

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

1 Основные понятия и обозначения

- Данные в задачах обучения по прецедентам
- Модели алгоритмов и методы обучения
- Функционалы качества
- Обобщающая способность и переобучение

2 Примеры прикладных задач

- Задачи классификации
- Задачи регрессии
- Задачи кластеризации
- Задачи ранжирования

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 .

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

$f_j: X \rightarrow D_j, j = 1, \dots, n$ — 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 $(f_1(x), \dots, f_n(x))$ — *признаковое описание* объекта x .

Матрица «объекты–признаки»

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

Ответы и типы задач

- $Y = \{1, \dots, M\}$ — 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 $g: X \times \Theta \rightarrow Y$ — фиксированная функция,
 Θ — множество допустимых значений параметра θ .

Пример.

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

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

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

Метод обучения

Метод обучения (learning algorithm) — это отображение вида

$$\mu: (X \times Y)^\ell \rightarrow A,$$

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

$\mathcal{L}(a, x)$ — 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

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

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 :

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

Сведение задачи обучения к задаче оптимизации

Метод минимизации эмпирического риска:

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

Пример. Задача восстановления регрессии: $Y = \mathbb{R}$;

n числовых признаков $f_j: X \rightarrow \mathbb{R}$, $j = 1, \dots, n$;

линейная модель регрессии $g(x_i, \theta) = \sum_{j=1}^n \theta_j f_j(x)$, $\theta \in \mathbb{R}^n$;

квадратичная функция потерь: $\mathcal{L}(a, x) = (a(x) - y^*(x))^2$.

Тогда метод минимизации эмпирического риска совпадает с *методом наименьших квадратов*:

$$\mu(X^\ell) = \arg \min_{\theta_1 \dots \theta_n} \sum_{i=1}^{\ell} (g(x_i, \theta) - y_i)^2.$$

Обобщающая способность

Контрольная выборка — новые данные, которые не были известны на этапе обучения: $X^k = (x'_i, y'_i)_{i=1}^{\ell}$, $y'_i = y^*(x_i)$.

Переобучение — это событие $Q(\mu(X^\ell), X^k) - Q(\mu(X^\ell), X^\ell) \geq \varepsilon$.

Формализации понятия «*обобщающей способности*» метода μ :

- эмпирическая оценка на отложенных данных (hold-out):
 $HO(\mu, X^\ell, X^k) = Q(\mu(X^\ell), X^k) \rightarrow \min;$
- эмпирическая оценка скользящего контроля (cross-validation):
 $CV(\mu, X^{\ell+k}) = \frac{1}{|N|} \sum_{n \in N} Q(\mu(X_n^\ell), X_n^k) \rightarrow \min;$
- теоретическая оценка вероятности переобучения:
 $Q_\varepsilon(\mu) = P[Q(\mu(X^\ell), X^k) - Q(\mu(X^\ell), X^\ell) \geq \varepsilon] \rightarrow \min;$
- оценка ожидаемой потери (или вероятности ошибки):
 $EQ(\mu(X^\ell), X^k) \rightarrow \min;$

Пример переобучения

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

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

Алгоритм полиномиальной регрессии

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

Обучение методом наименьших квадратов:

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

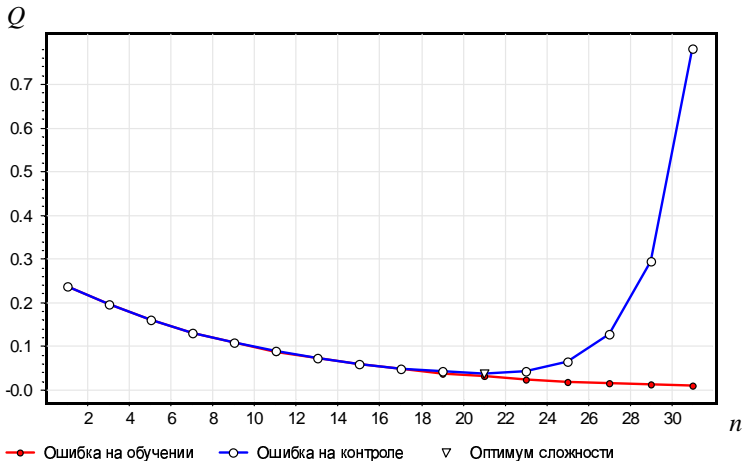
Обучающая выборка: $X^\ell = \{x_i = 4 \frac{i-1}{\ell-1} - 2 \mid i = 1, \dots, \ell\}$.

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

Что происходит с $Q(\mu(X^\ell), X^k)$ при увеличении n ?

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

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



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 оценку вероятности ошибки;
 - 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

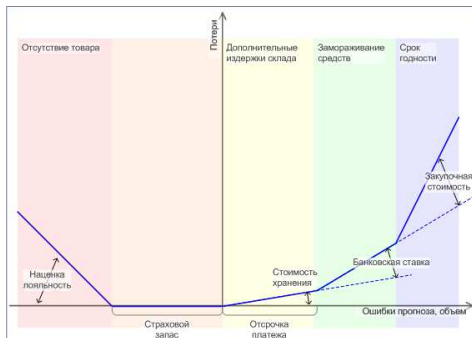
Объект — тройка ⟨товар, магазин, день⟩.

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 $\langle \text{запрос}, \text{документ} \rangle$.

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

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.

- **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.