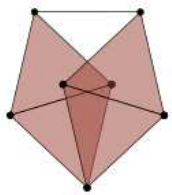
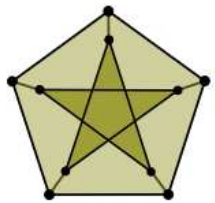
Теория и практика машинного обучения
• Лекция 1 •
Задачи и алгоритмы классификации

Воронцов Константин Вячеславович
МФТИ • МГУ • ВШЭ • ВЦ РАН • Яндекс • FORECSYS



Комбинаторика и алгоритмы
для школьников



• Летняя школа — 2015 •
18 августа 2015

Содержание

Восстановление зависимости 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* $\tilde{x}_i = (\tilde{x}_i^1, \ldots, \tilde{x}_i^n)$, $i = 1, \ldots, k$:

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

Типы признаков и типы задач

Типы признаков, $x_i^j \in D_j$, в зависимости от множества D_j:

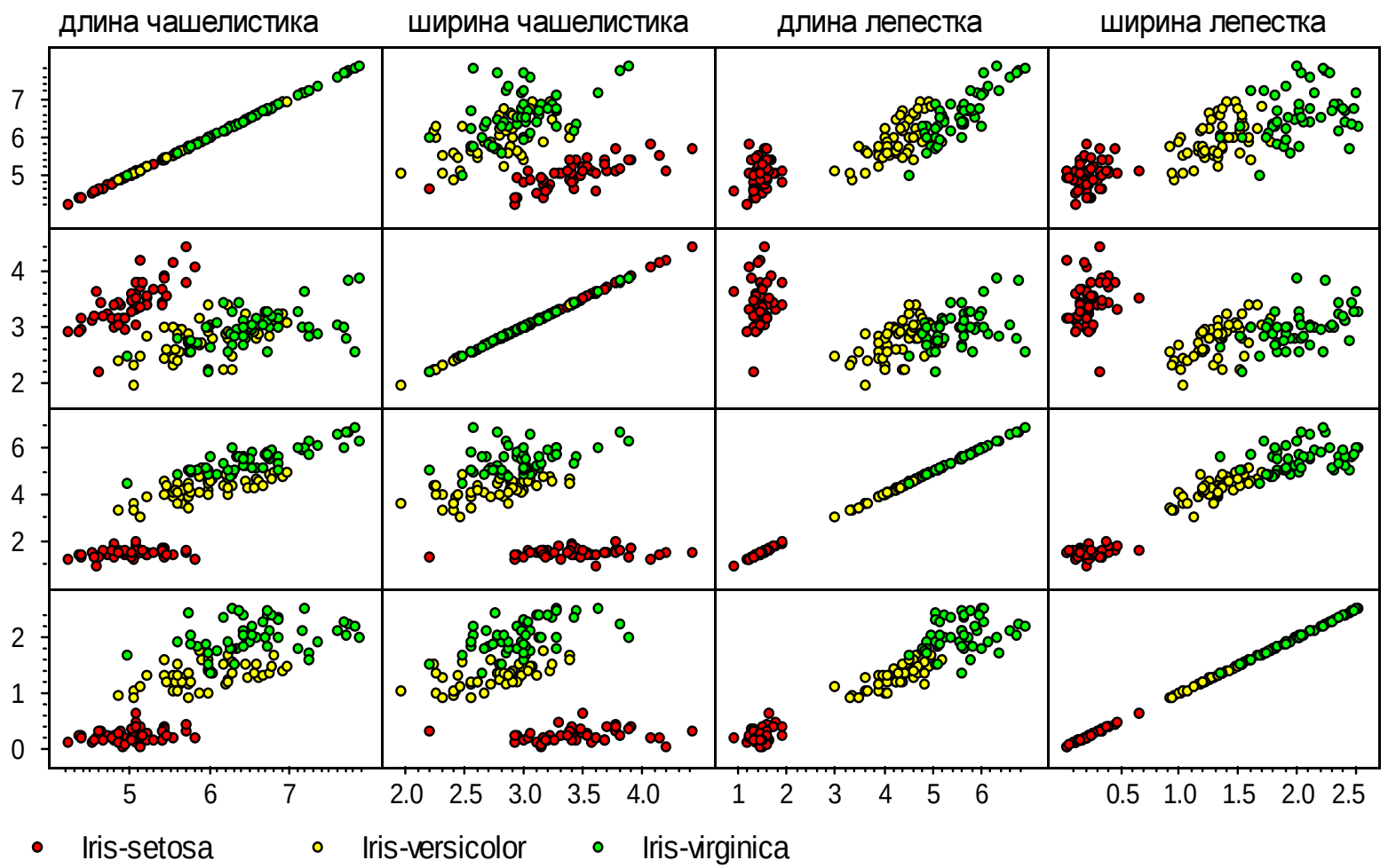
- $D_j = \{0, 1\}$ — *бинарный* признак;
- $|D_j| < \infty$ — *номинальный* признак;
- D_j упорядочено — *порядковый* признак;
- $D_j = \mathbb{R}$ — *количественный* признак.

Типы задач, $y_i \in Y$, в зависимости от множества Y:

- $Y = \{0, 1\}$ или $Y = \{-1, +1\}$ — *классификация* на 2 класса;
- $Y = \{1, \ldots, M\}$ — на M непересекающихся классов;
- $Y = \{0, 1\}^M$ — на M классов, которые могут пересекаться;
- $Y = \mathbb{R}$ — задача восстановления регрессии;
- Y упорядочено — задача ранжирования (learning to rank).

Задача классификации цветков ириса [Фишер, 1936]

$n = 4$ признака, $|Y| = 3$ класса, длина выборки $\ell = 150$.



Задачи медицинской диагностики

Объект — пациент в определённый момент времени.

Классы: диагнозы или способы лечения или исходы.

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;
- нужен интерпретируемый алгоритм классификации;
- нужна оценка вероятности ошибки.

Задача кредитного скоринга

Объект — заявка на выдачу 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 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.

Задача регрессии: прогноз стоимости недвижимости

Объект — квартира в Москве.

Примеры признаков:

- **бинарные:** наличие балкона, лифта, мусоропровода, охраны, гаража, чердака, 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 ранжирования поисковой выдачи

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

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

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:

- нужно строить признаки по разнородным сырым данным;
- оптимизируется не число ошибок, а качество ранжирования;
- сверхбольшие выборки.

Задача ранжирования в рекомендательных системах

Объект — пара $\langle$клиент, товар$\rangle$
(товары — книги, фильмы, музыка).

Предсказать: вероятность покупки или рейтинг товара.

Примеры признаков:

- **количественные:**
  частота покупок или средний рейтинг схожих товаров для данного клиента;
  частота покупок или средний рейтинг данного товара для схожих клиентов;
  оценки интересов клиента;
  оценки интересов товара;

