE-21-22
орг.вопросы по курсу: ml.cmc@mail.ru

ВМК МГУ • 14 декабря 2021

Содержание

Межотраслевой стандарт интеллектуального анализа данных

CRISP-DM: CRoss Industry Standard Process for Data Mining (1999)



Компании-инициаторы:

- SPSS
- Teradata
- Daimler AG
- NCR Corp.
- OHRA

Шаги процесса:

- понимание бизнеса
- понимание данных
- предобработка данных
- моделирование
- оценивание
- внедрение

Шкалы измерения

Измерительная шкала — множество Z допустимых значений, получаемых в результате измерения признака $f(x)$, $f: X \to Z$

Тип шкалы определяется множествами

- допустимых биективных преобразований $\psi: Z \to Z'$
- допустимых операций над значениями из шкалы Z

Классификация типов измерительных шкал по Стивенсу:

шкала	D	$\psi(z)$	операции
логическая (boolean)	$0, 1$	биективные	$\vee \wedge \neg$
номинальная (nominal)	$< \infty$	биективные	$= \neq \in$
порядковая (ordinal)	$< \infty$	монотонные	$= \neq \in < >$
интервальная (interval)	$\mathbb{R}$	$az + b$	$< > + -$
отношений (ratio)	$\mathbb{R}$	az	$< > + - \times \div$
абсолютная (absolute)	$\mathbb{R}$	z	любые

S.S.Stevens. On the Theory of Scales of Measurement // Science, 1946.

Примеры величин, измеряемых в различных шкалах

- **Логическая**
  наличие/отсутствие свойства, ответ «да/нет»
- **Номинальная** (можно переименовать или перенумеровать)
  идентификаторы классов, людей, регионов, фирм, товаров
- **Порядковая** (порядок частичный или линейный)
  уровень образования, тяжесть болезни, степень согласия
- **Ранговая** (частный случай порядковой: $1, 2, 3, \ldots, N$)
  оценка в баллах, шкалы Рихтера, Бофорта, Мооса
- **Интервальная** (можно сдвигать положение нуля)
  время, географическая широта, температура (°C, °F)
- **Отношений** (можно менять единицы измерения)
  масса, скорость, объём, сила, давление, заряд, яркость
- **Абсолютная**
  число предметов, частота события, оценка вероятности

Ослабление шкалы

Номинальный $\to$ много бинарных (one-hot-encoding):

- $f_v(x) = \left[f(x) = v\right]$, для всех значений v признака
- $f_A(x) = \left[f(x) \in A\right]$, индикаторный признак подмножества A

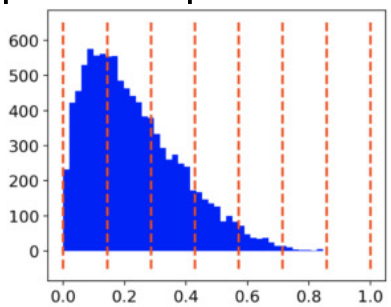
Числовой или порядковый $\to$ бинарный:

- $f_{a,b}(x) = \left[a \leqslant f(x) \leqslant b\right]$ для заданного отрезка $[a, b]$

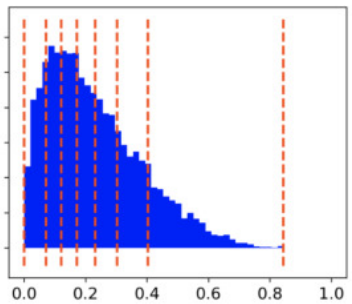
Числовой $\to$ ранговый (data binning, quantization):

- $f_a(x) = \sum_{k=1}^{K} \left[f(x) \geqslant a_k\right]$, номер интервала сетки $a_1, \ldots, a_K$

равномерная сетка



квантильная сетка



Ослабление шкалы всегда влечёт потерю информации

Усиление шкалы

Номинальный $\to$ числовой:

- категория заменяется частотой:
$$f'(x) = \frac{1}{\ell}\sum_{i=1}^{\ell}[f(x_i) = f(x)]$$
- условное среднее числового признака $g(x)$:
$$f'(x) = \text{mean}\big(g|f(x)\big) = \frac{\sum_{i=1}^{\ell} g(x_i)[f(x_i) = f(x)]}{\sum_{i=1}^{\ell}[f(x_i) = f(x)]},$$
- условное среднее целевой величины $y(x)$:
$f'(x) = \text{mean}\big(y|f(x)\big)$, возможно переобучение!

