Машинное обучение: вводная лекция

К. В. Воронцов
vokov@forecsys.ru

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Машинное обучение (курс лекций, К.В.Воронцов)»

февраль 2012

Содержание

Задача обучения по прецедентам

X — множество *объектов*;
Y — множество *ответов*;
$y^*: X \to Y$ — неизвестная зависимость (target function).

Дано:
$\{x_1, \ldots, x_\ell\} \subset X$ — *обучающая выборка* (training sample);
$y_i = y^*(x_i), \ \ i = 1, \ldots, \ell$ — известные ответы.

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

Далее мы определим более формально,

- как задаются объекты и какими могут быть ответы;
- как строится функция a;
- что значит «a приближает y^* на всём X».

Объекты и признаки

$f_j : X \to D_j$, $j = 1, \ldots, n$ — признаки объектов.

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 $\big(f_1(x), \ldots, f_n(x)\big)$ — *признаковое описание* объекта x.

Матрица «*объекты–признаки*»

$$F = \big\|f_j(x_i)\big\|_{\ell \times n} = \begin{pmatrix} f_1(x_1) & \ldots & f_n(x_1) \\ \ldots & \ldots & \ldots \\ f_1(x_\ell) & \ldots & f_n(x_\ell) \end{pmatrix}$$

Ответы и типы задач

- $Y = \{1, \ldots, M\}$ — задача *классификации* на M непересекающихся классов.
- $Y = \{0, 1\}^M$ — задача классификации на M классов, которые могут пересекаться.
- $Y = \mathbb{R}$ — задача восстановления *регрессии*.
- Y — конечное упорядоченное множество
  — задача ранговой регрессии или ранжирования.

Модель алгоритмов

Модель алгоритмов — параметрическое семейство отображений

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

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

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

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

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

Метод обучения

Метод обучения (learning algorithm) — это отображение вида

$$\mu\colon (X \times Y)^\ell \to A,$$

которое произвольной выборке $X^\ell = (x_i, y_i)_{i=1}^\ell$
ставит в соответствие некоторый алгоритм $a \in A$.

В задачах обучения по прецедентам всегда есть два этапа:

- Этап *обучения*:
  метод μ по выборке X^ℓ строит алгоритм $a = \mu(X^\ell)$.
- Этап *применения*:
  алгоритм a для новых объектов x выдаёт ответы $y = a(x)$.

Функционалы качества

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

Наиболее очевидные функции потерь, при $Y \subseteq \mathbb{R}$:

для задач классификации

- $\mathscr{L}(a, x) = \big[a(x) \neq y^*(x)\big]$ — индикатор ошибки;

для задач регрессии

- $\mathscr{L}(a, x) = \big|a(x) - y^*(x)\big|$ — абсолютное значение ошибки;
- $\mathscr{L}(a, x) = \big(a(x) - y^*(x)\big)^2$ — квадратичная ошибка.

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

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

Сведение задачи обучения к задаче оптимизации

Метод минимизации эмпирического риска:

$$\mu(X^\ell) = \arg\min_{a \in A} Q(a, X^\ell).$$

Пример: *метод наименьших квадратов* ($Y = \mathbb{R}$, $\mathscr{L}$ квадратична):

$$\mu(X^\ell) = \arg\min_{\theta} \sum_{i=1}^{\ell} \big(g(x_i, \theta) - y_i\big)^2.$$

Проблема обобщающей способности:

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

Пример переобучения

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

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

Алгоритм полиномиальной регрессии

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

Обучение методом наименьших квадратов:

$$Q(\theta, X^\ell) = \sum_{i=1}^{\ell} (\theta_0 + \theta_1 x_i + \cdots + \theta_n x_i^n - y_i)^2 \to \min_{\theta_0,\ldots,\theta_n}.$$

