

Машинное обучение

Задача обучения по прецедентам

X — множество объектов;

Y — множество ответов;

$y: X \rightarrow Y$ — неизвестная зависимость (target function).

Дано:

$\{x_1, \dots, x_\ell\} \subset X$ — обучающая выборка (training sample);

$y_i = y(x_i)$, $i = 1, \dots, \ell$ — известные ответы.

Найти:

$a: X \rightarrow Y$ — алгоритм, решающую функцию (decision function), приближающую y на всём множестве X .

Весь курс машинного обучения — это конкретизация:

- › как задаются объекты и какими могут быть ответы;
- › как строить функцию a ;
- › в каком смысле a должен приближать y .

Как задаются объекты. Признаковое описание

$f_j: X \rightarrow D_j, j = 1, \dots, n$ — признаки объектов (features).

Типы признаков:

- › $D_j = \{0, 1\}$ — бинарный признак f_j ;
- › $|D_j| < \infty$ — номинальный признак f_j ;
- › $|D_j| < \infty, D_j$ упорядочено — порядковый признак f_j ;
- › $D_j = \mathbb{R}$ — количественный признак f_j .

Вектор $(f_1(x), \dots, f_n(x))$ — признаковое описание объекта x .

Матрица «объекты–признаки» (feature data)

$$F = \parallel f_j(x_i) \parallel_{\ell \times n} = \begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_\ell) & \dots & f_n(x_\ell) \end{pmatrix}$$

Как задаются ответы. Типы задач

Задачи классификации (classification):

- › $Y = \{-1, +1\}$ — классификация на 2 класса.
- › $Y = \{1, \dots, M\}$ — на M непересекающихся классов.
- › $Y = \{0, 1\}^M$ — на M классов, которые могут пересекаться.

Задачи восстановления регрессии (regression):

- › $Y = \mathbb{R}$ или $Y = \mathbb{R}^m$.

Задачи ранжирования (ranking, learning to rank):

- › Y — конечное упорядоченное множество.

Предсказательная модель

Модель (predictive model) — параметрическое семейство функций

$$A = \{a(x) = g(x, \theta) \mid \theta \in \Theta\},$$

где $g: X \times \Theta \rightarrow Y$ — фиксированная функция,
 Θ — множество допустимых значений параметра θ .

Пример.

Линейная модель с вектором параметров $\theta = (\theta_1, \dots, \theta_n)$, $\Theta = \mathbb{R}^n$:

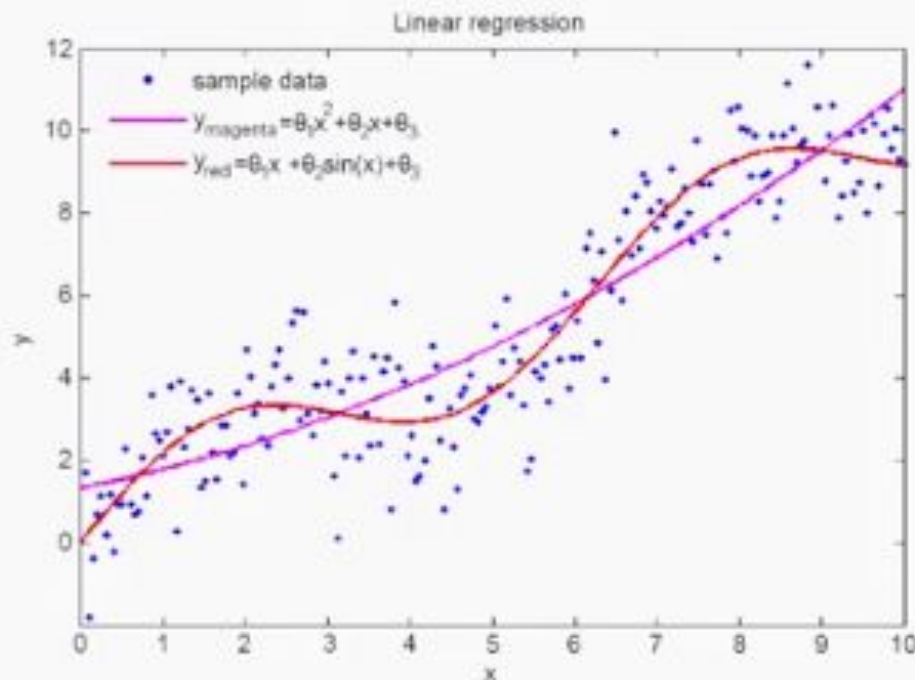
$$g(x, \theta) = \sum_{j=1}^n \theta_j f_j(x) \quad \text{— для регрессии и ранжирования, } Y = \mathbb{R};$$

