Восстановление зависимостей по эмпирическим данным

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

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

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

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

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

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 (риска | успеха | исхода).

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_i;
- документ может относиться к нескольким рубрикам;
- в каждом ребре дерева свой классификатор на 2 класса.