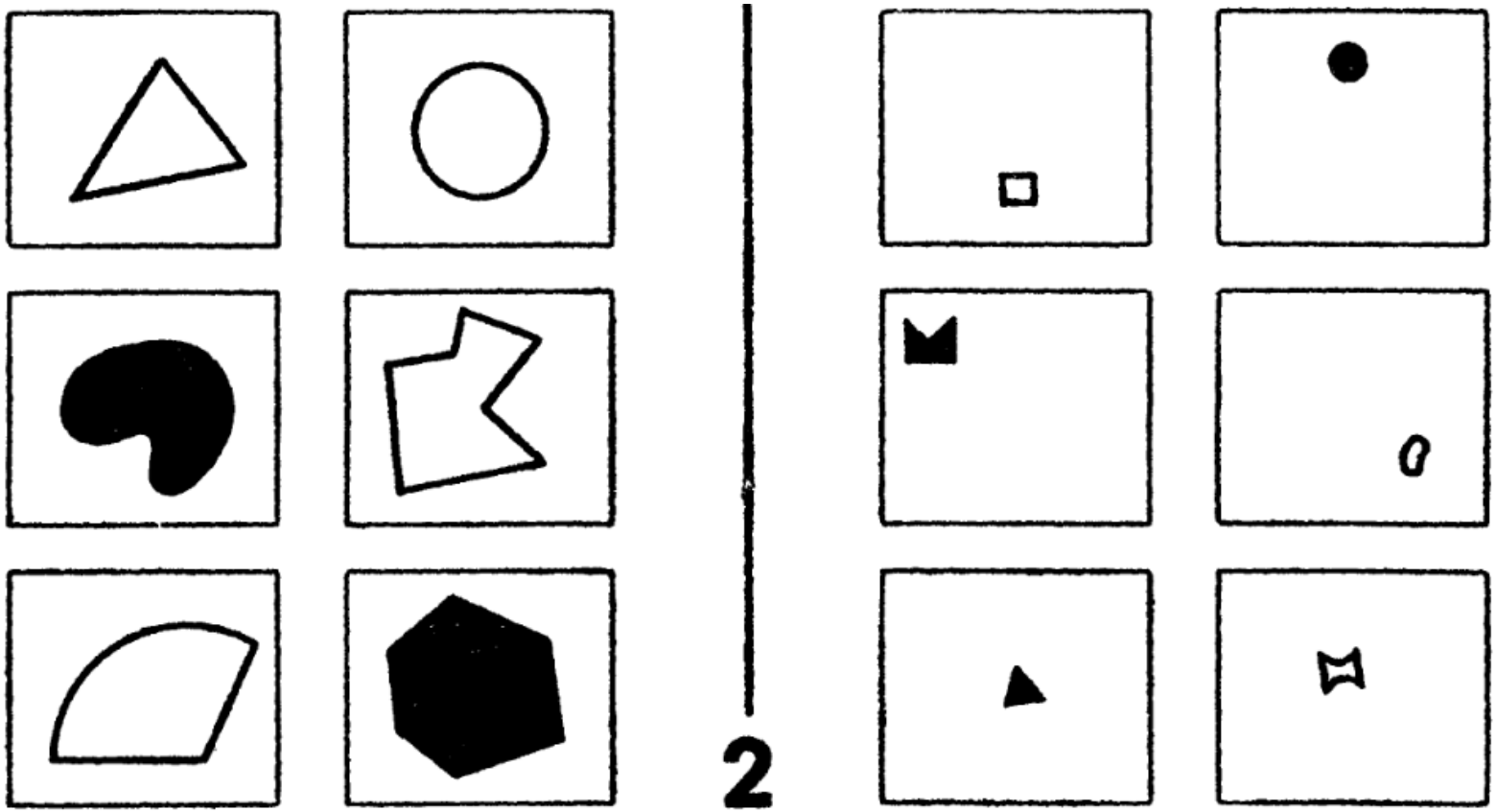
Особенности задачи:

- сверхбольшие разреженные данные;
- интересы скрыты, их надо сначала выявить.

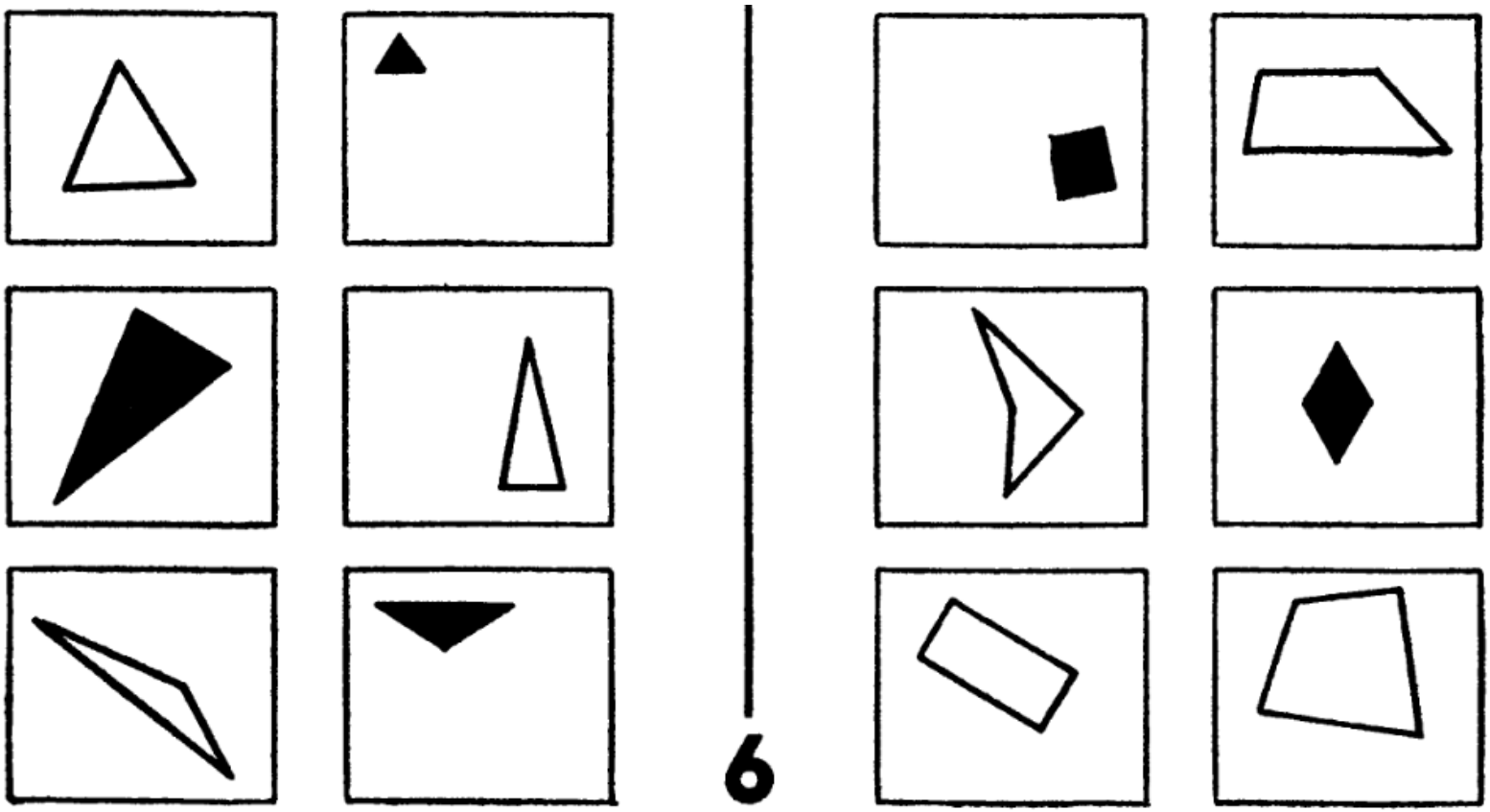
Примеры прикладных задач обучения по прецедентам

- Распознавание, классификация, принятие решений ($|\mathbb{Y}| < \infty$):
  - x — пациент; y — диагноз, рекомендуемая терапия;
  - x — заёмщик; y — вероятность дефолта;
  - x — геологич. объект; y — наличие полезного ископаемого;
  - x — абонент; y — вероятность ухода к другому оператору;
  - x — текстовое сообщение; y — спам / не спам;
  - x — документ; y — категория в рубрикаторе;
  - x — фрагмент белка; y — тип вторичной структуры;
  - x — фрагмент ДНК; y — функция: промотор / ген;
  - x — фотопортрет; y — идентификатор личности;
- Регрессия и прогнозирование ($\mathbb{Y} = \mathbb{R}$ или $\mathbb{R}^m$):
  - x — история продаж; y — прогноз объёма продаж;
  - x — пара $\langle$клиент,товар$\rangle$; y — рейтинг товара;
  - x — параметры технолог. процесса; y — свойство продукции;
  - x — структура хим. соединения; y — его свойство;
  - x — характеристики недвижимости; y — цена;

