Классификация текстов

Квантитативный анализ текста

Кирилл Александрович Маслинский

21.03.2022 / 05

НИУ ВШЭ Санкт-Петербург

Задача классификации текстов

Задача классификации текстов

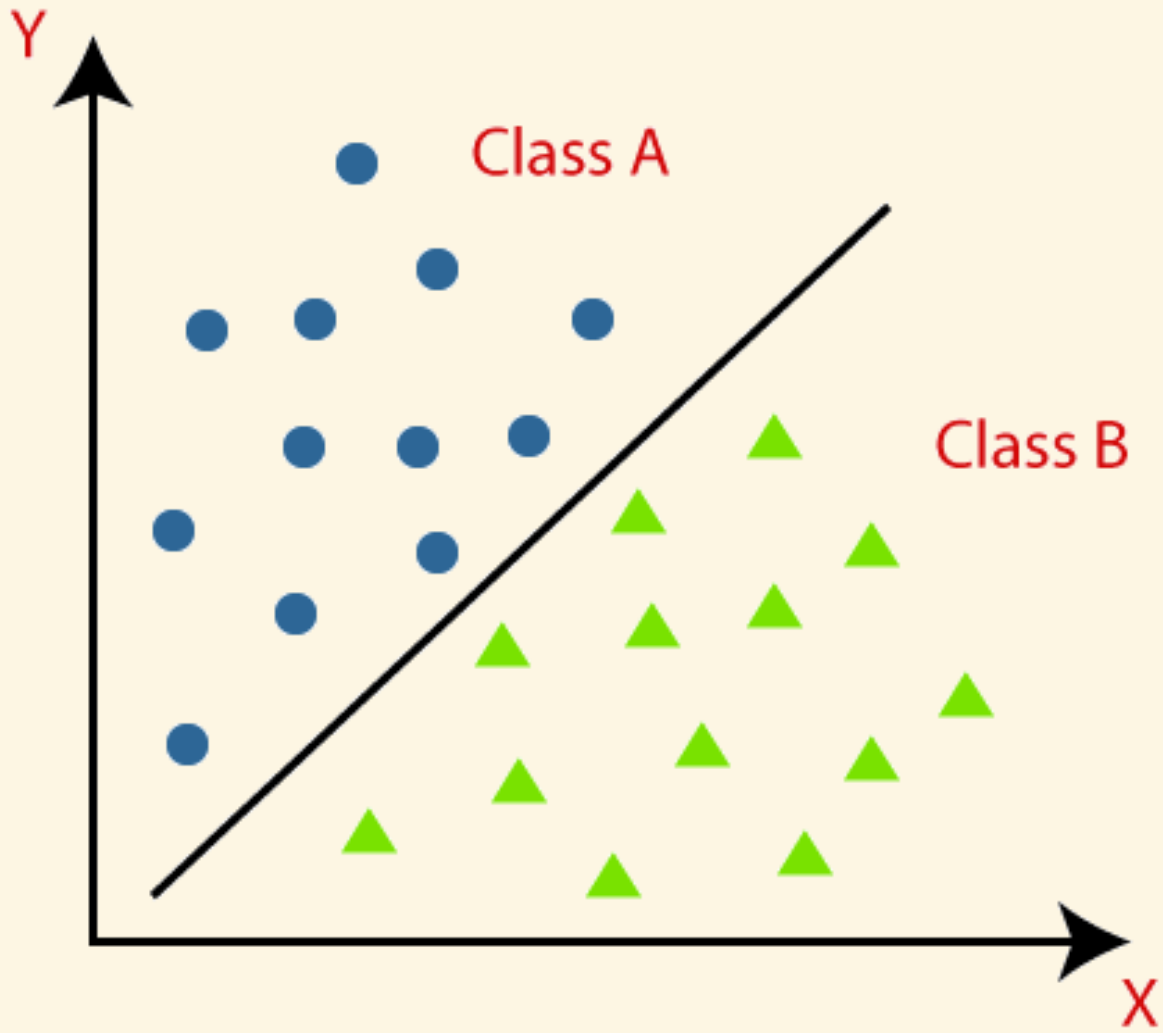
Задача:

- Заранее известен список классов
- Необходимо автоматически отнести каждый документ к одному из классов

Векторное представление документов

Docs	Terms вовочк	дет	класс	урок	учител	учи
1	0.5000000	0.0000000	0	0.2500000	0	0.
2	0.6666667	0.0000000	0	0.0000000	0	0.
3	0.2000000	0.2000000	0	0.0000000	0	0.
4	0.5000000	0.1666667	0	0.1666667	0	0.
5	0.7500000	0.0000000	0	0.0000000	0	0.
6	0.5000000	0.0000000	0	0.0000000	0	0.

