

# *Основные понятия Text Mining*

Курс «Интеллектуальные информационные системы»

Кафедра управления и информатики НИУ «МЭИ»

Осень 2017 г.

# Подходы Text Mining



## Лингвистический анализ

Системы ЛА обычно состоят из модели предметной области, содержащей основные тематические термины и их взаимосвязи, а также специализированной базы данных (БД) грамматических конструкций и семантических правил, свойственных конкретному языку – онтологий и тезаурусов. При этом модель предметной области обычно используется для проведения морфологического анализа, а специализированная БД – для синтаксического и семантического анализа

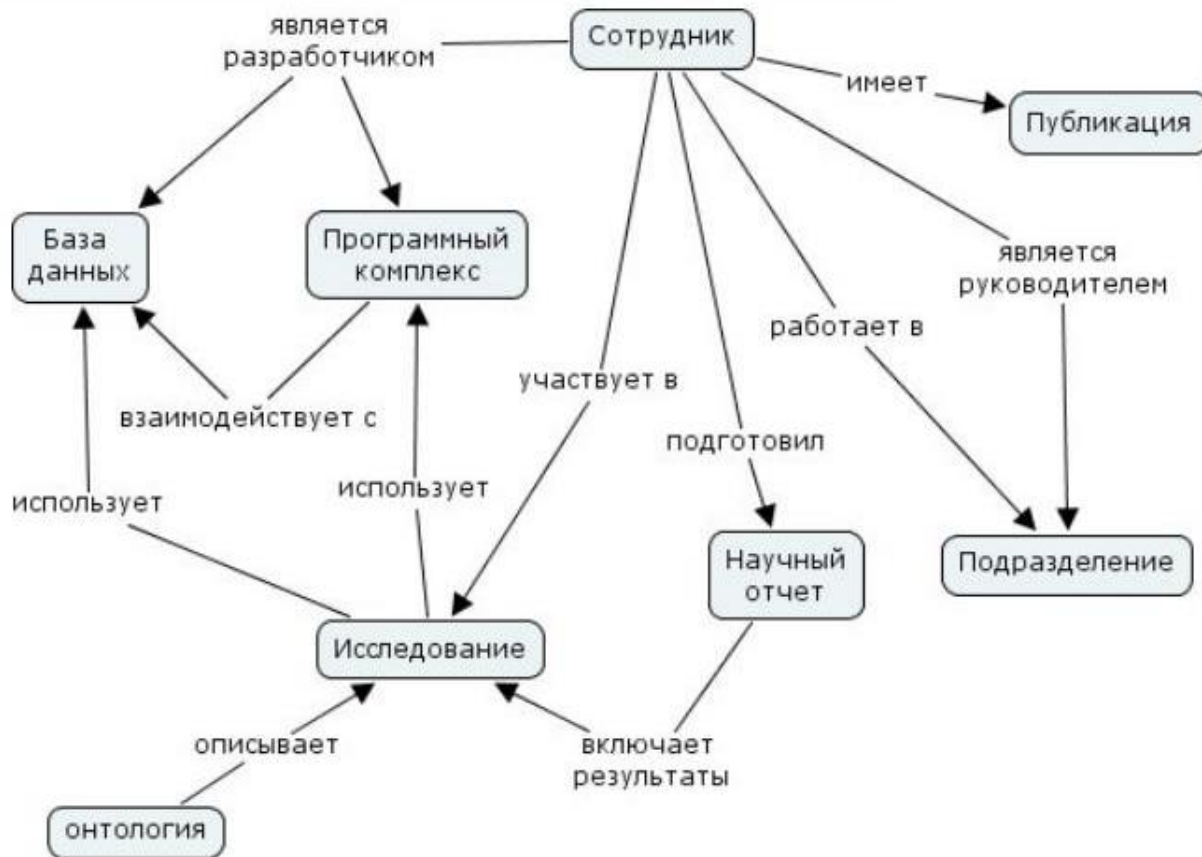


## Статистический анализ

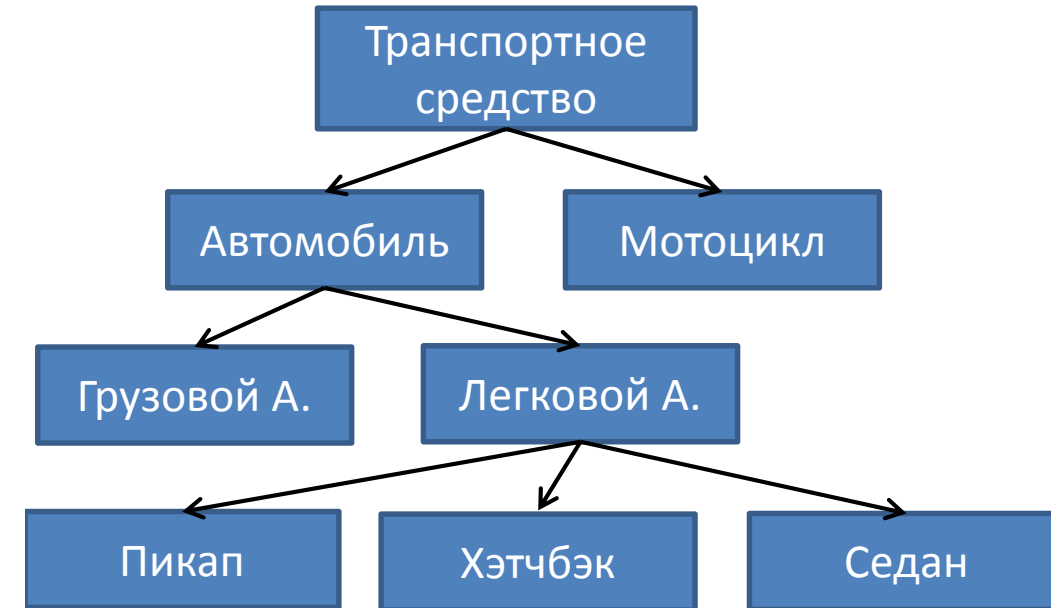
Текст – набор ключевых слов. Вес слов зависит от различных факторов, в частности – от частоты встречаемости термина в документе. Предполагается, что появление одних и тех же терминов в различных документах говорит об их подобии

# Онтологии и тезаурусы

Пример онтологии



Пример тезауруса



# Что такое текст?

- Текст – конечное множество слов (терминов), объединенных лексическими, грамматическими, смысловыми, частотными отношениями и образующих информативное сообщение.