$$g(x, \theta) = \text{sign} \sum_{j=1}^n \theta_j f_j(x) \quad \text{— для классификации, } Y = \{-1, +1\}.$$

Пример: задача регрессии, модельные данные

$X = Y = \mathbb{R}$, $\ell = 200$, $n = 3$ признака: $\{x, x^2, 1\}$ или $\{x, \sin x, 1\}$

$X = Y = \mathbb{R}$, $\ell = 200$, $n = 3$ признака: $\{x, x^2, 1\}$ или $\{x, \sin x, 1\}$



Вывод: признаковое описание можно задавать по-разному

Этап обучения (train):

Метод обучения (learning algorithm) $\mu: (X \times Y)^\ell \rightarrow A$
по выборке $X^\ell = (x_i, y_i)_{i=1}^\ell$ строит алгоритм $a = \mu(X^\ell)$:

$$\boxed{\begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_\ell) & \dots & f_n(x_\ell) \end{pmatrix} \xrightarrow{y} \begin{pmatrix} y_1 \\ \dots \\ y_\ell \end{pmatrix} \xrightarrow{\mu} a}$$

Этап применения (test):

алгоритм a для новых объектов x'_1, \dots, x'_k выдаёт ответы $a(x'_i)$.

$$\begin{pmatrix} f_1(x'_1) & \dots & f_n(x'_1) \\ \dots & \dots & \dots \\ f_1(x'_k) & \dots & f_n(x'_k) \end{pmatrix} \xrightarrow{a} \begin{pmatrix} a(x'_1) \\ \dots \\ a(x'_k) \end{pmatrix}$$

$\mathcal{L}(a, x)$ — функция потерь (loss function) — величина ошибки алгоритма $a \in A$ на объекте $x \in X$.

Функции потерь для задач классификации:

› $\mathcal{L}(a, x) = [a(x) \neq y(x)]$ — индикатор ошибки;

Функции потерь для задач регрессии:

› $\mathcal{L}(a, x) = |a(x) - y(x)|$ — абсолютное значение ошибки;

› $\mathcal{L}(a, x) = (a(x) - y(x))^2$ — квадратичная ошибка.

Эмпирический риск — функционал качества алгоритма a на X^ℓ :

$$Q(a, X^\ell) = \frac{1}{\ell} \sum_{i=1}^{\ell} \mathcal{L}(a, x_i).$$

Понятие обобщающей способности (generalization performance):

- › найдём ли мы «закон природы» или *переобучимся*, то есть подгоним функцию $g(x_i, \theta)$ под заданные точки?
- › будет ли $a = \mu(X^\ell)$ приближать функцию y на всём X ?
- › будет ли $Q(a, X^k)$ мало на новых данных — контрольной выборке $X^k = (x'_i, y'_i)_{i=1}^k$, $y'_i = y(x_i)$?

Восстановление зависимостей по эмпирическим данным

Задача восстановления зависимости $y = y(x)$ по точкам *обучающей выборки* (x_i, y_i) , $i = 1, \dots, \ell$.