Классификация векторов



Области применения классификации в NLP

- Документ целиком:
  - Определение языка текста
  - Определение тематики текста (из набора известных тем)
  - Sentiment classification (определение положительных/отрицательных отзывов)
  - Определение автора текста (из списка кандидатов)
- Отдельный токен (слово):
  - Разделение текста на предложения (классификация точек)
  - Определение части речи (part-of-speech tagging)
  - Снятие омонимии (выбор значения слова)
  - Извлечение именованных сущностей (Named entity recognition)
  - Извлечение отношений (Relations extraction)

Задача машинного обучения

Задача: научиться предсказывать трудно формализуемые, но важные для человека свойства объекта (текста).

target Определить набор интересующих нас меток

features Представить объект в виде набора свойств

model На основании статистики распределения свойств в текстах построить модель, предсказывающую метки новых объектов (которых модель еще не видела).

- PROFIT!

Задача машинного обучения

Задача: научиться предсказывать трудно формализуемые, но важные для человека свойства объекта (текста).

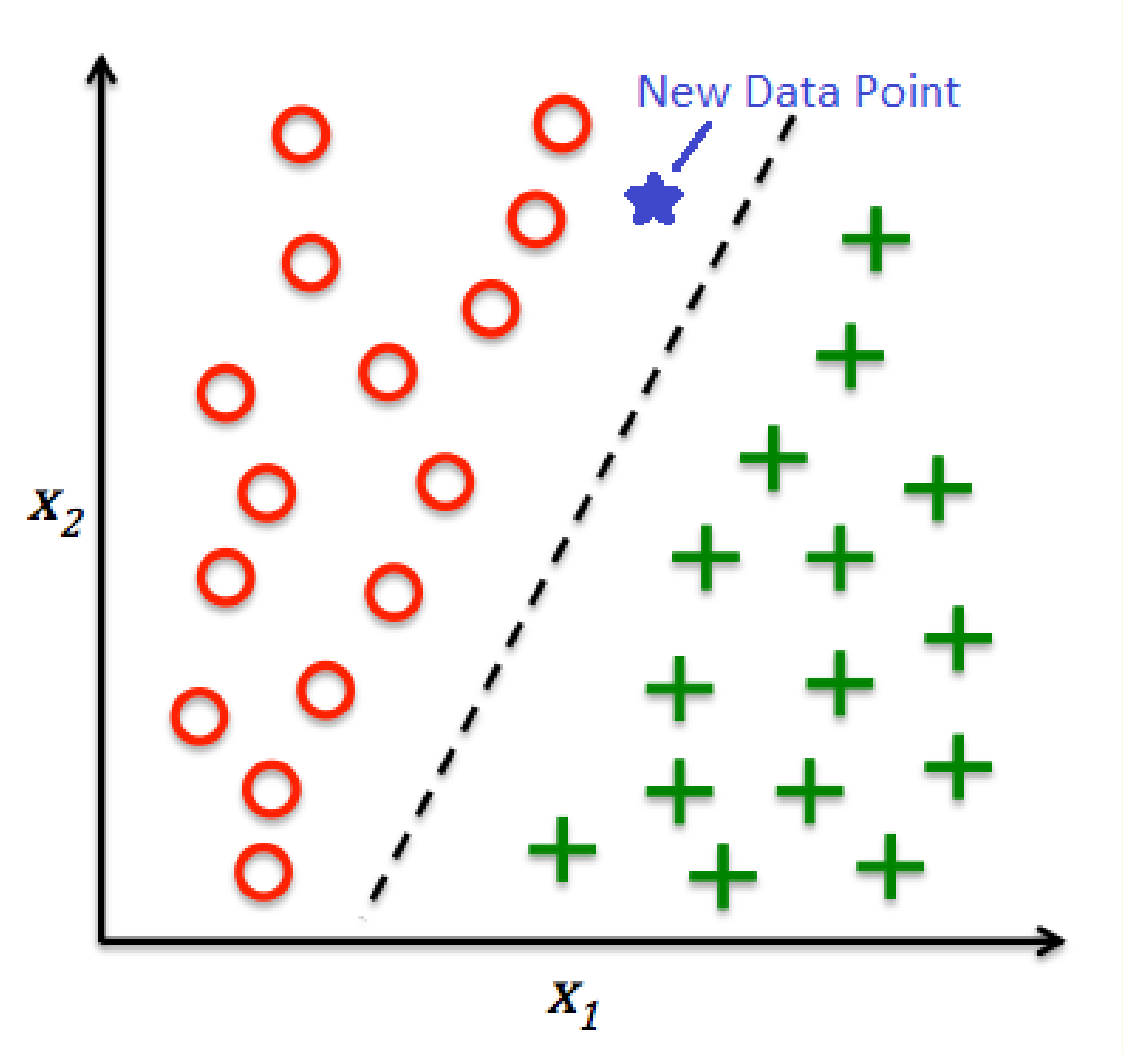
target Определить набор интересующих нас меток

features Представить объект в виде набора свойств

model На основании статистики распределения свойств в текстах построить модель, предсказывающую метки новых объектов (которых модель еще не видела).