- Главное в тексте – информация, новая для читателя, которая заключена в авторском изложении, и которую мы хотим извлечь.
- Чем больше информации извлечем – тем лучше.
- Не всегда большой текст = большому количеству информации

# Модели представления текстовых документов

- *Неструктурированная модель* – «мешок слов» (“bag of words”) – каждый термин рассматривается в качестве независимой случайной величины. Не учитываются возможные связи с другими словами в тексте.
- *Частично структурированная модель*
  - учет дополнительной информации о положении слова в тексте (заголовок, ключевые слова, первый абзац,...),
  - учет оформления слова (*курсив*, **полужирный**, подчеркивание,...),
  - выделение словосочетаний:  $w = \frac{w_{kj}}{w_k w_j}$
- *Полностью структурированная модель*
  - Использование информации из тезаурусов, онтологий, специальных словарей (WordNet)

# Как документ представляется в математическом виде?

Векторная модель:

Документ:

$$\vec{X}_j = \begin{bmatrix} x_j^{(1)} \\ \vdots \\ x_j^{(i)} \\ \vdots \\ x_j^{(M)} \end{bmatrix}$$

Матрица «Документ-термин»:

$$X = \begin{pmatrix} x_1^{(1)} & \dots & x_1^{(M)} \\ \vdots & \ddots & \vdots \\ x_N^{(1)} & \dots & x_N^{(M)} \end{pmatrix}$$