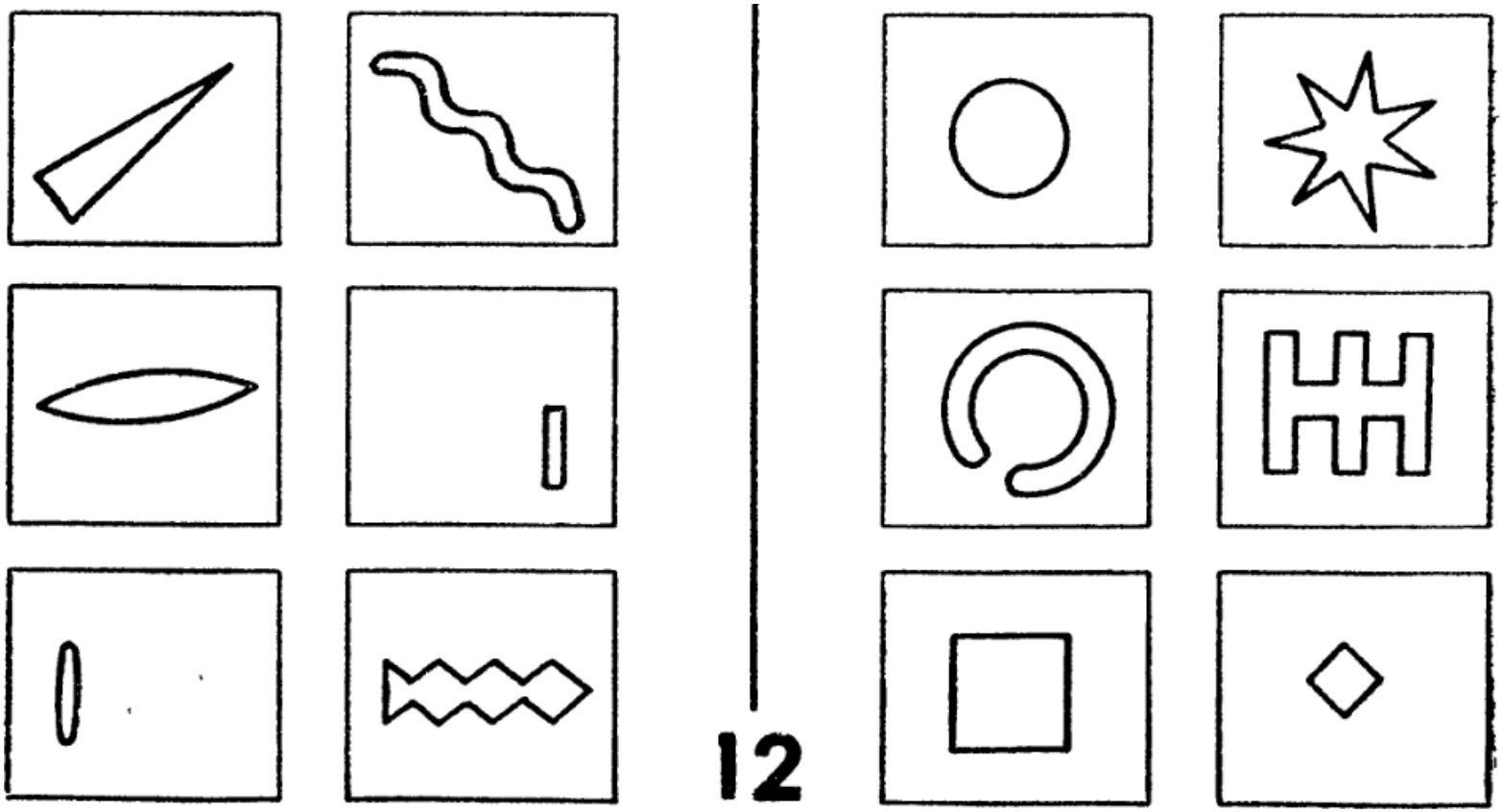
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



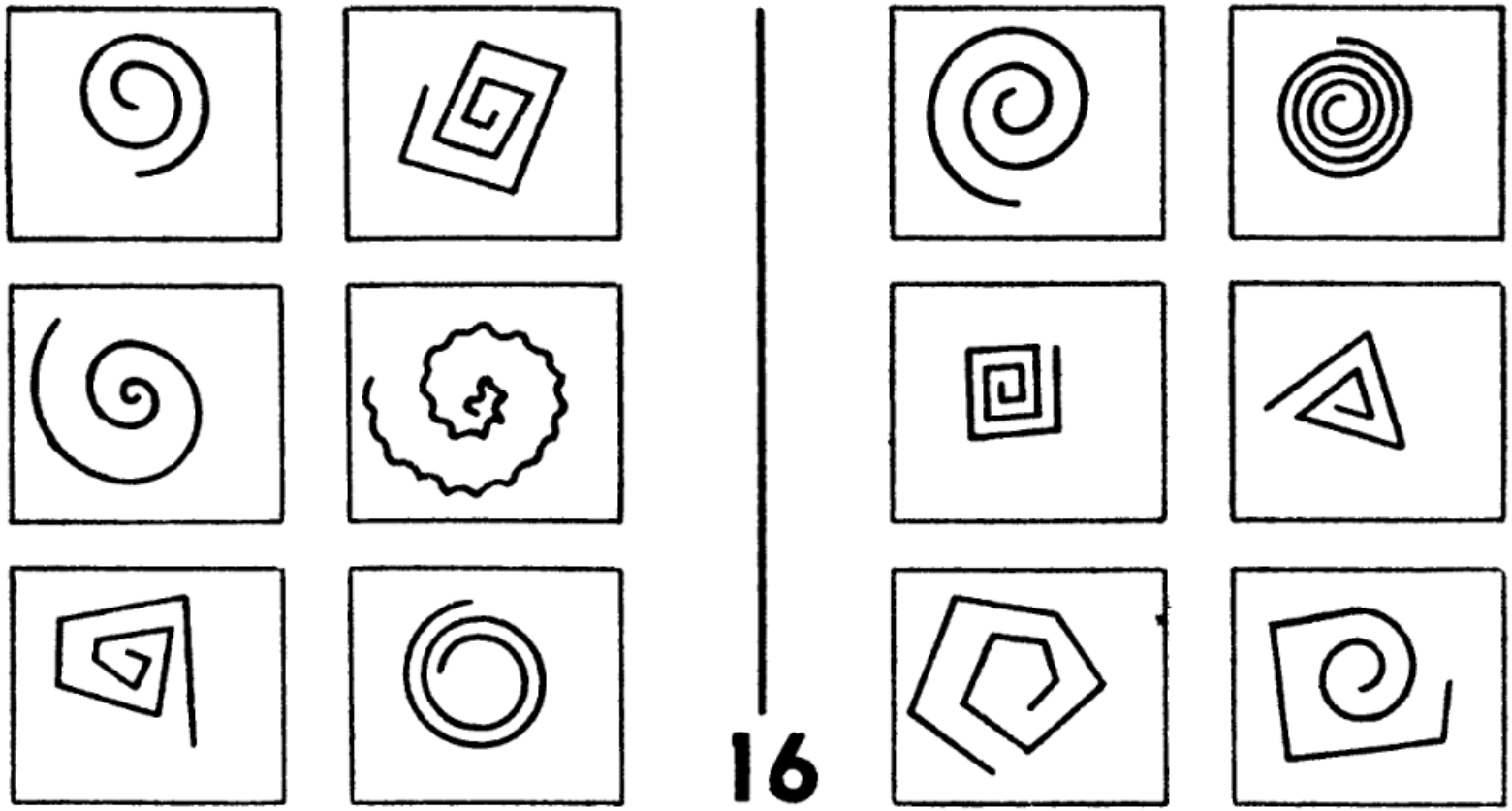
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