- PROFIT!

Классификация новых объектов



Разновидности машинного обучения

supervised Обучение с учителем.
Модель учится предсказывать, опираясь на образцы меток, поставленных человеком.

unsupervised Обучение без учителя.
Модель учится предсказывать на основании общих предположений о распределении свойств в текстах, без подготовленных человеком размеченных образцов.

semi-supervised Обучение с использованием внешних знаний.
Модель опирается на небольшое количество размеченных человеком образцов и активное использование знаний о предметной области:

Терминология

Обучающая выборка / training set Набор объектов с выставленными человеком «правильными» метками, на основании которых строится («обучается») модель.

Тестовая выборка / test set Набор объектов с выставленными человеком «правильными» метками, с помощью которых можно проверить, совпадают ли предложенные моделью метки с правильными.

Наивный Байес

Правило Байеса

$$P(A \text{ и } B) =$$

$$P(B)P(A|B) = P(A)P(B|A)$$

Правило Байеса

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

$$posterior = \frac{likelihood \cdot prior}{evidence} \tag{2}$$

- $P(\text{менеджер}|\text{застенчивый}) = ?$
- $P(\text{библиотекарь}|\text{застенчивый}) = ?$

- $P(\text{МЕНЕДЖЕР}) = \frac{10\text{млн}}{80\text{млн}} = 0.125$
- $P(\text{БИБЛИОТЕКАРЬ}) = \frac{0{,}5\text{млн}}{80\text{млн}} = 0.00625$

Задача классификации образов

Задача:

- Заранее известен список классов
- Необходимо автоматически отнести каждый объект к одному из классов
- Каждый объект представлен в виде набора признаков (features)

Правило Байеса применительно к классификации

$$P(\text{label}|\text{features}) = \frac{P(\text{features}|\text{label})P(\text{label})}{P(\text{features})} \tag{3}$$

Naive Bayes classifier

- Задача: Имея набор свойств (E) выбрать наиболее вероятную гипотезу (класс, H). Знаменатель P(E) — константа и не влияет на результат классификации.

$$P(\text{label}|\text{features}) \propto P(\text{features}|\text{label})P(\text{label})$$

Maximum a posteriory estimation

Упрощающее предположение

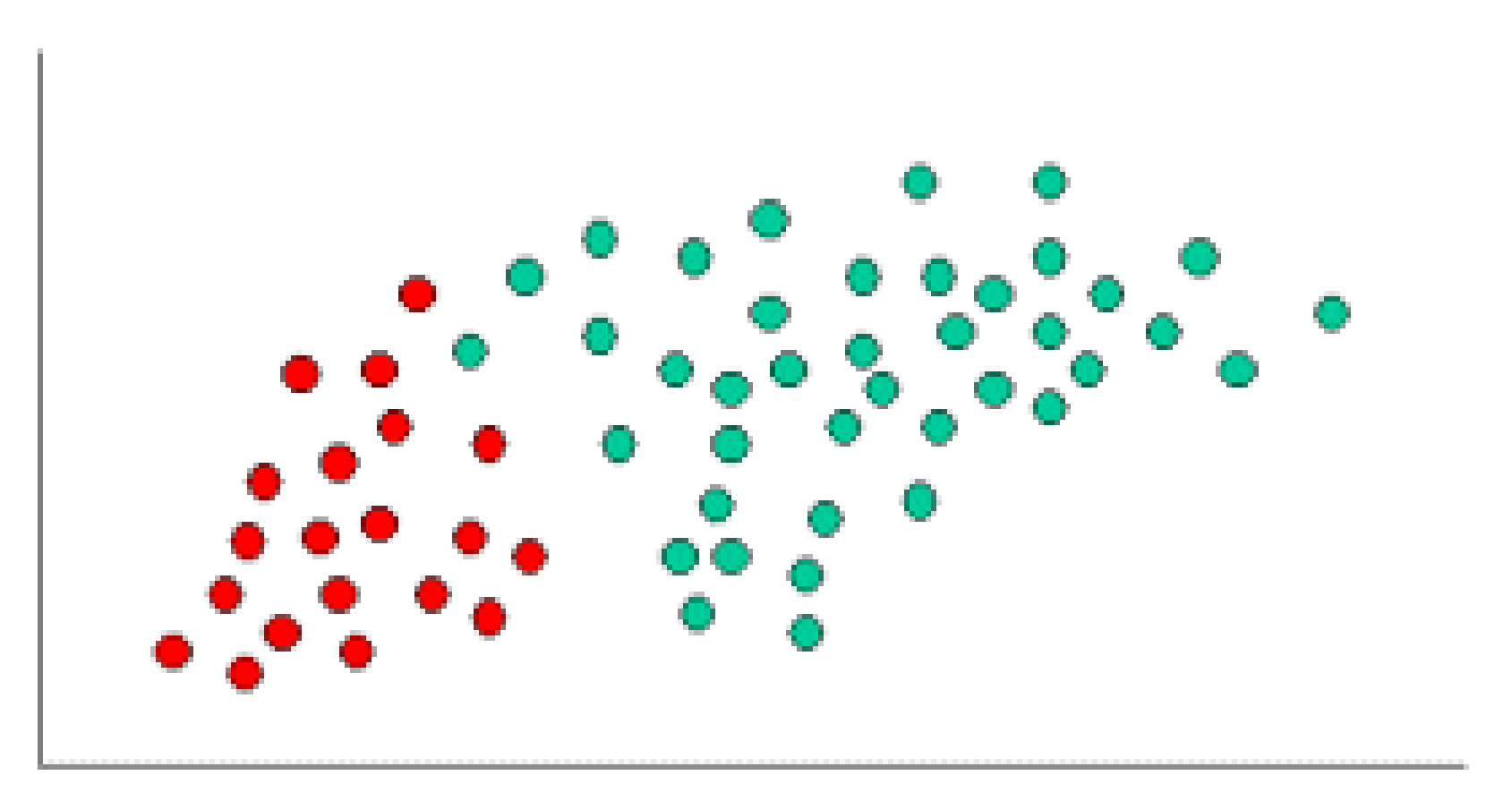
- Вычислить likelihood:

$$P(\text{features}|\text{label}) =$$

- Bayes assumption: Все свойства независимы друг от друга.

$$= \prod_{f \in \text{features}} P(f|\text{label})$$

Naive Bayes example

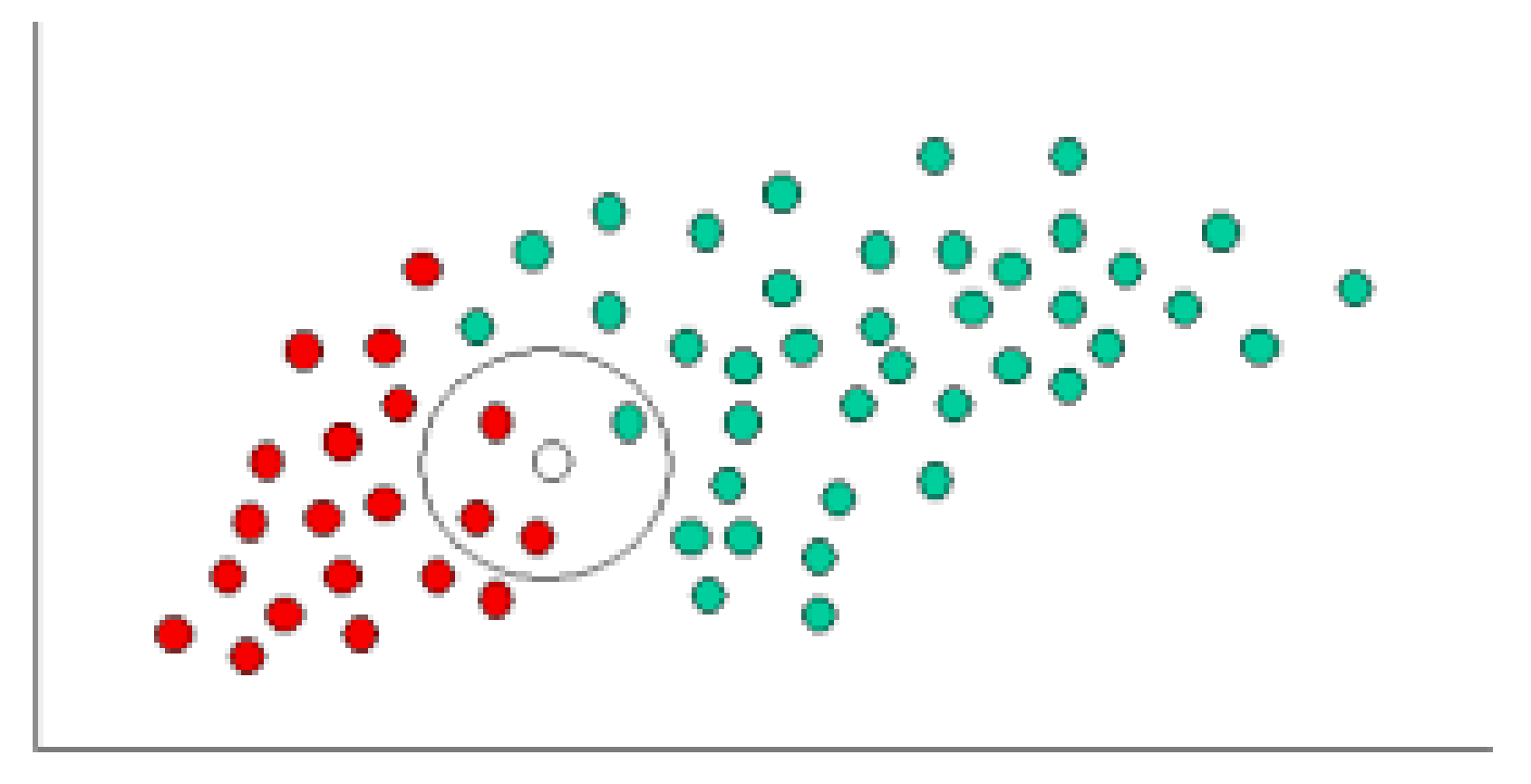


Naive Bayes example

$$\text{Prior}(\text{Green}) \propto \frac{f(\text{Green})}{\text{total}} = \frac{40}{60}$$

$$\text{Prior}(\text{Red}) \propto c\frac{f(\text{Red})}{\text{total}} = \frac{20}{60}$$

Naive Bayes example



Naive Bayes example

- $$\text{Likelihood}(X|\text{Green}) \propto \frac{f(\text{Green near X})}{f(\text{Green})} = \frac{1}{40}$$
- $$\text{Likelihood}(X|\text{Red}) \propto \frac{f(\text{Red near X})}{f(\text{Red})} = \frac{3}{20}$$