Дано: векторы $x_i = (x_i^1, \dots, x_i^n)$ — объекты обучающей выборки, $y_i = y(x_i)$ — правильные ответы, $i = 1, \dots, \ell$:

$$\begin{pmatrix} x_1^1 & \dots & x_1^n \\ \dots & \dots & \dots \\ x_\ell^1 & \dots & x_\ell^n \end{pmatrix} \xrightarrow{y} \begin{pmatrix} y_1 \\ \dots \\ y_\ell \end{pmatrix}$$

Найти: функцию $a(x)$, способную давать правильные ответы на *тестовых объектах* $\tilde{x}_i = (\tilde{x}_i^1, \dots, \tilde{x}_i^n)$, $i = 1, \dots, k$:

$$\begin{pmatrix} \tilde{x}_1^1 & \dots & \tilde{x}_1^n \\ \dots & \dots & \dots \\ \tilde{x}_k^1 & \dots & \tilde{x}_k^n \end{pmatrix} \xrightarrow{a?} \begin{pmatrix} a(\tilde{x}_1) \\ \dots \\ a(\tilde{x}_k) \end{pmatrix}$$

Задачи медицинской диагностики

Объект — пациент в определённый момент времени.

Классы: диагноз или способ лечения или исход заболевания.

Примеры признаков:

- › бинарные: пол, головная боль, слабость, тошнота, и т. д.
- › порядковые: тяжесть состояния, желтушность, и т. д.
- › количественные: возраст, пульс, артериальное давление, содержание гемоглобина в крови, доза препарата, и т. д.

Особенности задачи:

- › обычно много «пропусков» в данных;
- › как правило, недостаточный объём данных;
- › нужен интерпретируемый алгоритм классификации;
- › нужна оценка вероятности

Задача кредитного скоринга

Объект — заявка на выдачу банком кредита.

Классы — bad или good.

Примеры признаков:

- › бинарные: пол, наличие телефона, и т. д.
- › номинальные: место проживания, профессия, работодатель, и т. д.
- › порядковые: образование, должность, и т. д.
- › количественные: возраст, зарплата, стаж работы, доход семьи, сумма кредита, и т. д.

Особенности задачи:

- › нужно оценивать вероятность дефолта $P(\text{bad})$.

Задача предсказания оттока клиентов

Объект — абонент в определённый момент времени.

Классы — уйдёт или не уйдёт в следующем месяце.

Примеры признаков:

- › бинарные: корпоративный клиент, включение услуг, и т. д.
- › номинальные: тарифный план, регион проживания, и т. д.
- › количественные: длительность разговоров (входящих, исходящих, СМС, и т. д.), частота оплаты, и т. д.

Особенности задачи:

- › нужно оценивать вероятность ухода;
- › сверхбольшие выборки;
- › не ясно, какие признаки вычислять по «сырым» данным.

Задача категоризации текстовых документов

Объект — текстовый документ.

Классы — рубрики иерархического тематического каталога.

Примеры признаков:

- › номинальные: автор, издание, год, и т. д.
- › количественные: для каждого термина — частота в тексте, в заголовках, в аннотации, и т. д.

Особенности задачи:

- › лишь небольшая часть документов имеют метки y ;
- › документ может относиться к нескольким рубрикам;
- › в каждом ребре дерева свой классификатор на 2 класса.

Задача прогнозирования стоимости недвижимости

Объект — квартира в Москве.

Примеры признаков:

- › бинарные: наличие балкона, лифта, мусоропровода, охраны, и т. д.
- › номинальные: район города, тип дома (кирпичный/панельный/блочный/монолит), и т. д.
- › количественные: число комнат, жилая площадь, расстояние до центра, до метро, возраст дома, и т. д.

Особенности задачи:

- › выборка неоднородна, стоимость меняется со временем;
- › разнотипные признаки;
- › для линейной модели нужны преобразования признаков.

Задача прогнозирования объемов продаж

Объект — тройка (товар, магазин, день).

Примеры признаков:

- › бинарные: выходной день, праздник, промоакция, и т. д.
- › количественные: объёмы продаж в предшествующие дни.

Конкурс kaggle.com: TFI Restaurant Revenue Prediction

Объект — место для открытия нового ресторана.

Предсказать — прибыль от ресторана через год.

Задача ранжирования поисковой выдачи

Объект — пара (запрос, документ).

Классы — релевантен или не релевантен, разметка делается людьми — ассессорами.

Примеры признаков:

- › количественные:
 - частота слов запроса в документе,
 - число ссылок на документ,
 - число кликов на документ: всего, по данному запросу,
 - и т. д.

