

# Задачи и методы обучения по прецедентам

чл.-корр. РАН К. В. Рудаков,  
к.ф.-м.н. К. В. Воронцов (<http://www.ccas.ru/voron>)

3 марта 2010

## Содержание

- 1 Задачи обучения по прецедентам**
  - Основные понятия машинного обучения
  - Примеры прикладных задач обучения по прецедентам
- 2 Методология решения задач обучения по прецедентам**
  - Классическое и информационное моделирование
  - Проблема переобучения
  - Особенности реальных задач
- 3 Основные подходы и классы методов**
  - Принцип сходства
  - Минимизация эмпирического риска
  - Алгоритмические композиции
  - Поиск логических закономерностей

## Задача обучения по прецедентам (определения и обозначения)

$X$  — множество объектов;  $Y$  — множество ответов;

$\exists y: X \rightarrow Y$  — неизвестная зависимость.

**Дано:**  $X^\ell = (x_i, y_i)_{i=1}^\ell$  — обучающая выборка,  $y_i = y(x)$

$$\begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_\ell) & \dots & f_n(x_\ell) \end{pmatrix} \begin{pmatrix} y_1 \\ \dots \\ y_\ell \end{pmatrix}$$

**Построить** алгоритм — функцию  $a: X \rightarrow Y$ , приближающую неизвестную  $y(x)$  на всём множестве  $x \in X$ .

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

## Примеры прикладных задач обучения по прецедентам

- Распознавание, классификация, принятие решений ( $|Y| < \infty$ ):
  - $x$  — пациент;  $y$  — долгосрочный исход операции;
  - $x$  — заёмщик;  $y$  — вероятность дефолта;
  - $x$  — абонент;  $y$  — вероятность смены оператора связи;
  - $x$  — фотопортрет;  $y$  — идентификатор личности;
  - $x$  — фрагмент ДНК;  $y$  — функция: промотор / ген;
  - $x$  — фрагмент белка;  $y$  — тип вторичной структуры;
  - $x$  — текстовое сообщение;  $y$  — спам / не спам;
  - $x$  — документ;  $y$  — тематика (позиция в каталоге).
- Регрессия и прогнозирование ( $Y = \mathbb{R}$  или  $\mathbb{R}^m$ ):
  - $x$  — структура хим. соединения;  $y$  — его свойство;
  - $x$  — параметры технолог. процесса;  $y$  — свойство продукции;
  - $x$  — история продаж;  $y$  — прогноз потребительского спроса;
  - $x$  — характеристики недвижимости;  $y$  — цена;
  - $x$  — пара  $\langle \text{клиент, товар} \rangle$ ;  $y$  — рейтинг товара.

## Классическое и информационное моделирование

$A = \{a(x) = \varphi(x, w) : w \in \Omega\}$  — модель алгоритмов

Классическое моделирование:

- знаний о предметной области достаточно  $\Rightarrow \varphi$ ;
- число степеней свободы ( $\dim \Omega$ ) невелико;
- на выходе — достаточно сложное решение.

Информационное моделирование:

- мало знаний о предметной области;
- число степеней свободы ( $\dim \Omega$ ) велико;
- на выходе — простые решения или прогнозы.

## Проблема переобучения

$\mu: (X \times Y)^\ell \rightarrow A$  — метод обучения

$\nu(a, U)$  — частота ошибок алгоритма  $a$  на выборке  $U$

$X^k = (x'_i, y'_i)_{i=1}^k$  — независимая контрольная выборка

Теория статистического обучения (Statistical Learning Theory)  
занимается оценками вероятности переобучения:

$$P[\nu(\mu(X^\ell), X^k) - \nu(\mu(X^\ell), X^\ell) \geq \varepsilon]$$

Некоторые современные подходы к получению точных оценок:

- Margin based bounds [Bartlett, 1998]
- Rademacher complexity bounds [Koltchinskii, 2001]
- PAC-Bayes bounds [McAllester, 1999; Langford, 2005]
- Точные комбинаторные оценки [Воронцов, 2009]