Naive Bayes example

$$\text{Posterior}(\text{Green}|\text{X}) \propto \text{Likelihood}(\text{X}|\text{Green}) \times \text{Prior}(\text{Green}) = \frac{4}{6} \times \frac{1}{40} = \frac{}{6}$$

$$\text{Posterior}(\text{Red}|\text{X}) \propto \text{Likelihood}(\text{X}|\text{Red}) \times \text{Prior}(\text{Red}) = \frac{2}{6} \times \frac{3}{20} = \frac{1}{20}$$

Эксперимент

Эксперимент на людях

Наивный Байес — генеративный классификатор:

- строит модель каждого класса
- определяет вероятность, что наблюдаемые данные сгенерированы по модели данного класса

Наивный Байес: преимущества и недостатки

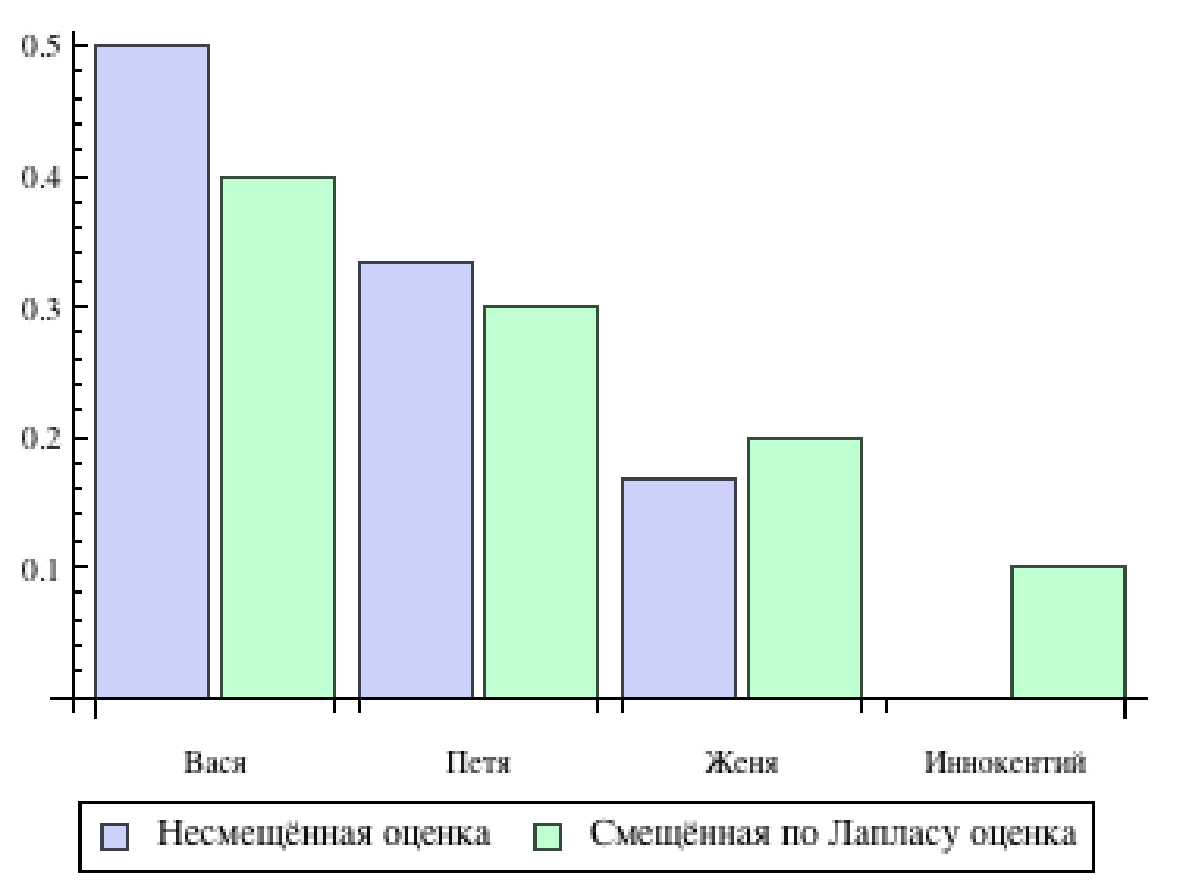
Преимущества:

- Годится для большого числа атрибутов, малой обучающей выборки
- На удивление хорошо работает в очень многих задачах
- Вычислительно эффективен (быстро учится и классифицирует)

Недостатки:

- Проблема нулевых значений (атрибут не встречается в обучающей выборке) — требует сглаживания

Аддитивное сглаживание



Проклятие размерности

Sparse data problem

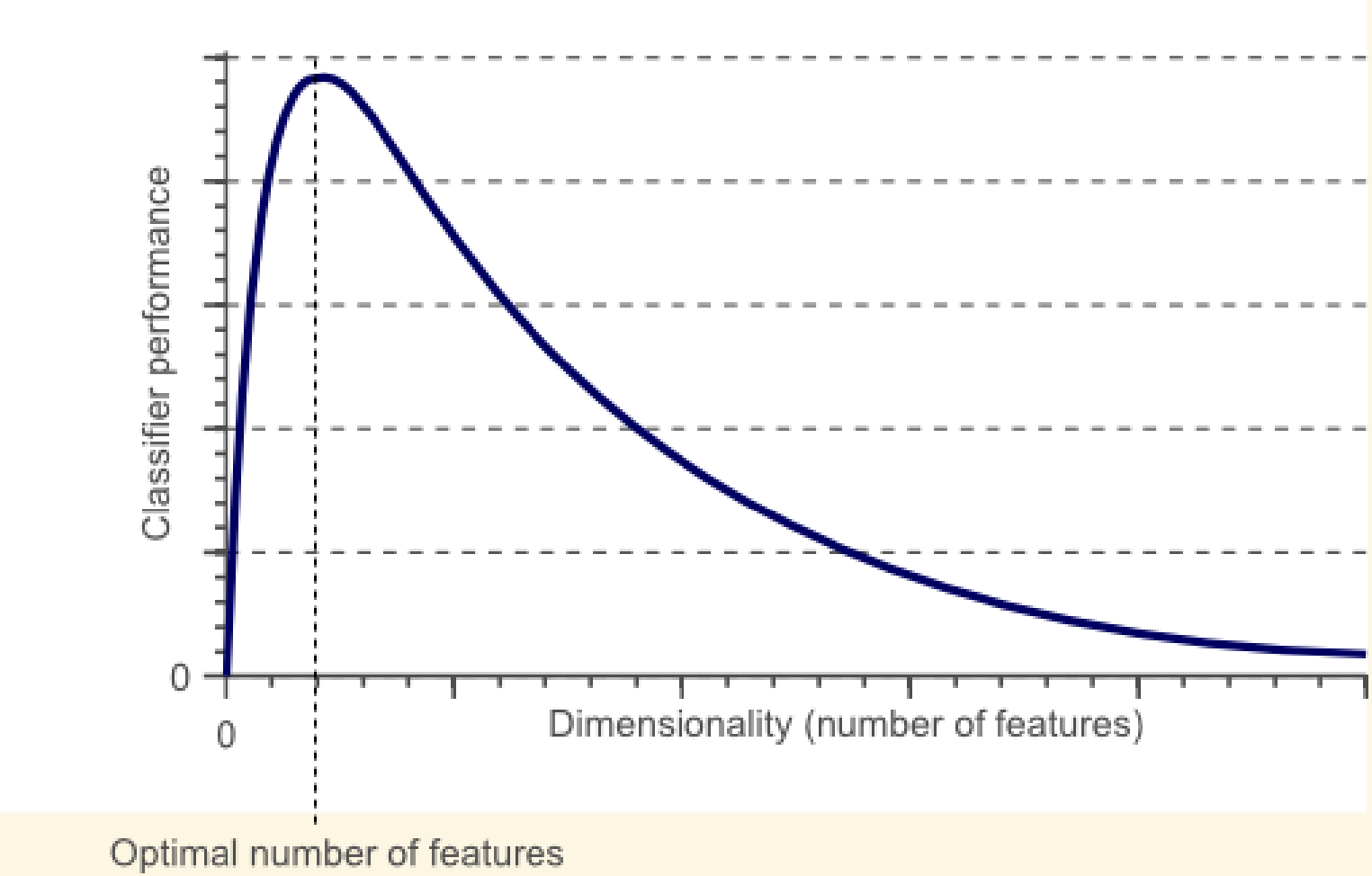
	Terms				
Docs	выгребать	выгребной	выгружать	выгрузка	выгрыза
1	0	0	0	0	
2	0	0	0	0	
3	0	0	0	0	
4	0	0	0	0	
5	0	0	0	0	