Особенности задачи:

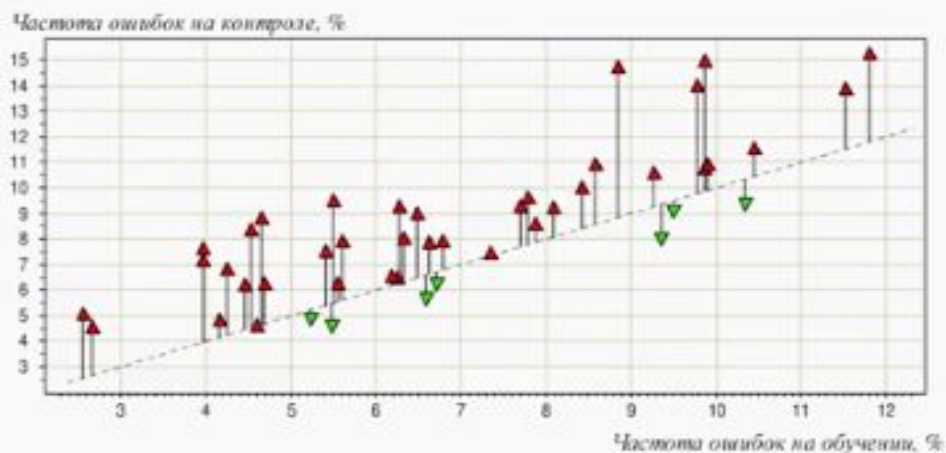
- › оптимизируется не число ошибок, а качество ранжирования;
- › сверхбольшие выборки;
- › проблема конструирования признаков по сырым данным.

Задача ранжирования в рекомендательных системах

Объект — пара (клиент, товар)
(товары — книги, фильмы, музыка).

Предсказать: вероятность покупки или рейтинг товара.

Пример. Переобучение в задаче медицинской диагностики

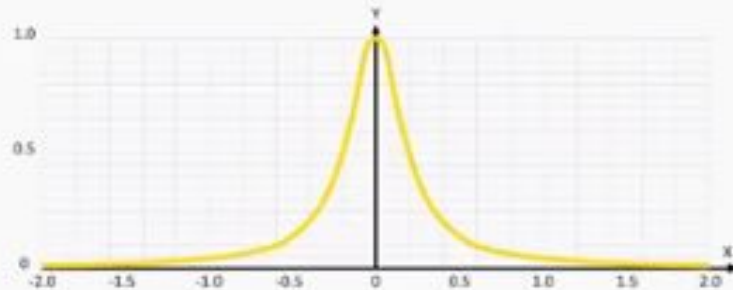


Задача предсказания отдалённого результата хирургического лечения атеросклероза.

Точки — различные алгоритмы.

Пример. Переобучение полиномиальной регрессии

Зависимость $y(x) = \frac{1}{1 + 25x^2}$ на отрезке $x \in [-2, 2]$.



Признаковое описание $x \mapsto (1, x^1, x^2, \dots, x^n)$.

Модель полиномиальной регрессии

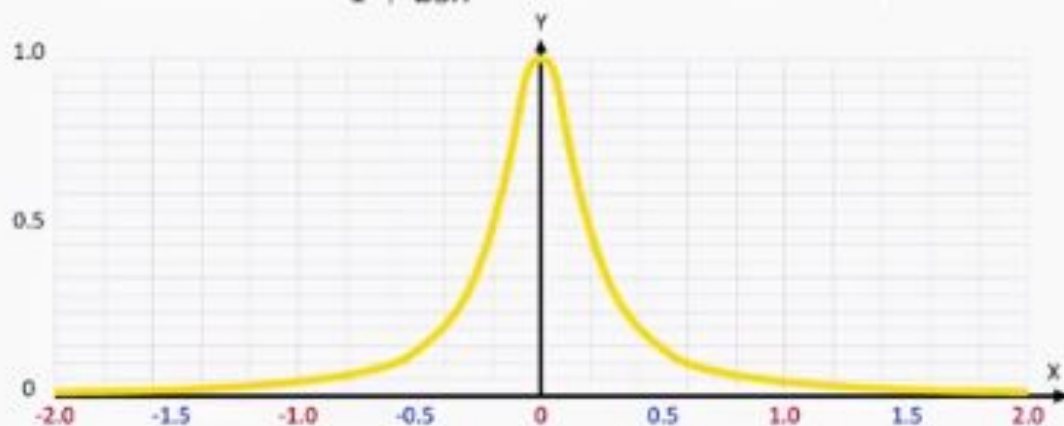
$$a(x, \theta) = \theta_0 + \theta_1 x + \dots + \theta_n x^n \text{ — полином степени } n.$$