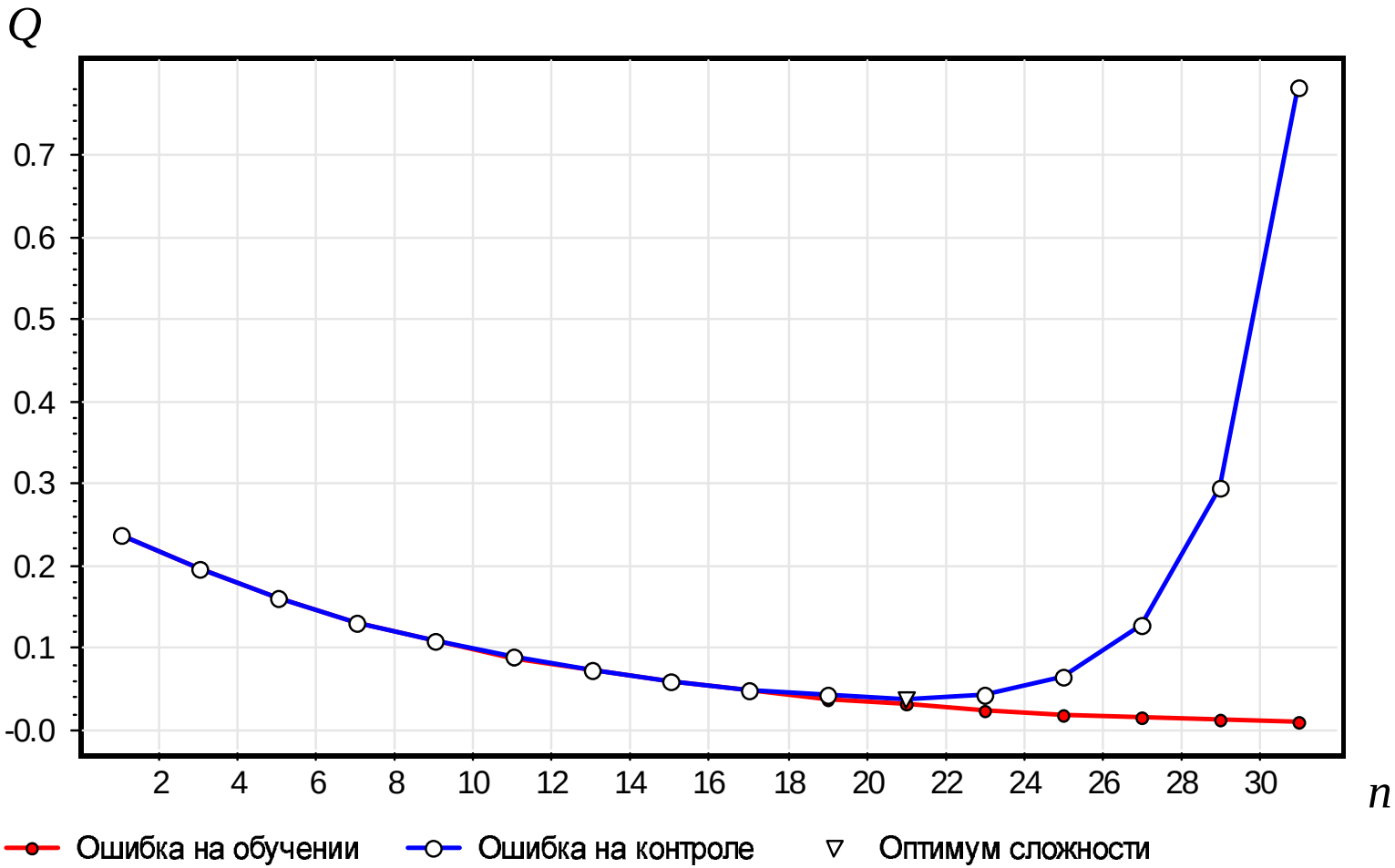
Обучающая выборка: $X^\ell = \left\{ x_i = 4\frac{i-1}{\ell-1} - 2 \;\middle|\; i = 1, \ldots, \ell \right\}$.

Контрольная выборка: $X^k = \left\{ x_i = 4\frac{i-0.5}{\ell-1} - 2 \;\middle|\; i = 1, \ldots, \ell-1 \right\}$.

Что происходит с $Q\left(\theta, X^\ell\right)$ и $Q\left(\theta, X^k\right)$ при увеличении n?