## Скольльзящий контроль

Скольльзящий контроль (cross-validation):

$X^L = X_n^\ell \sqcup X_n^k$ ,  $n = 1, \dots, N$  — разбиения на обучение и контроль

$$CV(\mu, X^L) = \frac{1}{N} \sum_{n=1}^N \nu(\mu(X_n^\ell), X_n^k).$$

Стандартная методология тестирования

- $t \times q$ -fold cross-validation;
- Матрица «алгоритмы  $\times$  задачи»;
- Выделение множества проверочных задач;
- Репозиторий реальных задач UCI:  
[archive.ics.uci.edu/ml](http://archive.ics.uci.edu/ml);
- Полигон алгоритмов классификации:  
[Poligon.MachineLearning.ru](http://Poligon.MachineLearning.ru).

## Особенности реальных задач

Особенности исходных данных:

- неполные данные (пропуски);
- неточные данные (погрешности, выбросы);
- разнородные данные;
- малые выборки;
- сверхбольшие выборки;
- несбалансированные классы;

Требования к алгоритмам:

- динамическое дообучение;
- универсальные ограничения;

## Некоторые методы классификации

- kNN — метод  $k$  ближайших соседей;
- ANN — искусственные нейронные сети;
- SVM — метод опорных векторов;
- RLR — логистическая регрессия с регуляризацией;
- RBF — метод радиальных базисных функций;
- КОРА, ТЕМП, ТЕСТ — логические алгоритмы;
- CART, C4.5, ADT — решающие деревья;
- HME — иерархические смеси алгоритмов;
- AdaBoost, TreeNet — композиции решающих деревьев;
- МГУА — метод группового учёта аргументов;
- MVR Composer — индуктивное порождение моделей;

## Метод $k$ ближайших соседей и его обобщения

**Гипотеза компактности:**

Схожие объекты, как правило, лежат в одном классе.

Пусть  $\rho(x, x')$  — функция расстояния на  $X$ .

$$a(u; X^\ell) = \arg \max_{y \in Y} \sum_{y_{i(u)}=y} w(i, u);$$

где  $w(i, u)$  — вес  $i$ -го соседа для классификации объекта  $u$ ,  
 $i(u)$  — номер  $i$ -го соседа объекта  $u$ .

**Важные проблемы:**

- Выбор адекватной модели сходства  $\rho$ ;
- «Проклятие размерности».

## Задача построения разделяющей поверхности

- Задача классификации с двумя классами,  $Y = \{-1, +1\}$ :  
 $a(x, w) = \text{sign } f(x, w)$ , где  
 $f(x, w)$  — дискриминантная функция,  
 $w$  — вектор параметров.

- $f(x, w) = 0$  — разделяющая поверхность;  
 $M_i(w) = y_i f(x_i, w)$  — отступ (margin) объекта  $x_i$ ;  
 $M_i(w) < 0 \iff$  алгоритм  $a(x, w)$  ошибается на  $x_i$ .

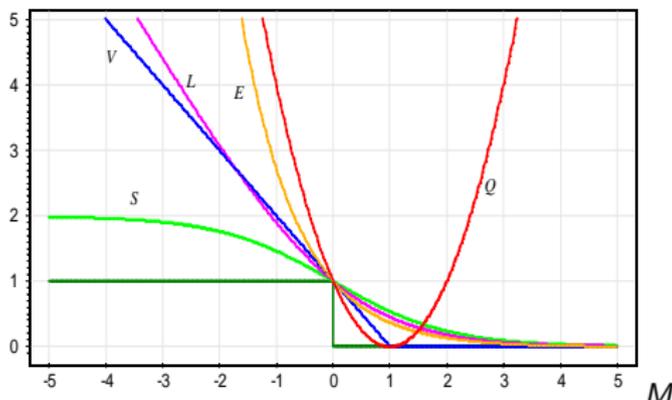
- Минимизация аппроксимированного эмпирического риска:

$$Q(w) = \sum_{i=1}^{\ell} [M_i(w) < 0] \leq \tilde{Q}(w) = \sum_{i=1}^{\ell} \mathcal{L}(M_i(w)) \rightarrow \min_w$$

функция потерь  $\mathcal{L}(M)$  невозрастающая, неотрицательная.