Обучение методом наименьших квадратов:

$$Q(a, X^\ell) = \sum_{i=1}^{\ell} (\theta_0 + \theta_1 x_i + \dots + \theta_n x_i^n - y_i)^2 \rightarrow \min_{\theta_0, \dots, \theta_n}$$

Пример. Переобучение полиномиальной регрессии

Зависимость $y(x) = \frac{1}{1 + 25x^2}$ на отрезке $x \in [-2, 2]$.



Обучающая выборка:

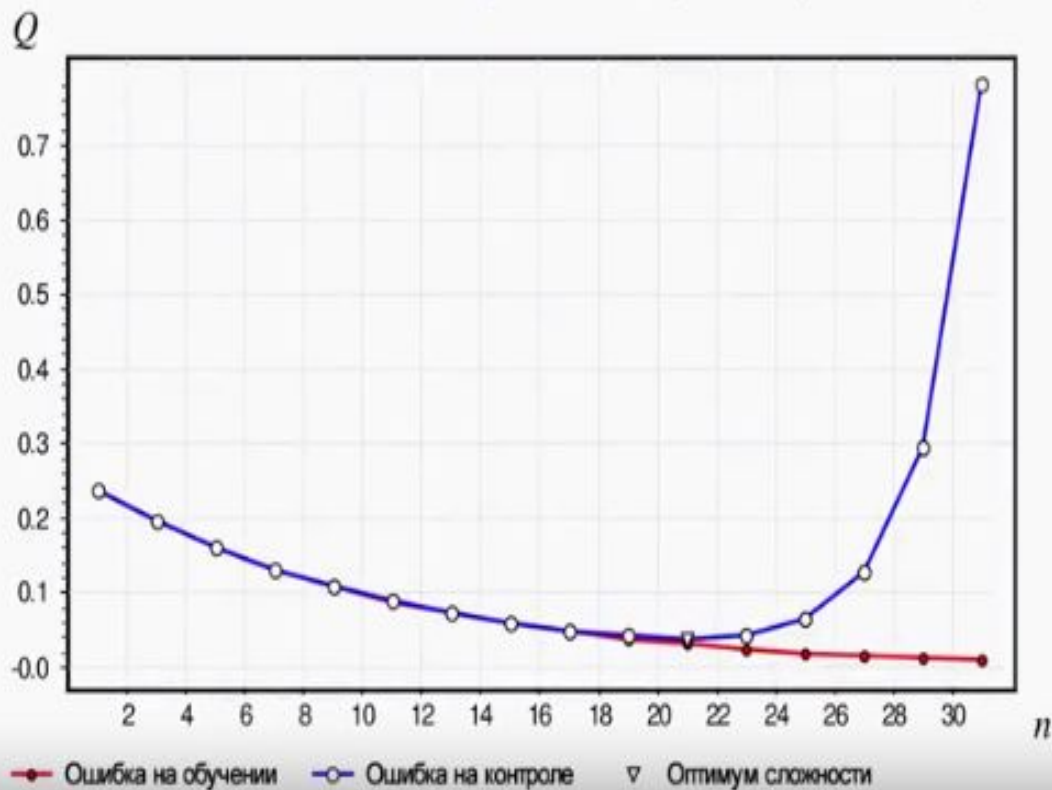
$$X^\ell = \left\{ x_i = 4 \frac{i-1}{\ell-1} - 2 \mid i = 1, \dots, \ell \right\}.$$