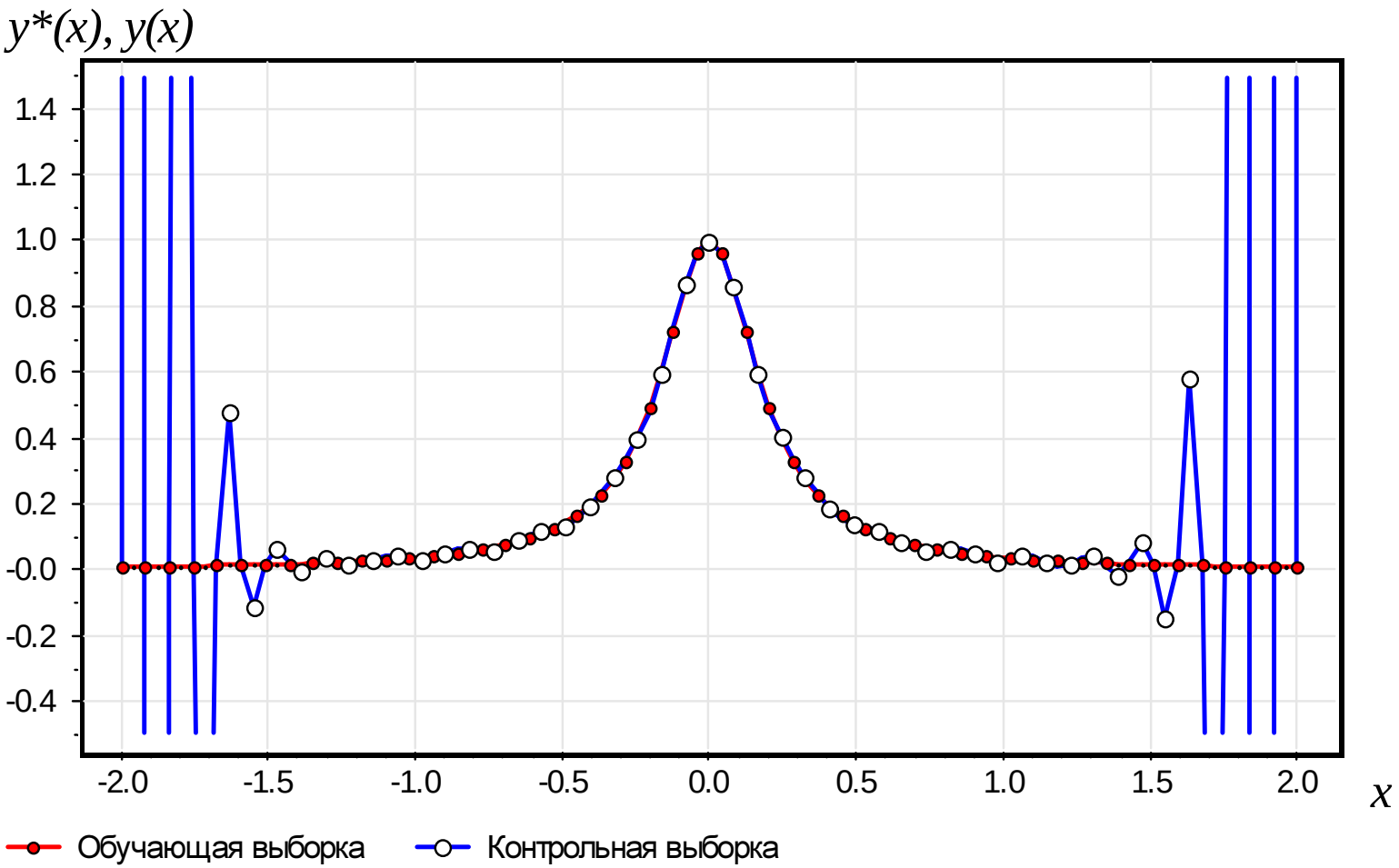
Пример переобучения: эксперимент при $\ell = 50$, $n = 1..31$

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



Пример переобучения: эксперимент при $\ell = 50$

Переобучение при степени полинома $n = 40$:



Различные формализации понятия «обобщающая способность»

- Эмпирическая оценка на отложенных данных (hold-out):

$$\mathrm{HO}(\mu, X^\ell, X^k) = Q(\mu(X^\ell), X^k) \to \min;$$

- Эмпирическая оценка скользящего контроля (cross-validation):

$$\mathrm{CV}(\mu, X^{\ell+k}) = \frac{1}{|N|} \sum_{n \in N} Q(\mu(X_n^\ell), X_n^k) \to \min;$$

- Теоретическая оценка вероятности переобучения:

$$Q_\varepsilon(\mu) = \mathrm{P}\Big[Q\big(\mu(X^\ell), X^k\big) - Q\big(\mu(X^\ell), X^\ell\big) \geqslant \varepsilon\Big] \to \min;$$

- Теоретическая оценка ожидаемой потери (переходит в вероятность ошибки, если функция потерь бинарная):

$$\mathrm{E}Q\big(\mu(X^\ell), X^k\big) \to \min;$$

Переобучение — одна из проблем машинного обучения

1. **Из-за чего возникает переобучение?**
   — избыточная сложность пространства параметров Θ,
   лишние степени свободы в модели $g(x, \theta)$ «тратятся»
   на чрезмерно точную подгонку под обучающую выборку.
   — переобучение есть всегда, когда есть оптимизация
   параметров по конечной (заведомо неполной) выборке.
2. **Как обнаружить переобучение?**
   — эмпирически, с помощью скользящего контроля.
3. **Избавиться от него нельзя. Как его минимизировать?**
   — минимизировать одну из теоретических оценок;
   — накладывать ограничения на θ (регуляризация);
   — минимизировать HoldOut или CV, но осторожно!

Задачи медицинской диагностики

Объект — пациент в определённый момент времени.

Классы: способы лечения или исходы заболевания.