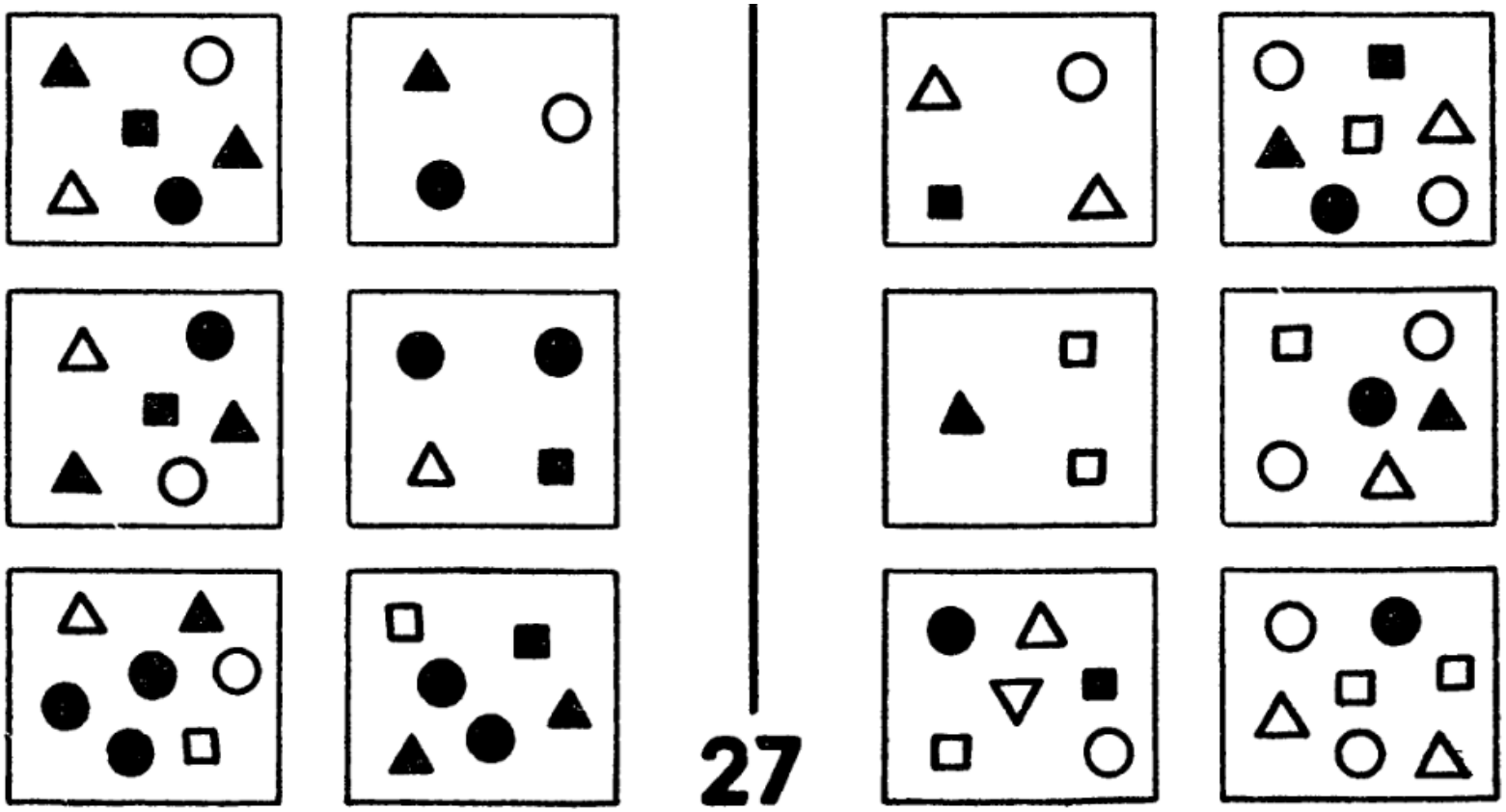
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



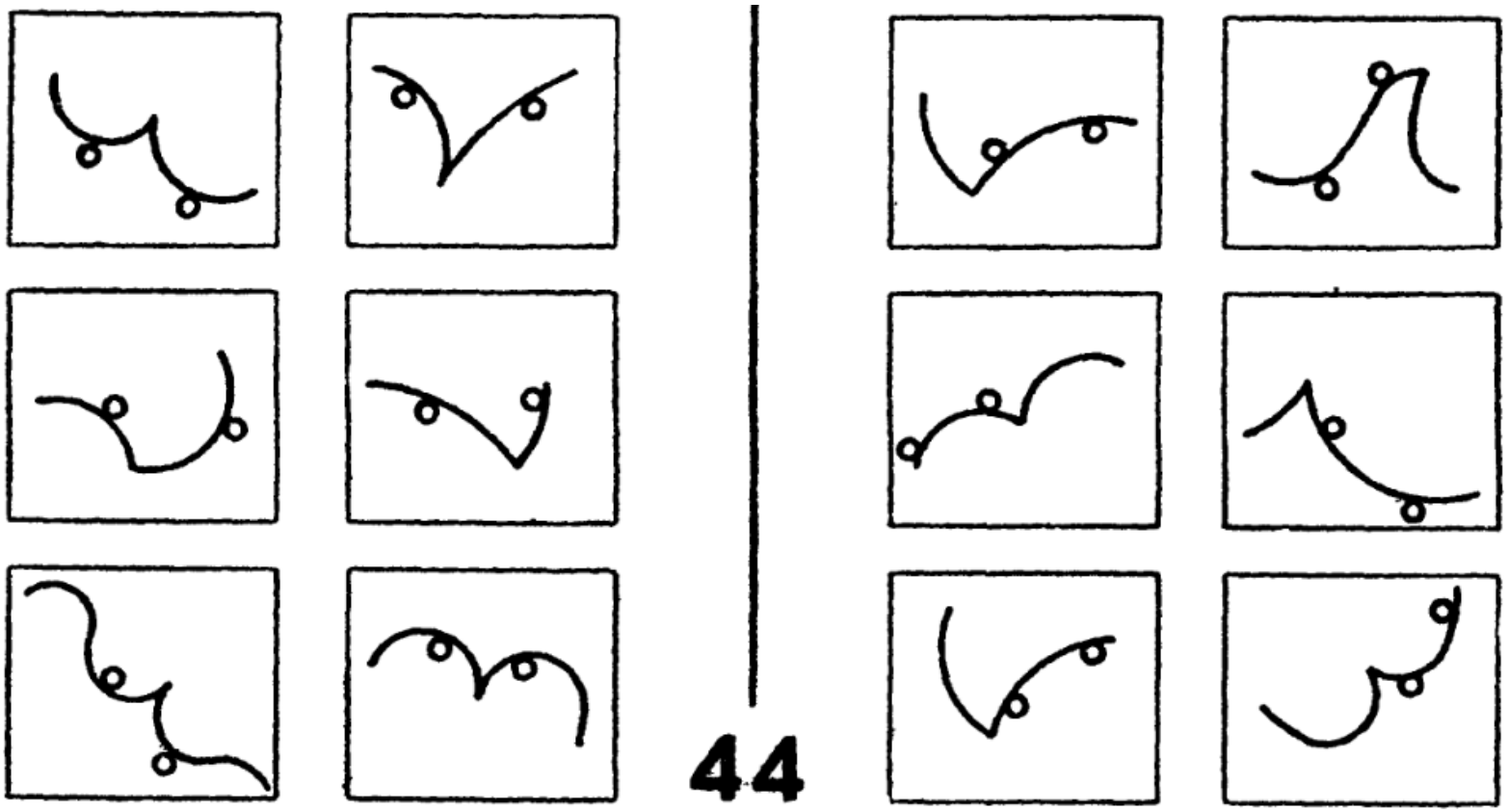
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