## Непрерывные аппроксимации пороговой функции потерь

Часто используемые функции потерь  $\mathcal{L}(M)$ :



- $Q(M) = (1 - M)^2$  — квадратичная (ЛДФ);  
 $V(M) = (1 - M)_+$  — кусочно-линейная (SVM);  
 $S(M) = 2(1 + e^M)^{-1}$  — сигмоидная (нейронные сети);  
 $L(M) = \log_2(1 + e^{-M})$  — логарифмическая (LR);  
 $E(M) = e^{-M}$  — экспоненциальная (AdaBoost).

## Метод опорных векторов SVM

Линейный классификатор:

$$a(x) = \text{sign}(\langle w, x \rangle - w_0), \quad w, x \in \mathbb{R}^n, \quad w_0 \in \mathbb{R}.$$

Задача максимизации зазора между классами приводит к аппроксимации и *регуляризации* эмпирического риска:

$$Q(w, w_0) = \sum_{i=1}^{\ell} (1 - M_i(w, w_0))_+ + \frac{1}{2C} \|w\|^2 \rightarrow \min_{w, w_0}.$$

Свойства решения этой задачи:

- $w$  зависит только от *опорных объектов*  $x_i$ ;
- $a(x)$  зависит только от  $\langle x, x_i \rangle \Rightarrow$  замена  $\langle x, x_i \rangle$  на любое неотрицательно определённое *ядро*  $K(x, x_i)$  приводит к нелинейному обобщению SVM.

## Обобщение: байесовская регуляризация

$p(x, y|w)$  — вероятностная модель данных;

$p(w; \gamma)$  — априорное распределение параметров модели;

$\gamma$  — вектор гиперпараметров;

Теперь не только появление выборки  $X^\ell$ ,  
но и появление модели  $w$  также полагается случайным.

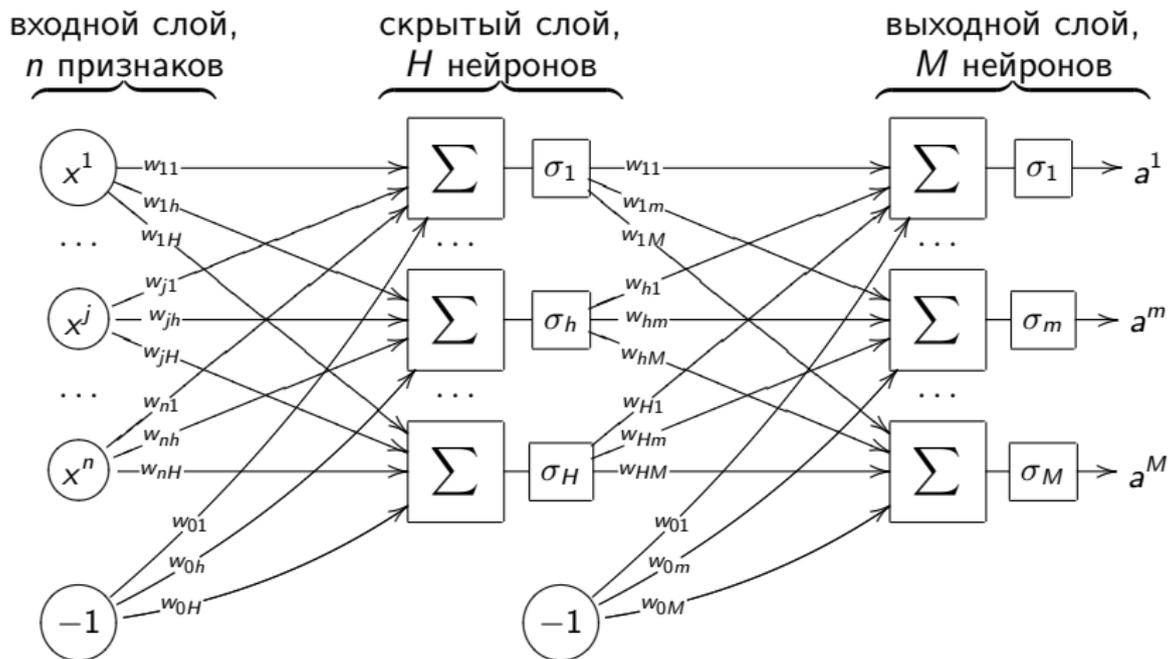
Совместное правдоподобие данных и модели:

$$p(X^\ell, w) = p(X^\ell|w) p(w; \gamma).$$

*Принцип максимума совместного правдоподобия:*

$$L(w, X^\ell) = \ln p(X^\ell, w) = \sum_{i=1}^{\ell} \ln p(x_i, y_i|w) + \underbrace{\ln p(w; \gamma)}_{\text{регуляризатор}} \rightarrow \max_w.$$

## Нейронные сети



## Стандартные градиентные методы обучения

Достоинства итерационного метода BackPropagation:

- функции потерь и архитектура сети — любые;
- вычисление градиента — эффективно;

Недостатки:

- выбор числа нейронов, начального приближения, правил остановки является искусством;

Способы устранения недостатков:

- регуляризация (против переобучения);
- удаление неинформативных связей (OBD, OBS);
- рандомизация (против «застревания» в локальных min);
- последовательное наращивание сети;

## Алгоритмические композиции

$b_t: X \rightarrow R$  — базовые классификаторы;

$F: R^T \rightarrow Y$  — корректирующая операция;

$a(x) = F(b_1(x), \dots, b_T(x))$  — композиция;

Взвешенное голосование (для случая  $Y = \{-1, +1\}$ ):

$$a(x) = \text{sign} \sum_{t=1}^T \alpha_t b_t(x);$$

Другие типы композиций:

- простое голосование (Simple Voting);
- решающий список (Decision List);
- решающее дерево (Decision Tree);
- смесь алгоритмов (Mixture of Experts);

## Бустинг (boosting)

Взвешенное голосование (для случая  $Y = \{-1, +1\}$ ):

$$a(x) = \text{sign} \sum_{t=1}^T \alpha_t b_t(x);$$

Бустинг основан на двух эвристиках:

- гладкая аппроксимация ( $\mathcal{L}(M) = e^{-M}$  — AdaBoost);
- последовательное «жадное» построение  $\alpha_t b_t$  при фиксации всех предыдущих  $\alpha_1 b_1, \dots, \alpha_{t-1} b_{t-1}$

Преимущества и особенности бустинга:

- есть доказательство сходимости;
- высокая обобщающая способность;
- бустинг над решающими деревьями — “best practice”;
- недостаток — избыточно большое  $T$ ;

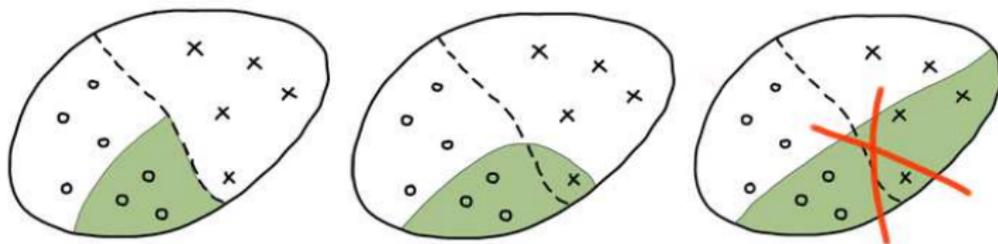
## Понятия закономерности и информативности