Примеры признаков:

- **бинарные:** пол, наличие головной боли, слабости, тошноты, 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;
- нужен интерпретируемый алгоритм классификации;
- нужна оценка вероятности ошибки.

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 прогнозирования объёмов продаж

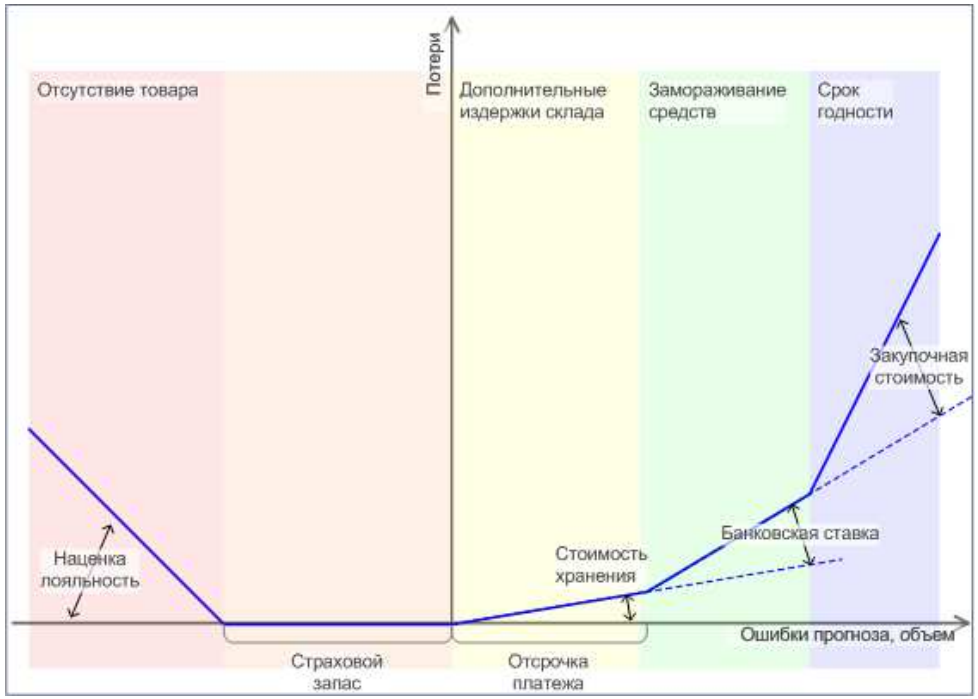
Объект — тройка $\langle$товар, магазин, день$\rangle$.

Примеры признаков:

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

Особенности задачи:

- функция потерь не квадратична и даже не симметрична;
- разреженные данные.



Задача кластеризации регионов

Объекты — регионы Российской Федерации.

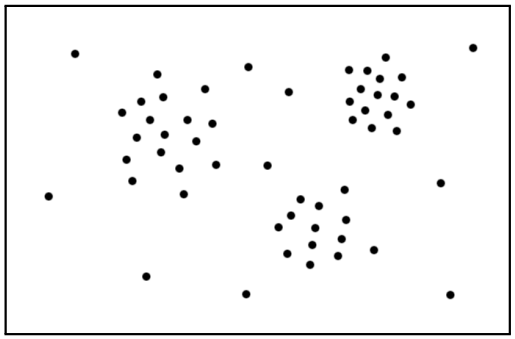
Классы y_i не заданы, неизвестно даже число классов $|Y|$.
Объекты надо сгруппировать по *сходству*.

Примеры признаков:

- **количественные:** население, доля городского населения, уровень безработицы, уровень преступности, и т. д.

Особенности задачи:

- результат зависит от того, как задать «*сходство*»;
- результат необходимо визуализировать и интерпретировать



Задача каталогизации 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.

Задача ранжирования поисковой выдачи

Объект — пара $\langle$запрос, документ$\rangle$.

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

Примеры признаков:

- **количественные:** частоты слов запроса в документе, число ссылок на документ, частота обращений к документу, и т. д.

Особенности задачи:

- минимизировать надо не число ошибок, а более сложные функционалы качества ранжирования;
- сверхбольшие выборки;
- проблема конструирования признаков по сырым данным.

Резюме в конце лекции

- **Основные понятия машинного обучения:**
  объект, ответ, признак, алгоритм, модель алгоритмов,
  метод обучения, эмпирический риск.

- **Этапы решения задач машинного обучения:**
  — понимание задачи и данных;
  — предобработка данных и изобретение признаков;
  — построение модели;
  — сведение обучения к оптимизации;
  — решение проблем переобучения и эффективности;
  — оценивание качества;
  — внедрение и эксплуатация.

- **Прикладные задачи машинного обучения:**
  очень много, очень разных,
  во всех областях бизнеса, науки, производства.