Проклятие размерности

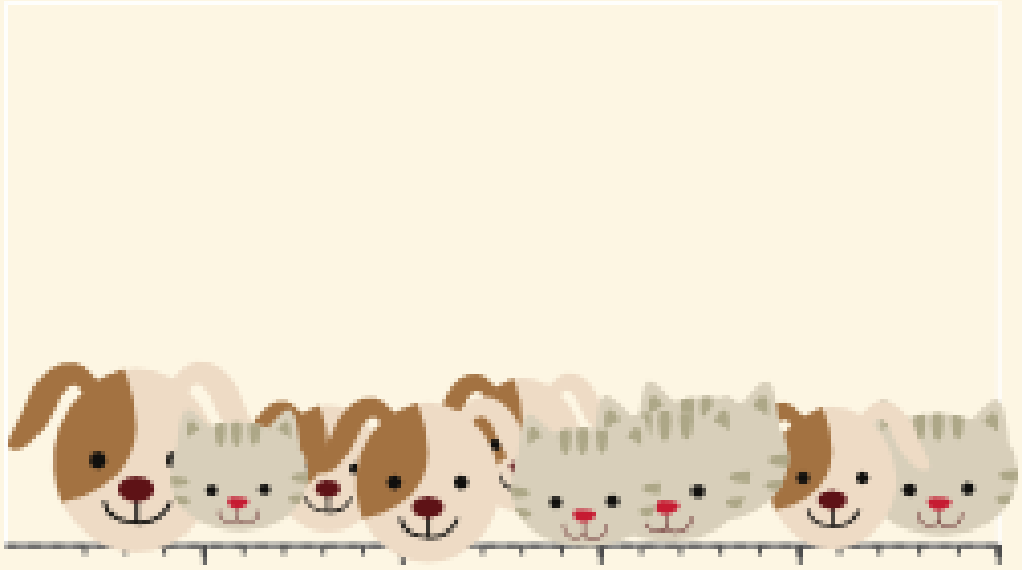
```
A document-term matrix (1530 documents, 13322 terms

Non-/sparse entries: 68859/20313801
Sparsity           : 100%
Maximal term length: 66
Weighting          : term frequency (tf)
```