Закономерность класса  $y$  — это предикат  $r: X \rightarrow \{0, 1\}$ :

- $r(x)$  описывается на естественном языке;
- $r(x) = 1$  преимущественно на объектах класса  $y$ :

$$p_y(r) = \#\{x_i: r(x_i)=1 \text{ и } y_i=y\} \rightarrow \max;$$

$$n_y(r) = \#\{x_i: r(x_i)=1 \text{ и } y_i \neq y\} \rightarrow \min;$$



## Что значит «легко интерпретируемая формула»?

### Пример 1: распознавание спама

Если текст содержит «Иностранцы работают»  
и искажённый телефонный номер,  
то это спам.

### Пример 2: прогноз исхода операции

Если возраст пациента  $> 65$   
и увеличение скорости кровотока в артерии  $> 50\%$   
и перепад концентрации лимфоцитов  $< 20\%$ ,  
то высок риск повторной операции.

**Интерпретируемость:** формула  $r$  зависит от небольшого числа признаков и выражается на естественном языке.

## Виды закономерностей

Параметрическое семейство *конъюнкций пороговых термов*:

$$r(x) = \bigwedge_{j \in \omega} [\alpha_j \leq f_j(x) \leq \beta_j].$$

Конъюнкция — слишком жёсткое условие.

$$r(x) = \left[ \sum_{j \in \omega} [\alpha_j \leq f_j(x) \leq \beta_j] \geq w_0 \right],$$

где  $w_0 \in \{0, \dots, |\omega|\}$  — минимальное число выполненных термов.

Синдромы  $\omega$  объективно обнаруживаются во многих прикладных областях: медицинской диагностике, геологической разведке, кредитном скоринге и др.

## Методы поиска закономерностей

**Основная проблема** — отбор признаков  $\omega \subseteq \{1, \dots, n\}$ .  
Это задача дискретной оптимизации.

Основные методы:

- поиск в глубину, метод ветвей и границ (КОРА);
- поиск в ширину (ТЭМП);
- стохастический локальный поиск (SLIPPER);
- генетические (эволюционные) алгоритмы (DMEL);
- случайный поиск с адаптацией;

Оптимизация параметров закономерностей  $\alpha_j, \beta_j$ :

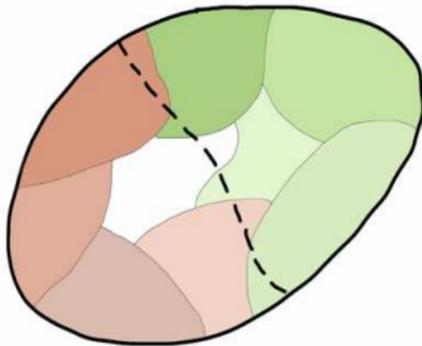
- полный перебор, выбор по критерию информативности;

## Классификатор — композиция закономерностей

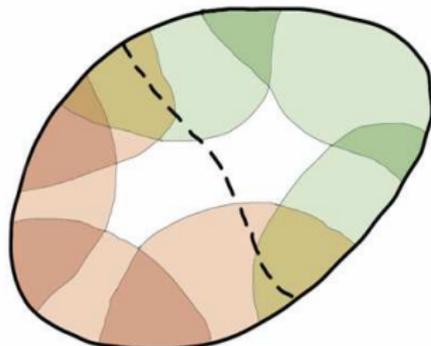
Каждое правило выделяет лишь небольшую область объектов.  
Следовательно, правил нужно много.

Две основные идеи:

покрытие  
(комитет старшинства)



голосование  
(комитет большинства)



Спасибо за внимание!

Воронцов Константин Вячеславович  
voron@forecsys.ru  
[www.MachineLearning.ru/wiki](http://www.MachineLearning.ru/wiki), «Участник:Vokov»