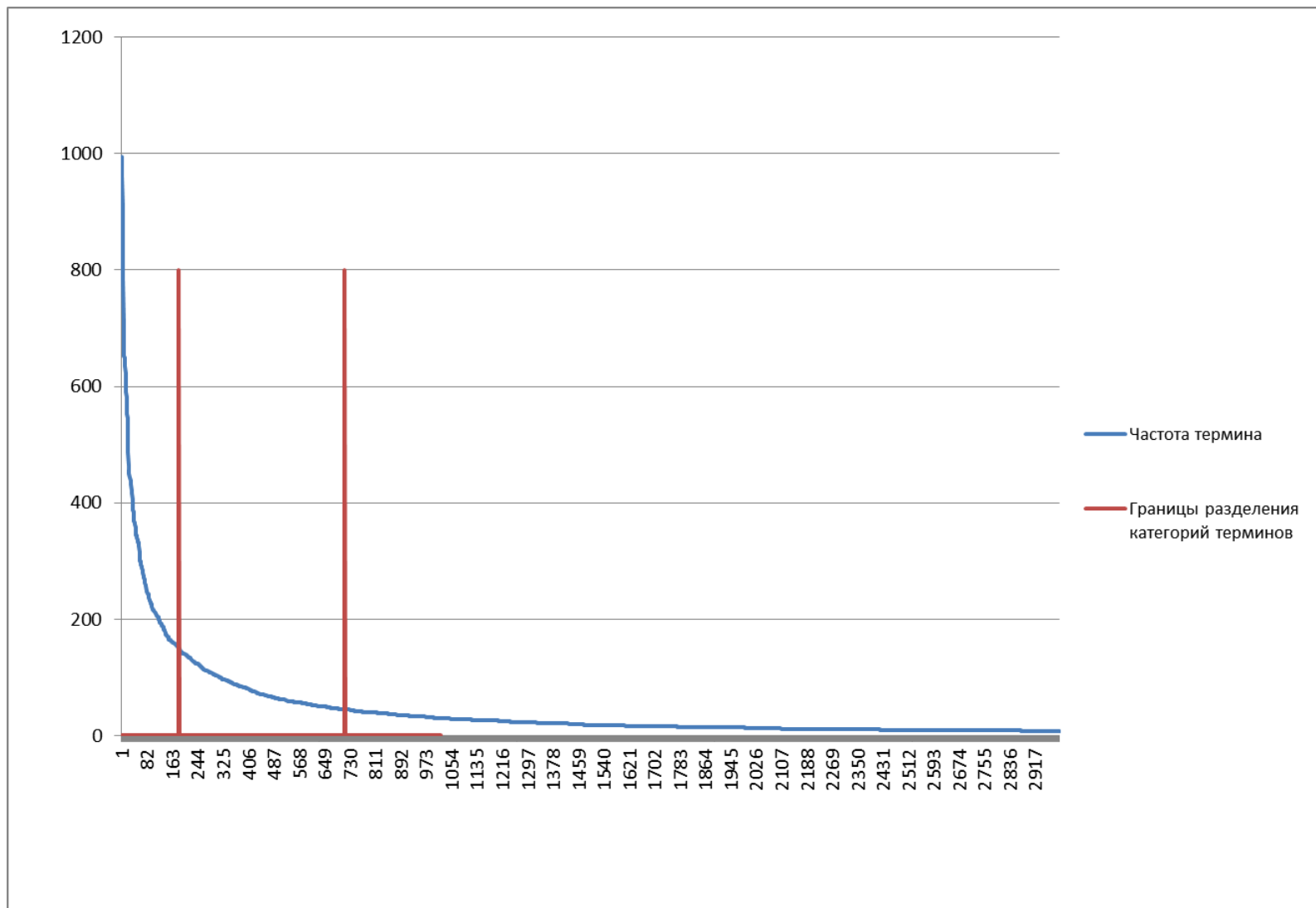
$x_j^{(i)}$  - Вес термина  $i$  в документе  $j$   
 $i = 1..M, j = 1..N$

Размерность матрицы крайне высокая,  $M \rightarrow 10^4 - 10^5$

\* Вместо термина (или слова) могут использоваться n-граммы - последовательность из  $n$  элементов:

Триграммы Hello world: Hel, ell, llo, lo ,o w, wo, orl, rld.