Контрольная выборка:

$$X^k = \left\{ x_i = 4 \frac{i-0.5}{\ell-1} - 2 \mid i = 1, \dots, \ell - 1 \right\}.$$

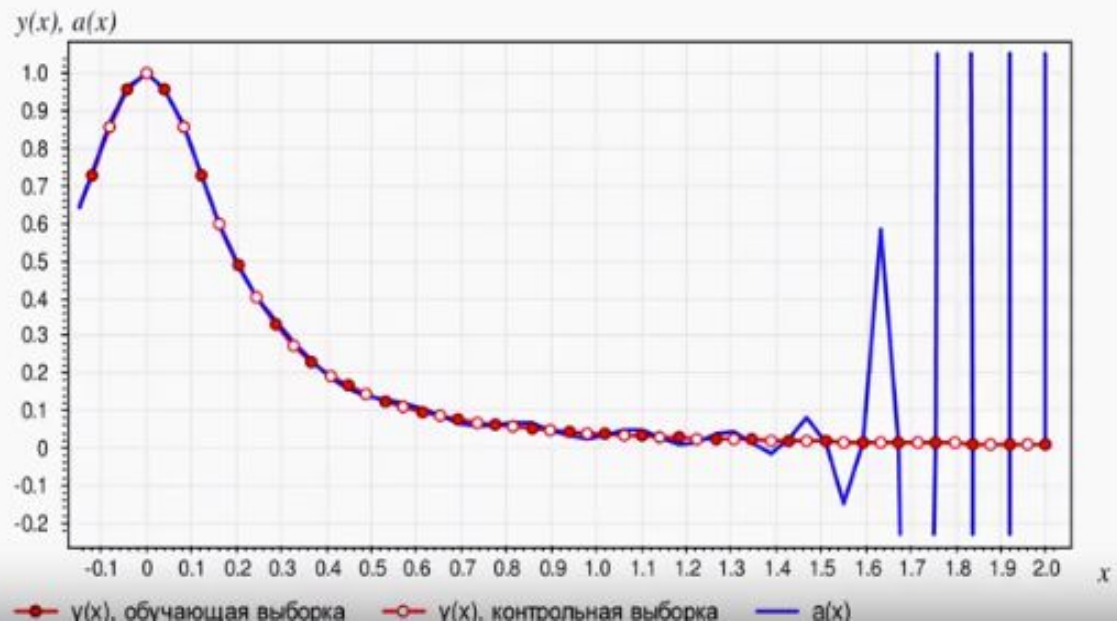
Пример переобучения: эксперимент при $l = 50$, $n = 1, \dots, 31$

Переобучение — это когда $Q(\mu(X^\ell), X^k) \gg Q(\mu(X^\ell), X^\ell)$:



Пример переобучения: эксперимент при $l = 50$, $n = 38$

$$y(x) = \frac{1}{1 + 25x^2}; \quad a(x) \text{ — полином степени } n = 38$$



- Эмпирический риск на тестовых данных (hold-out):

$$\text{HO}(\mu, X^\ell, X^k) = Q(\mu(X^\ell), X^k) \rightarrow \min$$

- Скользющий контроль (leave-one-out), $L = \ell + 1$:

$$\text{LOO}(\mu, X^L) = \frac{1}{L} \sum_{i=1}^L \mathcal{L}(\mu(X^L \setminus \{x_i\}), x_i) \rightarrow \min$$

- Кросс-проверка (cross-validation) по N разбиениям, $X^L = X_n^\ell \sqcup X_n^k$, $L = \ell + k$:

$$\text{CV}(\mu, X^L) = \frac{1}{N} \sum^N Q(\mu(X_n^\ell), X_n^k) \rightarrow \min$$

Эксперименты на реальных данных

Эксперименты на конкретной прикладной задаче:

- › цель — решить задачу как можно лучше
- › важно понимание задачи и данных
- › важно придумывать информативные признаки
- › конкурсы по анализу данных: <http://www.kaggle.com>

Эксперименты на наборах прикладных задач:

- › цель — протестировать метод в разнообразных условиях
- › нет необходимости (и времени) разбираться в сути задач : (
- › признаки, как правило, уже кем-то придуманы
- › репозиторий UC Irvine Machine Learning Repository
<http://archive.ics.uci.edu/ml> (308 задач, 09-02-2015)

Используются для тестирования новых методов обучения.
Преимущество — мы знаем истинную $y(x)$ (ground truth)

Эксперименты на модельных (synthetic) данных:

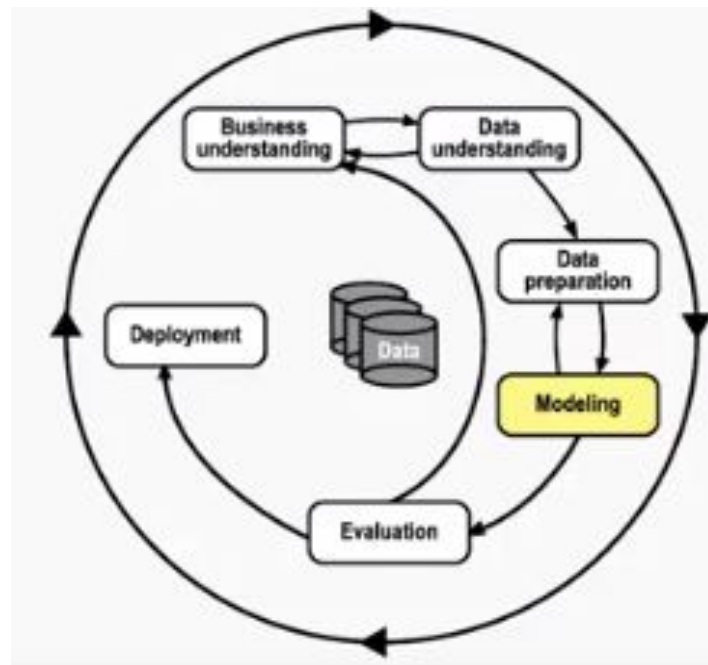
- › цель — отладить метод, выявить границы применимости
- › объекты x_i из придуманного распределения (часто 2D)
- › ответы $y_i = y(x_i)$ для придуманной функции $y(x)$
- › двумерные данные + визуализация выборки

Эксперименты на полумодельных (semi-synthetic) данных:

- › цель — протестировать помехоустойчивость модели
- › объекты x_i из реальной задачи (+ шум)
- › ответы $y_i = a(x_i)$ для полученного решения $a(x)$ (+ шум)

CRISP-DM: Cross Industry Standard Process for Data Mining

CRISP-DM — межотраслевой стандарт решения задач интеллектуального анализа данных



Резюме

Этапы решения задач машинного обучения:

- › понимание задачи и данных;
- › предобработка данных и изобретение признаков;
- › построение модели;
- › сведение обучения к оптимизации;
- › решение проблем оптимизации и переобучения;
- › оценивание качества решения;
- › внедрение и эксплуатация.

Задача классификации (обучение с учителем)

Задача восстановления зависимости $y: X \rightarrow Y$, $|Y| < \infty$
по точкам обучающей выборки (x_i, y_i) , $i = 1, \dots, \ell$.

Дано: векторы $x_i = (x_i^1, \dots, x_i^n)$ — объекты обучающей выборки,
 $y_i = y(x_i)$ — классификации, ответы учителя, $i = 1, \dots, \ell$:

$$\begin{pmatrix} x_1^1 & \dots & x_1^n \\ \dots & \dots & \dots \\ x_\ell^1 & \dots & x_\ell^n \end{pmatrix} \xrightarrow{y^*} \begin{pmatrix} y_1 \\ \dots \\ y_\ell \end{pmatrix}$$

Найти: функцию $a(x)$, способную классифицировать объекты
произвольной тестовой выборки $\tilde{x}_i = (\tilde{x}_i^1, \dots, \tilde{x}_i^n)$, $i = 1, \dots, k$:

$$\begin{pmatrix} \tilde{x}_1^1 & \dots & \tilde{x}_1^n \\ \dots & \dots & \dots \\ \tilde{x}_k^1 & \dots & \tilde{x}_k^n \end{pmatrix} \xrightarrow{a?} \begin{pmatrix} a(\tilde{x}_1) \\ \dots \\ a(\tilde{x}_k) \end{pmatrix}$$