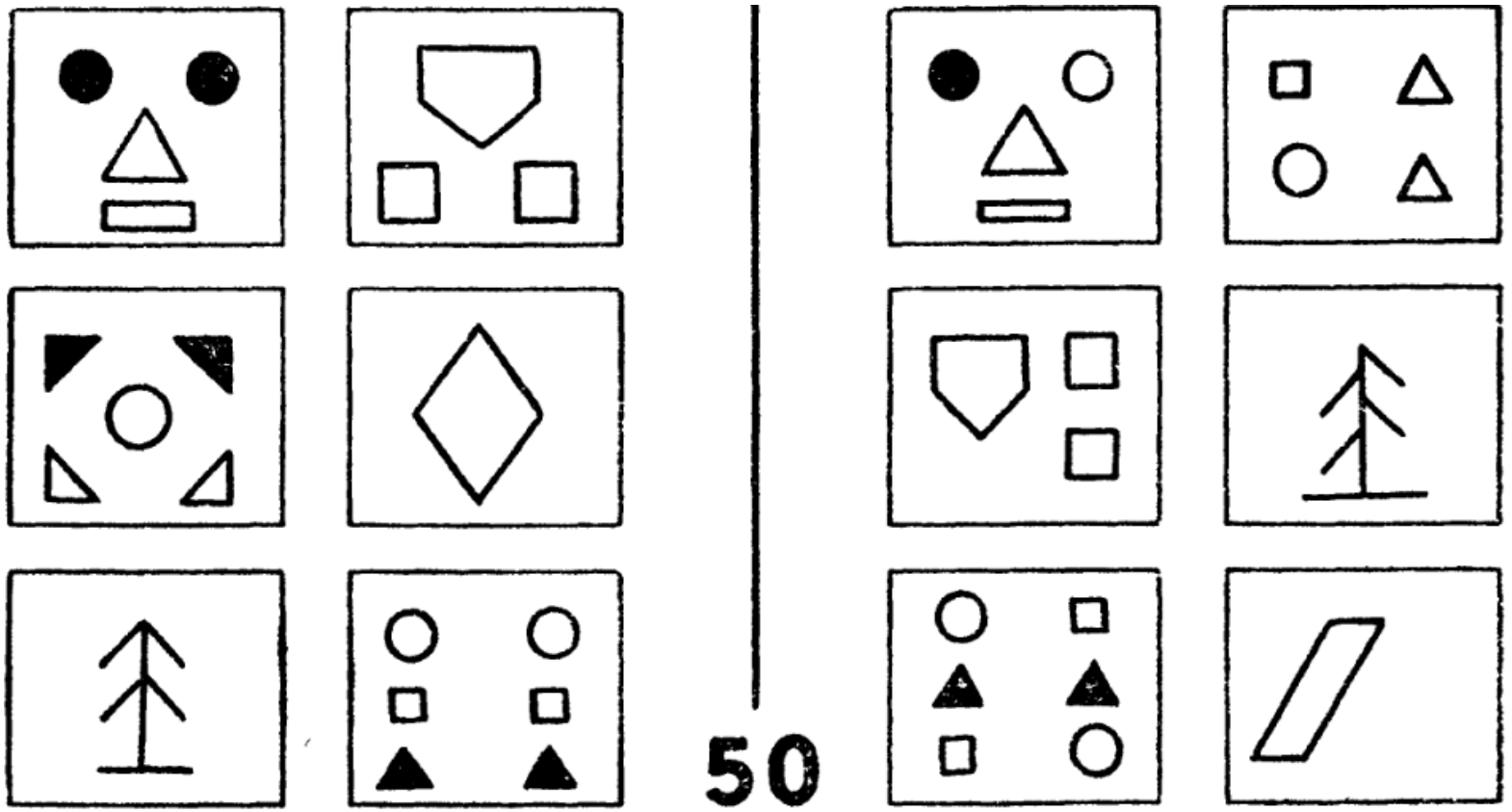
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



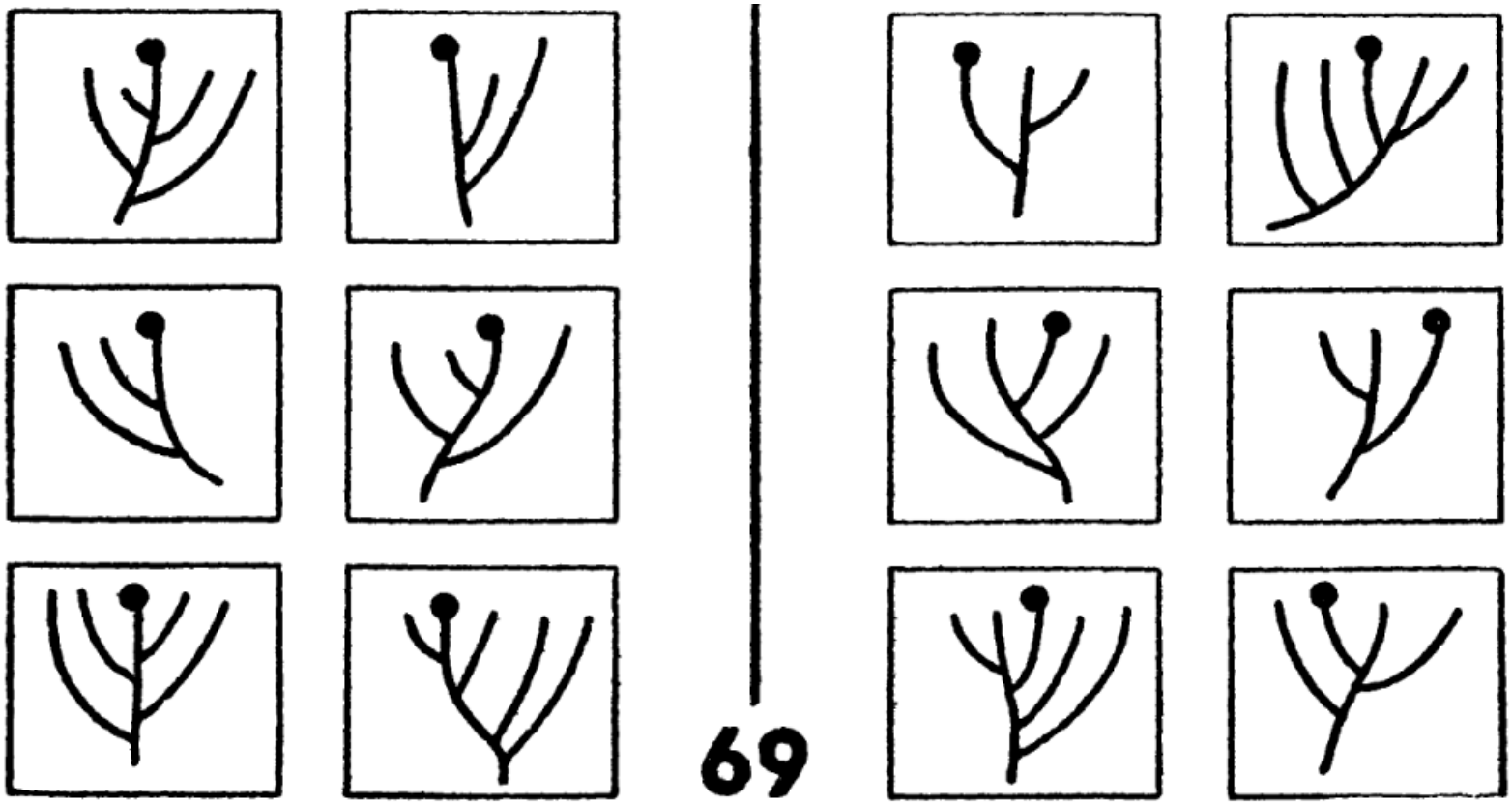
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



Тесты М. М. Бонгарда [Проблема узнавания, 1967]



Тесты М. М. Бонгарда [Проблема узнавания, 1967]



Этапы решения задач машинного обучения

1. Понимание задачи и источника «сырых» данных
2. Измерение (изобретение, извлечение) признаков
3. Изобретение параметрической модели зависимости
4. Оптимизация (обучение) параметров модели по данным
5. Проверка качества предсказаний на тестовой выборке
6. Если «всё плохо», то GOTO 1 или 2 или 3 или 4

Пример задачи: распознавание языка текста

Объект — текст.
Классы — языки.
Признаки — ?