# Выявление информативных признаков



Закон Ципфа:  
 $w_n = w_1/n$

# Предварительная обработка документов

## Предварительная обработка данных



A first attempt to solve this problem is to exploit the information provided by external knowledge sources such as bilingual dictionaries, to collapse all the rows representing translation pairs in this setting. The similarity of this problem...

A
first
Attempt
to
solve
this
problem
is
to
exploit
the
information
provided
by
...

first
attempt
solve
problem
exploit
information
provided
external
knowledge
sources
bilingual
dictionaries
collapse
rows
...

first
attempt
solve
problem
exploit
information
provid
extern
knowledge
source
bilingual
dictionar
collapse
row
...

term	23
text	20
problem	19
word	11
information	10
translate	8
resource	6
source	3
Knowledge	3
dictionar	2
exploit	2
improve	1
collapse	1
row	1
...	

term	23
text	20
problem	19
word	11
information	10
translate	8
resource	6
source	3
knowledge	3
dictionar	2
exploit	2
improve	1
collapse	1
row	1
...	



# Результат проведения стемминга и отсеечения стоп-слов

## Семантическая интерпретация в системах компьютерного анализа текста

Описывается подход к построению семантического компонента в системах компьютерного анализа текста на естественном языке. Подход основан на применении специальных шаблонов к сети синтактико-семантических отношений между словами текста, которая строится синтаксическим анализатором. Шаблоны определяют способ интерпретации фрагментов сети в заданные фреймы с идентификацией участников ситуаций и их ролей.

**Ключевые слова:** компьютерный анализ текста, семантическая интерпретация, семантическая сеть, синтаксический анализ, фреймы.

## Semantic Interpretation in Computer Text Analysis Systems

The article describes an approach to semantic component building in computer text analysis systems for a natural language text. The approach is based on applying special patterns to a net of syntactic and semantic relations between words in a text, which is formed by a syntactic parser. The patterns define the way to interpret parts of the net according to given frames with identification of participants of the situation and their roles.

**Keywords:** text mining, semantic interpretation, semantic network, syntactic parser, frames.

# Определение весов терминов

Название	Формула
Логическое взвешивание	$x_j^{(i)} = \begin{cases} 1, & f_{ij} > 0 \\ 0, & f_{ij} = 0 \end{cases}$
Взвешивание частотой слова ( <i>term frequencies, tf</i> )	$x_j^{(i)} = f_{ij}$
<i>tf-idf</i> - взвешивание ( <i>term frequencies – inverse document frequencies</i> )	$x_j^{(i)} = f_{ij} \log\left(\frac{N}{N_i}\right)$
<i>tfc</i> - взвешивание	$x_j^{(i)} = \frac{f_{ij} \log\left(\frac{N}{N_i}\right)}{\sqrt{\sum_{i=1}^M \left[ f_{ij} \log\left(\frac{N}{N_i}\right) \right]^2}}$

# Определение весов терминов (2)

Название	Формула
<i>lfc</i> – взвешивание. Данный подход заключается в использовании логарифма частоты слова вместо $f_{ij}$ . Это позволяет сократить характерный для большинства текстовых документов существенный разброс в частотах различных терминов	$x_j^{(i)} = \frac{\log(f_{ij} + 1) \log\left(\frac{N}{N_i}\right)}{\sqrt{\sum_{i=1}^M \left[ \log(f_{ij} + 1) \log\left(\frac{N}{N_i}\right) \right]^2}}$
<i>atc</i> – взвешивание. При таком взвешивании веса будут изменяться от 0,5 до 1, что в ряде случаев приводит к улучшению качества классификации, позволяя учесть значимые термины, имеющие редкую встречаемость в конкретной выборке	$x_j^{(i)} = \frac{(0,5 + 0,5 \frac{f_{ij}}{\max f}) \log\left(\frac{N}{N_i}\right)}{\sqrt{\sum_{k=1}^{Mk} \left[ (0,5 + 0,5 \frac{f_{ij}}{\max f}) \log\left(\frac{N}{N_k}\right) \right]^2}}$

Кроме взвешивания применяются и другие методы выявления информативных терминов:

- Факторный и компонентный анализ (переход к новой системе признаков)
- Статистический подход (Хи-квадрат критерий)
- Теоретико-информационный подход

## *Факторный анализ (ФА) и Метод Главных Компонент (МГК)*

- ФА: различные признаки являются одним и тем же явлением., следовательно можно создать новые переменные – «факторы», позволяющие «вскрыть» логическую структуру выборки.
- МГК: переход к новым переменным, которые являются линейной комбинацией исходных.