Порядковый $\to$ числовой:

- значение заменяется частотой:
$$f'(x) = \frac{1}{\ell}\sum_{i=1}^{\ell}[f(x_i) \leqslant f(x)]$$

Нормализация и стандартизация числовых шкал

Многие методы накапливают меньше вычислительных погрешностей, если признаки приведены к одному масштабу

- $f'_j(x) = \dfrac{f_j(x) - f_j^{\min}}{f_j^{\max} - f_j^{\min}}$ — нормализация, приведение к $[0, 1]$
- $f'_j(x) = \dfrac{f_j(x)}{|f_j|^{\max}}$ — масштабирование с сохранением нуля
- $f'_j(x) = \dfrac{f_j(x) - \mu_j}{\sigma_j}$ — стандартизация

$f_j^{\max}$, $|f_j|^{\max}$, $f_j^{\min}$, μ_j, σ_j определяются по обучающей выборке

Для повышения устойчивости к выбросам можно отбрасывать 5% наименьших и наибольших значений признака

Трансформация вида распределения

F_j — функция распределения (c.d.f.) признака f_j
Эмпирическая функция распределения (кусочно-постоянная):

$$\hat{F}_j(z) = \frac{1}{\ell}\sum_{i=1}^{\ell}\big[f_j(x_i) \leqslant z\big]$$

- $f'_j(x) = F_j(f_j(x))$ — преобразование $f_j(x)$ в равномерную на отрезке $[0,1]$ случайную величину
- $f'_j(x) = \Phi^{-1}\big(F_j(f_j(x))\big)$ — преобразование $f_j(x)$ в случайную величину с заданной функцией распределения Φ (например, в нормальную)
- $f'_j(x) = \ln\big(1 + f_j(x)\big)$ — преобразование случайной величины «с тяжёлым правым хвостом» (объёмы производства, перевозок, продаж)

Подходы к обработке пропущенных значений

- Игнорировать объекты или признаки с пропусками :(
- Заполнить пропущенные значения признака f:
  — средним или медианным значением $\bar{f}$
- Прогнозировать значения признака f по остальным:
  — регрессия для вещественного признака f
  — классификация для дискретного признака f
  — матричные разложения, например, разреженный SVD
- Использовать модели, способные обрабатывать пропуски:
  — решающие деревья
  — голосование низкоразмерных базовых предикторов
- Ввести бинарный признак $f'(x) = [f(x) \text{ не известно}]$

Непараметрическая регрессия для заполнения пропусков

Формула Надарая–Ватсона, ядерное сглаживание:

$$\hat{f}_j(x_i) = \frac{\sum_u f_j(u) S(u, x_i)}{\sum_u S(u, x_i)}$$

где $\sum_u$ — сумма по всем объектам $u \in X^\ell$ с известным $f_j(u)$

Возможные конструкции функций сходства $S(u, x)$:

- $S(u,x) = K\left(\frac{\rho(u,x)}{h}\right), \quad \rho^2(u,x) = \frac{1}{|J_{ux}|} \sum_{j \in J_{ux}} \big(f_j(u) - f_j(x)\big)^2$
- $S(u,x) = \frac{1}{|J_{ux}|} \sum_{j \in J_{ux}} f_j(u) f_j(x)$ — скалярное произведение
- $S(u,x) = \frac{\sum_{j \in J_{ux}} f_j(u) f_j(x)}{\sqrt{\sum_{j \in J_{ux}} f_j^2(u)} \sqrt{\sum_{j \in J_{ux}} f_j^2(x)}}$ — косинусная ф.сх.