Декларация прав человека. На каких языках?

Статья 1. Все люди рождаются свободными и равными в своем достоинстве и правах. Они наделены разумом и совестью и должны поступать в отношении друг друга в духе братства.

Стаття 1. Всі люди народжуються вільними і рівними у своїй гідності та правах. Вони наділені розумом і совістю і повинні діяти у відношенні один до одного в дусі братерства.

Article 1. All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.

Article 1. Tous les êtres humains naissent libres et égaux en dignité et en droits. Ils sont doués de raison et de conscience et doivent agir les uns envers les autres dans un esprit de fraternité.

Декларация прав человека. На каких языках?

rus: Russian

Статья 1. Все люди рождаются свободными и равными в своем достоинстве и правах. Они наделены разумом и совестью и должны поступать в отношении друг друга в духе братства.

ukr: Ukrainian

Стаття 1. Всі люди народжуються вільними і рівними у своїй гідності та правах. Вони наділені розумом і совістю і повинні діяти у відношенні один до одного в дусі братерства.

eng: English

Article 1. All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.

frn: French

Article 1. Tous les êtres humains naissent libres et égaux en dignité et en droits. Ils sont doués de raison et de conscience et doivent agir les uns envers les autres dans un esprit de fraternité.

Декларация прав человека. На каких языках?

Artikel 1. Alle Menschen sind frei und gleich an Würde und Rechten geboren. Sie sind mit Vernunft und Gewissen begabt und sollen einander im Geiste der Brüderlichkeit begegnen.

Artikel 1. Alle menslike wesens word vry, met gelyke waardigheid en regte, gebore. Hulle het rede en gewete en behoort in die gees van broederskap teenoor mekaar op te tree.

Artículo 1. Todos los seres humanos nacen libres e iguales en dignidad y derechos y. Dotados como están de razón y conciencia, deben comportarse fraternalmente los unos con los otros.

Artigo 1. Todos os seres humanos nascem livres e iguais em dignidade e em direitos. Dotados de razão e de consciência, devem agir uns para com os outros em espírito de fraternidade.

Декларация прав человека. На каких языках?

ger: German

Artikel 1. Alle Menschen sind frei und gleich an Würde und Rechten geboren. Sie sind mit Vernunft und Gewissen begabt und sollen einander im Geiste der Brüderlichkeit begegnen.

afk: Afrikaans

Artikel 1. Alle menslike wesens word vry, met gelyke waardigheid en regte, gebore. Hulle het rede en gewete en behoort in die gees van broederskap teenoor mekaar op te tree.

spn: Spanish

Artículo 1. Todos los seres humanos nacen libres e iguales en dignidad y derechos y. Dotados como están de razón y conciencia, deben comportarse fraternalmente los unos con los otros.

por: Portuguese

Artigo 1. Todos os seres humanos nascem livres e iguais em dignidade e em direitos. Dotados de razão e de consciência, devem agir uns para com os outros em espírito de fraternidade.

Декларация прав человека. На каких языках?

Artikla 1. Kaikki ihmiset syntyvät vapaina ja tasavertaisina arvoltaan ja oikeuksiltaan. Heille on annettu järki ja omatunto, ja heidän on toimittava toisiaan kohtaan veljeyden hengessä.

Artikkel 1. Kõik inimesed sünnivad vabadena ja võrdsetena oma väärikuselt ja õigustelt. Neile on antud mõistus ja südametunnistus ja nende suhtumist üksteisesse peab kandma vendluse vaim.

Artikel 1. Alla människor är födda fria och lika i värde och rättigheter. De har utrustats med förnuft och samvete och bör handla gentemot varandra i en anda av gemenskap.

Artikkel 1. Alle menneske er fødde til fridom og med same menneskeverd og menneskerettar. Dei har fått fornuft og samvit og skal leve med kvarandre som brør.

Декларация прав человека. На каких языках?

fin: Finnish

Artikla 1. Kaikki ihmiset syntyvät vapaina ja tasavertaisina arvoltaan ja oikeuksiltaan. Heille on annettu järki ja omatunto, ja heidän on toimittava toisiaan kohtaan veljeyden hengessä.

est: Estonian

Artikkel 1. Kõik inimesed sünnivad vabadena ja võrdsetena oma väärikuselt ja õigustelt. Neile on antud mõistus ja südametunnistus ja nende suhtumist üksteisesse peab kandma vendluse vaim.

swd: Swedish

Artikel 1. Alla människor är födda fria och lika i värde och rättigheter. De har utrustats med förnuft och samvete och bör handla gentemot varandra i en anda av gemenskap.

nrn: Norwegian

Artikkel 1. Alle menneske er fødde til fridom og med same menneskeverd og menneskerettar. Dei har fått fornuft og samvit og skal leve med kvarandre som brør.

Задача «Language Identification»

Как обучить машину определять язык текста автоматически?

Зачем это нужно:

- Поисковые системы
- Системы агрегации контента
- Системы машинного перевода

Определения и обозначения

x_i — обучающая выборка текстов, $i = 1, \ldots, \ell$,
$y_i \in \{0, 1\}$ — два класса: «язык 0» и «язык 1»
(будем сравнивать языки попарно).

Векторизация текста — это преобразование текста
(символьной последовательности произвольной длины)
в числовой вектор признаков фиксированной размерности.

Основной принцип векторизации:
признаки должны содержать важную информацию о классах.

N-грамма — подстрока текста из N последовательных букв,
K — число букв в алфавите, $n = K^N$ — число N-грамм,

x_i^j — частота N-граммы j в тексте x_i, $j = 1 \ldots n$, $i = 1 \ldots \ell$,

$\left(x_i^j\right)_{j=1}^n$ — n-мерный вектор признаков текста x_i.

Модель классификации текстов

Основное предположение:
каждый язык имеет уникальное распределение частот N-грамм

Линейная модель классификации:

$$a(x) = \big[\langle x, w\rangle \geqslant w_0\big], \qquad \langle x, w\rangle = \sum_{j=1}^{n} w_j x^j,$$

x^j — частота N-граммы j в тексте x,
w_j — вес N-граммы j:

- $w_j > 0$, N-грамма более специфична для языка 1
- $w_j < 0$, N-грамма более специфична для языка 0
- $w_j = 0$, N-грамма не различает эти языки

Методы *машинного обучения* позволяют настраивать веса w_j по обучающей выборке автоматически

Методы обучения линейных классификаторов

Линейная модель классификации:

$$a(x) = \big[\langle x, w\rangle \geqslant w_0\big], \qquad \langle x, w\rangle = \sum_{j=1}^{n} w_j x^j,$$

Методы обучения весов w_j в линейных классификаторах

- SVM — Support Vector Machine
- LR — Logistic Regression
- RLR — Regularized Logistic Regression
- LASSO — Least Absolute Shrinkage and Selection Operator
- NB — Naïve Bayes Classifier
- и др.

Но в данной задаче простые *эвристики* уже работают хорошо.