Проведение снижения размерности с помощью ФА и МГК особенно эффективно для отображения объектов в трехмерное пространство и на плоскость.

## Статистический подход выявления информативных признаков

Q	$Q_1$	$Q_2$	...	$Q_K$	$\sum_{k=1}^K n_{ik} = n_{i*}$
X					
$x^{(1)}$	$n_{11}$	$n_{12}$	...	$n_{1K}$	$n_{1*}$
$x^{(2)}$	$n_{21}$	$n_{22}$	...	$n_{2K}$	$n_{2*}$
...	...	...	...	...	...
$x^{(M)}$	$n_{M1}$	$n_{M2}$	...	$n_{MK}$	$n_{M*}$
$\sum_{i=1}^M n_{ik} = n_{*k}$	$n_{*1}$	$n_{*2}$	...	$n_{*K}$	$n_{**} = N$

$n_{ik}$  — клеточная частота — число объектов в выборке, обладающих данным сочетанием переменных  $\{x^{(i)}, Q_k\}$

Проверяется гипотеза  $H_0: n_{ik} - \hat{n}_{ik} = 0$

$$\hat{n}_{ik} = P(x^{(i)}, Q_k) = P(x^{(i)})P(Q_k) = N \frac{n_{i*}}{N} \cdot \frac{n_{*k}}{N} = \frac{n_{i*}n_{*k}}{N}$$

$$\chi^2 = \sum_{i=1}^M \sum_{k=1}^K \frac{(n_{ik} - \hat{n}_{ik})^2}{\hat{n}_{ik}}$$

Гипотеза о независимости отвергается с уровнем значимости  $\alpha$ , если рассчитанная величина  $\chi^2$  превышает критическое значение  $\chi_{\alpha, S}^2$

где  $S=(M-1)(K-1)$

## Частный случай Хи-квадрат критерия

$X \backslash Q_k$	Принадлежность классу $Q_k$	Непринадлежность классу $Q_k$	$\Sigma$
Наличие признака $x^{(i)}$	$A$	$B$	$A+B$
Отсутствие признака $x^{(i)}$	$C$	$D$	$C+D$
$\Sigma$	$A+C$	$B+D$	$N$

$$\chi^2(x^{(i)}, Q_k) = N \frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1*}n_{2*}n_{*1}n_{*2}} = N \frac{(AD - CB)^2}{(A+B)(C+D)(A+C)(B+D)}$$

$$\chi_{\text{средний}}^2(x^{(i)}) = \sum_{k=1}^K P(Q_k) \chi^2(x^{(i)}, Q_k)$$

$$\chi_{\text{max1}}^2(x^{(i)}) = \max_{k=1}^K P(Q_k) \chi^2(x^{(i)}, Q_k)$$

$$\chi_{\text{max2}}^2(x^{(i)}) = \max_{k=1}^K \chi^2(x^{(i)}, Q_k)$$

$$P(Q_k) = \frac{N_k}{N}$$

Недостатки  $\chi^2$  - критерия:

- Вычислительная сложность
- Невысокая точность для редких терминов

## Критерий взаимной информации (Mutual information)

Взаимная информация, как среднее количество информации, содержащееся в  $X$  относительно  $Q$ :

$I(X, Q) = H(X) - H(X | Q)$  , где  $H(X), H(X | Q)$  – соответственно энтропия и условная энтропия.

$$I(X, Q) = -\sum_{i=1}^M P(x^{(i)}) \log P(x^{(i)}) + \sum_{i=1}^M \sum_{k=1}^K P(x^{(i)}, Q_k) \log \frac{P(x^{(i)}, Q_k)}{P(Q_k)}$$

$$P(x_i) = P(Q_1)P(x^{(i)} | Q_1) + P(Q_2)P(x^{(i)} | Q_2) + \dots + P(Q_K)P(x^{(i)} | Q_K) = P(x^{(i)}) = \sum_{k=1}^K P(x^{(i)}, Q_k)$$