где J_{ux} — множество признаков j с известными $f_j(x)$ и $f_j(u)$

Разреженное низкоранговое матричное разложение

Дано: матрица $F = \big(f_{ij} = f_j(x_i)\big)_{\ell \times n}$, $\Omega \subseteq \{1, \ldots, \ell\} \times \{1, \ldots, n\}$

Найти: матрицы $G = (g_{it})_{\ell \times k}$ и $U = (u_{jt})_{n \times k}$ такие, что

$$\big\|F - GU^{\mathsf{T}}\big\| = \sum_{(i,j)\in\Omega} \underbrace{\big(f_{ij} - \langle g_i, u_j \rangle\big)}_{\varepsilon_{ij}}^2 = \sum_{(i,j)\in\Omega} \Big(f_{ij} - \sum_{t=1}^{k} g_{it} u_{jt}\Big)^2 \to \min_{G,U}$$

Классический SVD неприменим для разреженной задачи.

Метод стохастического градиента: перебираем $(i,j) \in \Omega$ в случайном порядке, делаем градиентные шаги $(\varepsilon_{ij})^2 \to \min\limits_{g_i, u_j}$:

$$g_{it} := g_{it} + \eta \varepsilon_{ij} u_{jt}, \quad t = 1, \ldots, k$$

$$u_{jt} := u_{jt} + \eta \varepsilon_{ij} g_{it}, \quad t = 1, \ldots, k$$

$\hat{f}_j(x_i) = \langle g_i, u_j \rangle$ — восстановление пропущенных значений

g_{it} — новые признаки x_i в пространстве размерности k

Классические подходы к генерации признаков

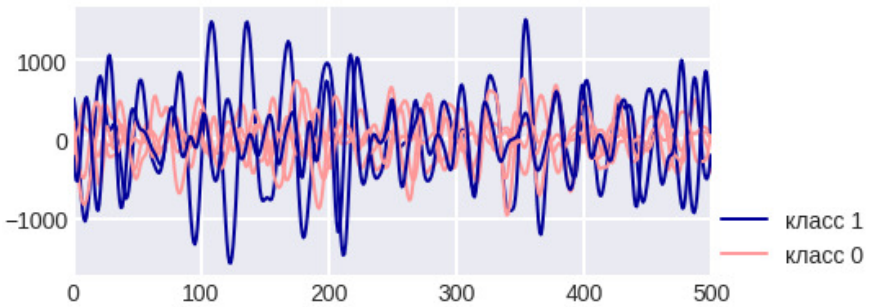
Feature Engineering: признаки вычисляются по формулам, которые зависят от задачи, требуют изобретательности и знаний предметной области. Долго, дорого.

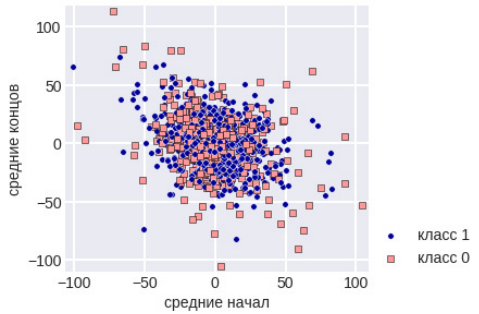
- Прогнозирование временных рядов:
  признаки агрегируются по предыстории различной глубины
- Распознавание лиц:
  признаки размера и формы черт лица
- Классификация и поиск текстов:
  признаки частоты слов, терминов, названий, синонимов
- Распознавание речи:
  спектральные, фонетические, лингвистические признаки

Иногда удачные признаки решают задачу без ML

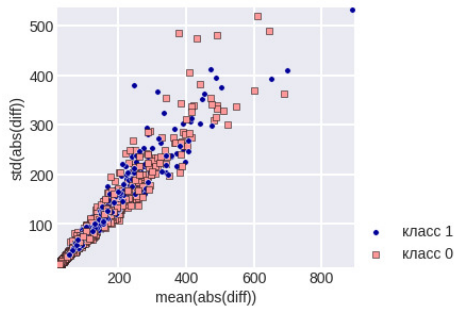
Соревнование «Ford Classification Challenge» (2008)
Задача детектирования поломок по сигналу датчика

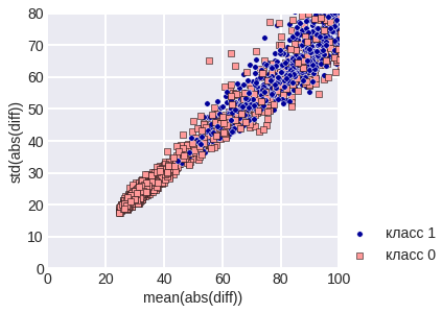
Признаки, генерируемые по исходным временным рядам, слабы:

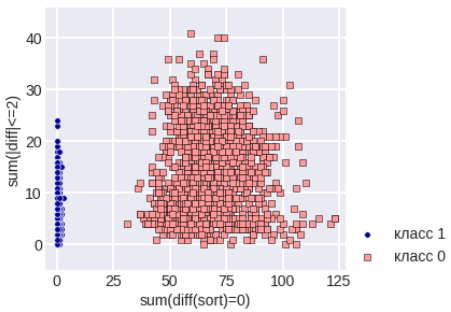




Среди признаков рядов их производных оказывается идеальный:







https://dyakonov.org/2018/06/28/простые-методы-анализа-данных/

Общий подход — автоматическая векторизация данных

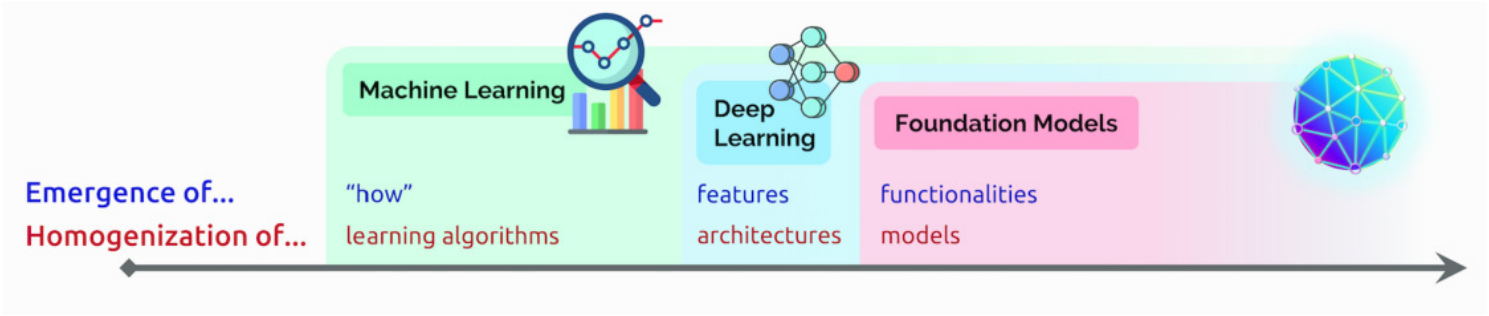
Глубокие нейронные сети объединили этапы векторизации данных и обучения предсказательной модели

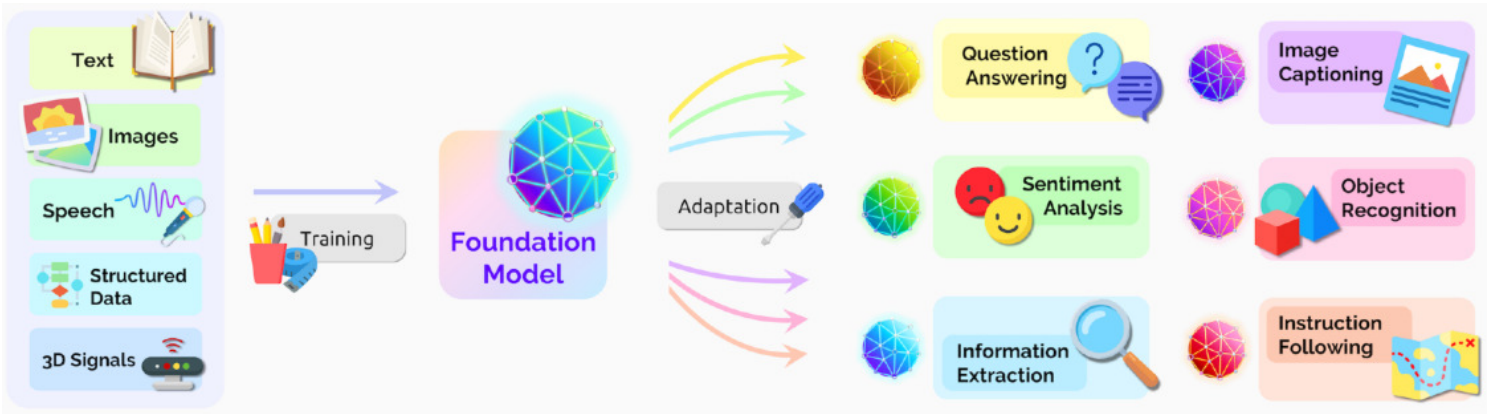
- Компьютерное зрение
- Обработка текстов естественного языка
- Распознавание и синтез речи
- Анализ транзакционных данных
- Анализ сигналов
- Прогнозирование временных рядов
- Анализ графов
- ...