Определение бинарного решающего дерева

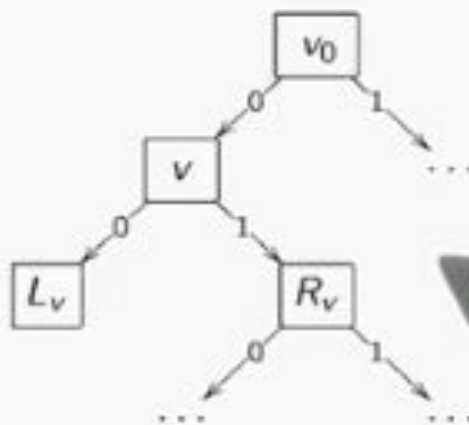
Бинарное решающее дерево — алгоритм классификации $a(x)$, задающийся бинарным деревом:

1) $\forall v \in V_{\text{внутр}} \rightarrow$ предикат $\beta_v : X \rightarrow \{0, 1\}$, $\beta_v \in \mathcal{B}$,

2) $\forall v \in V_{\text{лист}} \rightarrow$ имя класса $c_v \in Y$,

где \mathcal{B} — множество бинарных признаков или предикатов (например, вида $\beta(x) = [x^j \geq \theta_j]$, $x^j \in \mathbb{R}$)

- 1: $v := v_0$;
- 2: **пока** $v \in V_{\text{внутр}}$
- 3: **если** $\beta_v(x) = 1$ **то**
- 4: переход вправо: $v := R_v$;
- 5: **иначе**
- 6: переход влево: $v := L_v$;
- 7: **вернуть** c_v .



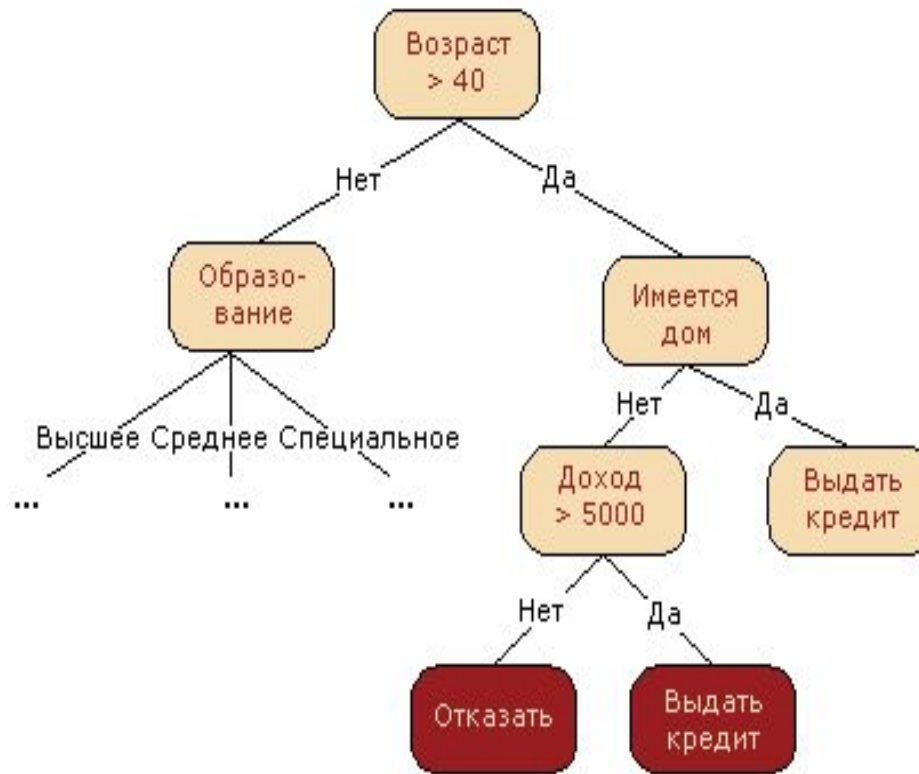


Рисунок 1