## Критерий взаимной информации (Mutual information) (2)

$$I(X, Q) = -\sum_{i=1}^M P(x^{(i)}) \log P(x^{(i)}) + \sum_{i=1}^M \sum_{k=1}^K P(x^{(i)}, Q_k) \log \frac{P(x^{(i)}, Q_k)}{P(Q_k)} =$$

$$= \sum_{i=1}^M \sum_{k=1}^K P(x^{(i)}, Q_k) \log \frac{P(x^{(i)}, Q_k)}{P(Q_k) P(x^{(i)})}.$$

$$P(x^{(i)}) = \frac{A + B}{N}$$

$$P(Q_k) = \frac{A + C}{N}$$

$$P(x^{(i)}, Q_k) = \frac{A}{N}$$

$$MI(x^{(i)}, Q_k) = \log_2 \frac{AN}{(A + B)(A + C)}$$



$$I_{\text{сред}}(X, Q) = \sum_{k=1}^K P(Q_k) MI(x^{(i)}, Q_k)$$

$$I_{\text{max}}(X, Q) = \max_{k=1}^K \{MI(x^{(i)}, Q_k)\}$$

Данный критерий, в отличие от Хи-квадрат, большие веса дает редким признакам



# Меры близости и расстояния

Евклидово расстояние

$$d(\vec{X}_j, \vec{X}_l) = \sqrt{\sum_{i=1}^M (x_j^{(i)} - x_l^{(i)})^2}$$

Расстояние городских кварталов

$$d(\vec{X}_j, \vec{X}_l) = \sum_{i=1}^M |x_j^{(i)} - x_l^{(i)}|$$

Косинусоидальная мера близости.

Показывает косинус угла между векторами.

Стремится к 1, когда документы похожи между собой

$$d(\vec{X}_j, \vec{X}_l) = \cos(\vec{X}_j, \vec{X}_l) = \frac{\sum_{i=1}^M x_j^{(i)} x_l^{(i)}}{\sqrt{\sum_{i=1}^M (x_j^{(i)})^2 \sum_{i=1}^M (x_l^{(i)})^2}}$$

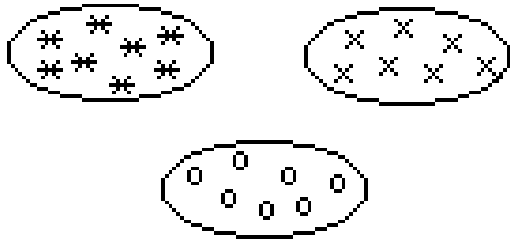
## *Формирование обучающих и экзаменационных выборок*

- Эффективность методов Machine Learning сильно зависит от того, как были сформированы обучающие выборки.
- Выборки должны быть:
  - Независимо извлеченными из генеральной совокупности
  - Представительными (репрезентативными)
  - Содержать минимум нетипичных объектов
- Не так важно, как выглядит генеральная совокупность во всем пространстве признаков. Гораздо важнее, как она выглядит в районе границы между двумя классами

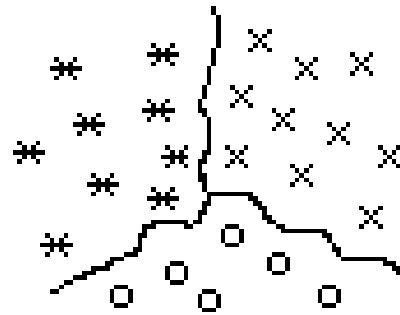
Неидеальность разметки документов – разные эксперты могут отнести документ к разным классам. Как поступать?