(в следующем семестре)

Концепция фундаментальных моделей (Foundation Models)

Обучаемая векторизация данных — глобальный тренд в ML





R.Bommasani et al. (Stanford University) On the opportunities and risks of foundation models // CoRR, 20 August 2021.

Методология анализа ошибок

$\mathscr{L}(x_i, a)$ — функция потерь (чем меньше, тем лучше)
Критерий средней потери алгоритма a на выборке U:

$$Q(a) = \frac{1}{|U|} \sum_{x_i \in U} \mathscr{L}(x_i, a) \;\to\; \min$$

- Ранжировать объекты по убыванию потерь $\mathscr{L}_i = \mathscr{L}(x_i, a)$
- Сравнить распределения потерь на обучении и тесте.
- Если сильно отличаются, то надо устранять переобучение.
- Есть ли объекты со сверхбольшими потерями, много ли их?
- Можно ли от них избавиться? Может, это выбросы?
- Если нет, то что общего у объектов с большими потерями?
- Как модифицировать модель, чтобы уменьшить потери?
- Может, что-то не так с функцией потерь?

A/B тестирование (A/B testing, Split Testing)

Две модели, «базовая A» и «улучшенная B»,
построенные по историческим данным X^ℓ,
тестируются по метрике качества Q на новых данных X^k

В чём отличия A/B тестирования от обычного hold-out?

- X^k — это именно будущие данные (out-of-time), а не часть прошлых данных, исключённых из обучения (out-of-sample)
- больше реализма: за это время могут измениться свойства потока данных, реальные данные не обязаны быть i.i.d.
- однократный выбор модели почти не переобучается
- накопление данных X^k может потребовать много времени
- работа модели может влиять на формирование потока данных (например, в рекомендательных системах)

Мета-обучение (meta-learning, learning to learn)

Дано: выборка «задача, метод» $\to$ критерии качества

Найти: модель предсказания, каким методом решать задачу

Критерий: точность предсказания оптимального метода

Признаки:

- размерные характеристики задачи
- характеристики пространства признаков:
  типы, выбросы, пропуски, корреляции
- результаты быстрых низкоразмерных методов

Joaquin Vanschoren. Meta-learning Architectures: Collecting, Organizing and Exploiting Meta-knowledge. 2009.

Joaquin Vanschoren. Meta-Learning: A Survey. 2018.

Автоматический выбор моделей и гиперпараметров (AutoML)

Проблема:
подбор структуры модели (архитектуры нейросети)
и гиперпараметров требует слишком много ресурсов

Дано: выборка «задача, структура» $\to$ критерии качества

Найти: какой следующий эксперимент провести с моделью

Критерий:
минимизация затрат ресурсов на автоматический поиск
оптимальной модели, сопоставимой по качеству с моделями,
построенными профессиональными исследователями

Близкая классическая задача — *планирование экспериментов*

Xin He, Kaiyong Zhao, Xiaowen Chu. AutoML: A Survey of the State-of-the-Art. 2019

https://github.com/sberbank-ai-lab/LightAutoML — AutoML от Сбербанка

Резюме

- Культура анализа данных:
  — смотреть на данные глазами
  — делать анализ ошибок
  — в случае неудачи порождать новые гипотезы
  — учитывать сильные и слабые стороны методов
- Автоматизация распространяется по схеме CRISP-DM, охватывая не только моделирование, но и предобработку, оценивание и выбор моделей
- Современный подход к векторизации данных — глубокие нейронные сети (об этом в следующем семестре)