Простые эвристики для выбора весов

Средняя частота N-граммы j в текстах класса y:

$$S_y^j = \frac{1}{\ell_y} \sum_{i=1}^{\ell} [y_i = y] x_i^j, \quad \ell_y = \sum_{i=1}^{\ell} [y_i = y]$$

Эвристика: вес N-граммы j должен быть тем больше, чем больше S_1^j и чем меньше S_0^j

Можно пробовать разные формулы для весов:

$$w_j = \frac{S_1^j + \gamma}{S_0^j + \gamma} \qquad w_j = \log \frac{S_1^j + \gamma}{S_0^j + \gamma}$$

$$w_j = \sqrt{S_1^j} - \sqrt{S_0^j} \qquad w_j = \sqrt{S_1^j/\ell_1} - \sqrt{S_0^j/\ell_0}$$

... и разрешается фантазировать!

Эксперимент с текстами Декларации прав человека

Цели эксперимента — проверить:

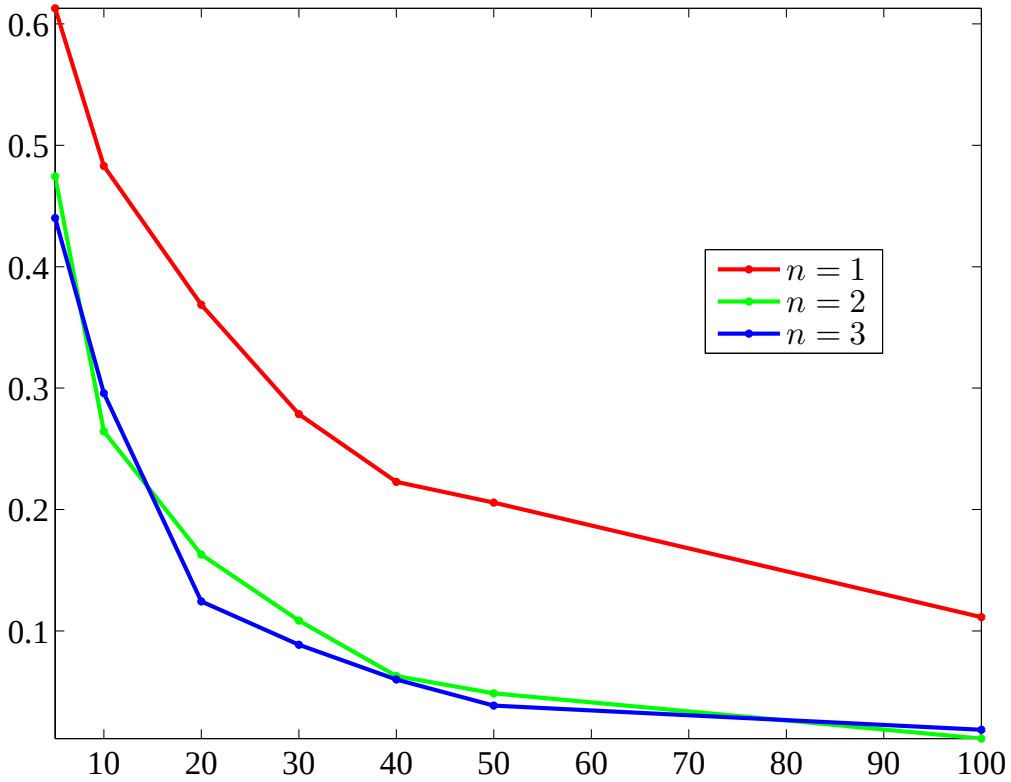
- действительно ли частоты триграмм распознают язык?
- как точность распознавания зависит от длины текста?

Методика эксперимента:

- используем тексты Декларации на 7 языках ($\sim 10^4$ букв)
- в каждом тексте 100 раз случайным образом выбираем обучающий фрагмент длины ℓ и не пересекающийся с ним контрольный фрагмент длины k
- коэффициенты w_j определяем по обучающему фрагменту текста длины ℓ для каждого языка
- точность измеряем как долю ошибочных распознаваний языка по контрольным фрагментам

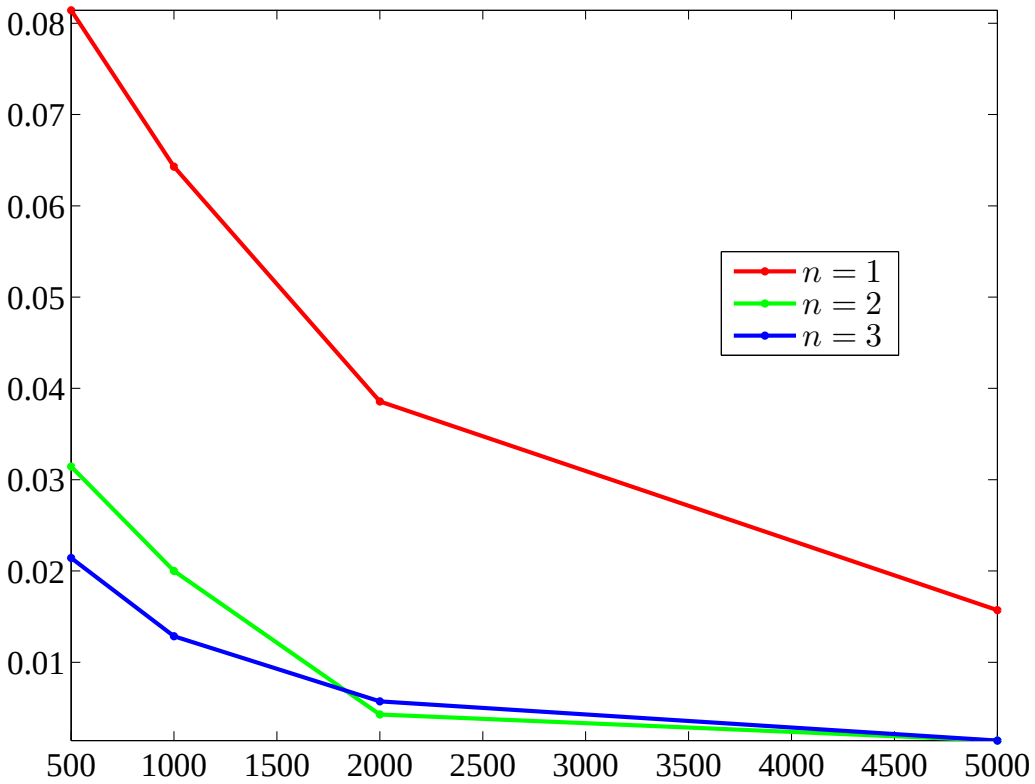
Результаты эксперимента

Зависимость доли ошибок на контроле
от длины контрольных текстов для $N = 1, 2, 3$-грамм
(длина обучающих текстов $\ell = 2000$ символов)