# Как оценить выборку?

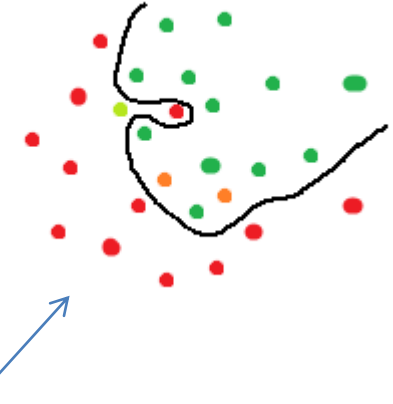
Ядерная (центроидная) модель



Модель рассеяния



Модель засорения



Переобученность (Overtraining)

Средняя сумма внутриклассовой дисперсии:

$$Q_1 = \frac{1}{N_k} \sum_{j=1}^{N_k} d^2(\vec{X}_j, \vec{X}_k) \quad \text{или} \quad Q_1^* = \frac{1}{M} \sum_{k=1}^M \frac{1}{N_k} \sum_{j=1}^{N_k} d^2(\vec{X}_j, \vec{X}_k),$$

Средняя сумма квадратов внутриклассовых попарных расстояний

$$Q_2 = \frac{1}{N_k} \sum_{l=1}^{N_k} \sum_{j=1, j \neq l}^{N_k} d^2(\vec{X}_l, \vec{X}_j) \quad \text{или} \quad Q_2^* = \frac{1}{M} \sum_{k=1}^M \frac{1}{N_k} \sum_{l=1}^{N_k} \sum_{j=1, j \neq l}^{N_k} d^2(\vec{X}_l, \vec{X}_j)$$

## Как оценить выборку? (2)

Средняя сумма квадратов  
межклассовых попарных  
расстояний

$$Q_3 = \frac{1}{N_k N_s} \sum_{l=1}^{N_k} \sum_{j=1, j \neq l}^{N_s} d^2(\vec{X}_l, \vec{X}_j) \quad \text{или} \quad Q_3 = \frac{1}{M} \sum_{k=1}^M \sum_{s=1, s \neq k}^M \frac{1}{N_k N_s} \sum_{l=1}^{N_k} \sum_{j=1, j \neq l}^{N_s} d^2(\vec{X}_l, \vec{X}_j)$$

Обобщенный функционал

$$Q_4 = \frac{Q_3}{Q_2}$$

На основе такого анализа исследователь может: 1) объединить несколько близких небольших классов в один; 2) удалить “нехарактерные” шумовые элементы, расположенные вдалеке от центра классов (модель засорения); 3) заново сформировать выборку, увеличив (уменьшив) количество классов или количество элементов.

# Свойства сформированных выборок

- любая обучающая выборка конечного размера не является полной, т.е. не содержит необходимого количества элементов для проведения безошибочной классификации;
- элементы обучающей выборки обычно имеют произвольное распределение в пространстве признаков и, как следствие, получаемые решающие правила могут обладать неодинаковой достоверностью в различных областях изменения параметров;
- базы данных текстовых документов, из которых чаще всего составляются обучающие выборки, как правило, содержат шумовые (нерелевантные, не относящиеся к указанным классам) элементы, другую противоречивую или ошибочную информацию, которая так или иначе попадает в обучающую выборку.

# Оценка точность классификации в задачах Text Mining

Часть размеченных документов оставляют для обучения, часть – для оценки точности метода. Обычно используют следующие методы оценки:

- *Оценка точности по экзаменационным выборкам.  $N_{обуч} > N_{экзамен}$*
- *Оценка точности с помощью скользящего контроля (или «метод складного ножа», «Jackknife») – для небольших выборок*
- *Оценка точности с помощью  $v$ -кратной перекрестной проверки ( $v$ -fold cross validation)*

# Оценка точность классификации в задачах Text Mining (2)

Ошибка классификации – несовпадение метки, назначенной классификатором с меткой, назначенной экспертом (учителем).

Точность (правильность, аккуратность)

$$\text{Accuracy} = \frac{P}{N}$$

P- количество документов, по которым классификатор принял правильное решение

$$\text{Точность } Precision = \frac{TP}{TP+FP}$$

$$\text{Полнота Recall} = \frac{TP}{TP+FN}$$

$$F\text{—measure} = \frac{2(Precision*Recall)}{Precision+Recall}$$

	Оценка эксперта	
Оценка системы	Положительная	Отрицательная
Положительная	TP	FP
Отрицательная	FN	TN