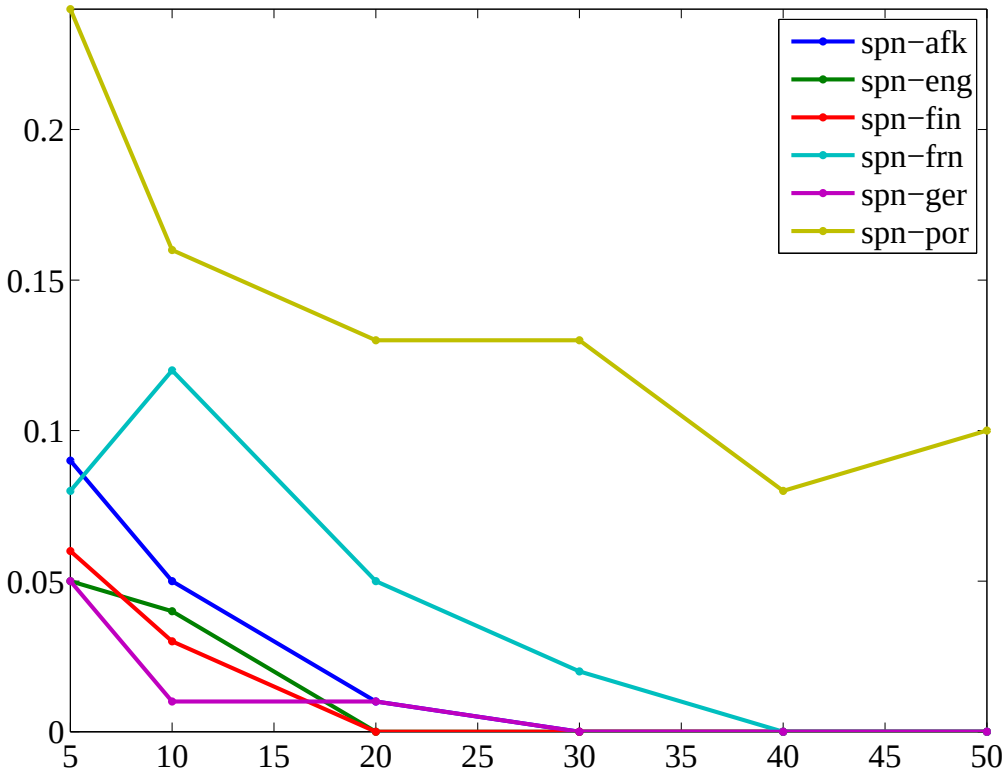
Результаты эксперимента

Зависимость доли ошибок на контроле
от длины обучающих текстов для $N = 1, 2, 3$-грамм
(длина контрольных текстов $k = 200$ символов)



Результаты эксперимента

По оси X — длина контрольной выборки
По оси Y — доля случаев, когда испанский язык был перепутан с другим языком
(3-граммы, длина обучающих текстов 2000 символов)



Результаты эксперимента

В столбцах — истинные классы, в строках — предсказанные
В ячейках — число случаев (из общего числа $100 \cdot 7 = 700$)
(3-граммы, длина обучения 2000, длина контроля 10)

	afk	eng	fin	frn	ger	por	spn
afk	69	7	9	6	14	1	5
eng	6	75	3	4	2	1	4
fin	4	1	82	3	0	1	3
frn	3	4	1	66	1	5	12
ger	15	4	1	4	80	1	1
por	1	6	2	5	1	62	16
spn	2	3	2	12	2	29	59

Результаты эксперимента

В столбцах — истинные классы, в строках — предсказанные
В ячейках — число случаев (из общего числа $100 \cdot 7 = 700$)
(3-граммы, длина обучения 2000, длина контроля 50)

	afk	eng	fin	frn	ger	por	spn
afk	100	0	0	0	0	0	0
eng	0	98	0	1	1	0	0
fin	0	0	100	0	0	0	0
frn	0	1	0	98	1	1	0
ger	0	1	0	0	98	0	0
por	0	0	0	0	0	89	10
spn	0	0	0	1	0	10	90

Результаты эксперимента

В столбцах — истинные классы, в строках — предсказанные
В ячейках — число случаев (из общего числа $100 \cdot 7 = 700$)
(3-граммы, длина обучения 2000, длина контроля 100)

	afk	eng	fin	frn	ger	por	spn
afk	100	0	0	0	0	0	0
eng	0	100	0	0	0	0	0
fin	0	0	100	0	0	0	0
frn	0	0	0	100	0	0	0
ger	0	0	0	0	100	0	0
por	0	0	0	0	0	92	5
spn	0	0	0	0	0	8	95

Результаты эксперимента

В столбцах — истинные классы, в строках — предсказанные
В ячейках — число случаев (из общего числа $100 \cdot 7 = 700$)
(3-граммы, длина обучения 2000, длина контроля 1000)

	afk	eng	fin	frn	ger	por	spn
afk	100	0	0	0	0	0	0
eng	0	100	0	0	0	0	0
fin	0	0	100	0	0	0	0
frn	0	0	0	100	0	0	0
ger	0	0	0	0	100	0	0
por	0	0	0	0	0	99	0
spn	0	0	0	0	0	1	100

20 самых частых триграмм в 7 языках

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
afk	ie-	die	-di	en-	ing	ng-	an-	et-	-re	reg	eg-	e-r	-en	nie	van	-ni	een	el-	e-o	n-h
eng	-an	and	nd-	the	-th	he-	ion	of-	-of	tio	al-	to-	-to	on-	ent	ati	-in	e-e	ll-	t-t
fin	ise	sta	an-	en-	ta-	ais	aan	la-	ell	ist	ike	kai	keu	oik	-ta	lla	on-	tai	-oi	ast
frn	-de	es-	de-	le-	et-	ion	nt-	tio	-et	te-	ent	e-d	e-p	ne-	on-	ati	a-d	e-s	la-	oit
ger	en-	ein	er-	der	ine	nd-	cht	ung	-un	ich	und	ech	gen	ht-	ng-	sei	ver	-ei	-ha	-se
por	de-	-de	os-	-e-	em-	o-d	to-	-a-	-di	dir	-co	-pe	ire	as-	ito	o-e	-se	eit	ess	e-d
spn	os-	-de	-la	de-	la-	-y-	es-	-a-	ent	ien	en-	al-	as-	ere	e-l	-el	-lo	cia	el-	los

Упрощения, сделанные в этом эксперименте:

- использовались только языки на основе латиницы,
- все диакритические знаки и пробел были заменены на «-»,
- использовались только триграммные признаки,
- использовалась только линейная модель.

Выводы

- Языки можно распознавать автоматически,
- с очень высокой надёжностью,
- используя частоты триграмм или биграмм,
- причём точность распознавания быстро увеличивается с ростом длины контрольного текста, и сотни символов уже хватает для распознавания даже близких языков.
- Для распознавания языка коротких текстов (SMS, twitter) надо использовать также *словарные признаки*,
- Аналогичным образом решаются и другие задачи, например:
  — обнаружение аномалий в полётах самолётов,
  — диагностика болезней по электрокардиограмме

Внимание, конкурс!

Дано:
обучающая выборка — тексты Декларации на 6 языках,
тестовая выборка — 1560 фрагментов фраз на тех же языках,
всё в двух кодировках: UTF-8 и ASCII (без акцентов)
http://www.MachineLearning.ru/wiki?title=User:Vokov

Найти:
язык каждой тестовой фразы

Критерий:
число ошибок классификации

Решение прислать мне: voron@forecsys.ru в текстовом файле,
с тем же порядком строк, что в файле тестовой выборки,
в каждой строке — номер языка тестовой фразы.

Крайний срок — вечер 21 августа.

Как выглядят тестовые данные (UTF8)

de l'ocre rouge et un colorant noir
of gamelan onto the western staff
manera horizontal melodía y vertical armonía
otras que no lo son
equipara infancia con el arte prehistórico
slegs in die natuur gevind kan word
a squiggle on a horizontal staff
en dos ejes uno horizontal
dialektik und rhetorik das quadrivium umfasste geometrie
não era tida como arte
et de trouver du nouveau
sugerido la supresión total de los impuestos
van rykdom of mag

... всего 1560 фрагментов на 6 языках

Подсказки

- использование букв с акцентами (диакритическими знаками) улучшает качество, но придётся разобраться с кодировкой UTF-8 или применить конвертер, чтобы построить объединённый алфавит
- чтобы использовать метрический классификатор, не обязательно вычислять признаки — можно сравнивать сами тексты
- функцию близости двух текстов естественного языка можно определить как длину максимальной общей подпоследовательности
- или как суммарную частоту N-грамм одного текста во втором тексте
- униграммы, биграммы, ... N-граммы можно использовать совместно
- если вес признака близок к нулю, то лучше этот признак вообще не использовать — когда таких признаков становится много, все вместе они привносят шум и способны испортить модель
- веса признаков для взвешенной евклидовой метрики можно определить так же, как для линейного классификатора
- смелее придумывайте собственные эвристики!

Воронцов Константин Вячеславович
voron@forecsys.ru

www.MachineLearning.ru • Участник:Vokov

Если что-то было не понятно,
не стесняйтесь подходить и спрашивать :)