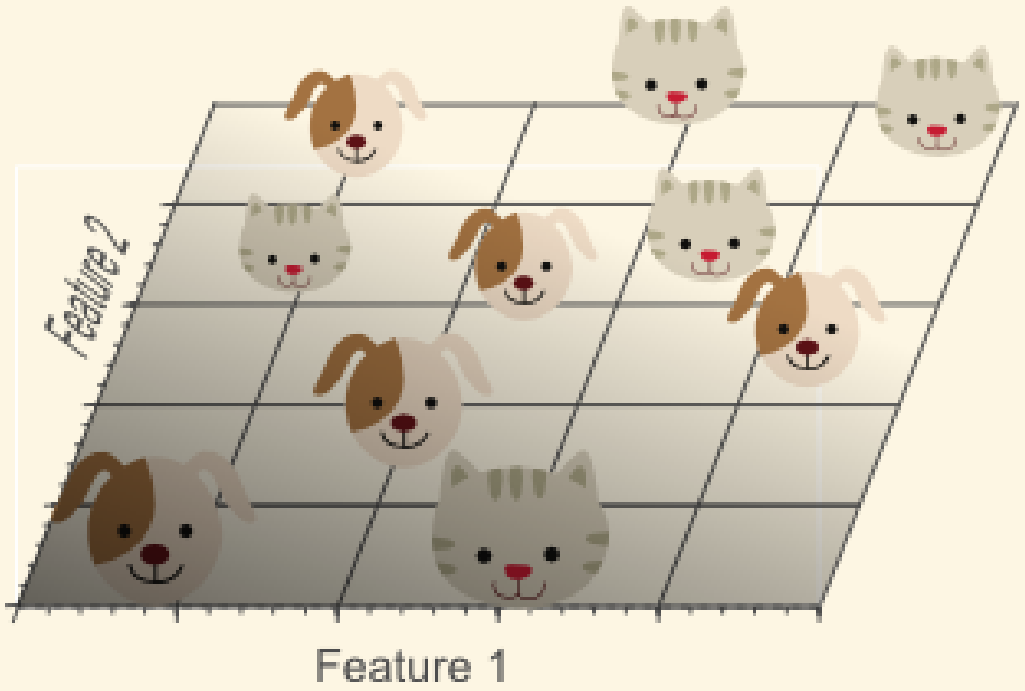
Hughes phenomenon



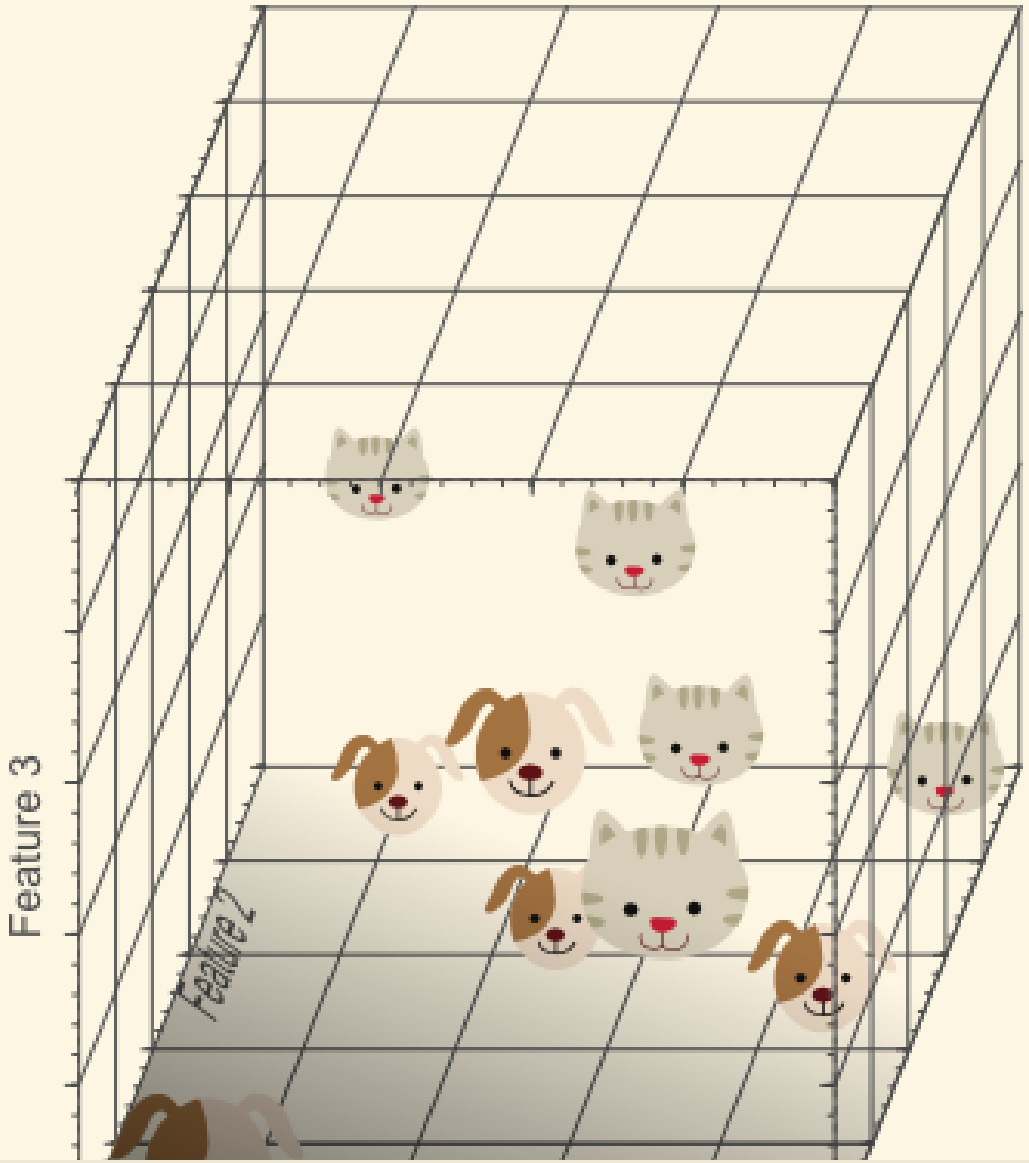
Проклятие размерности на кошках



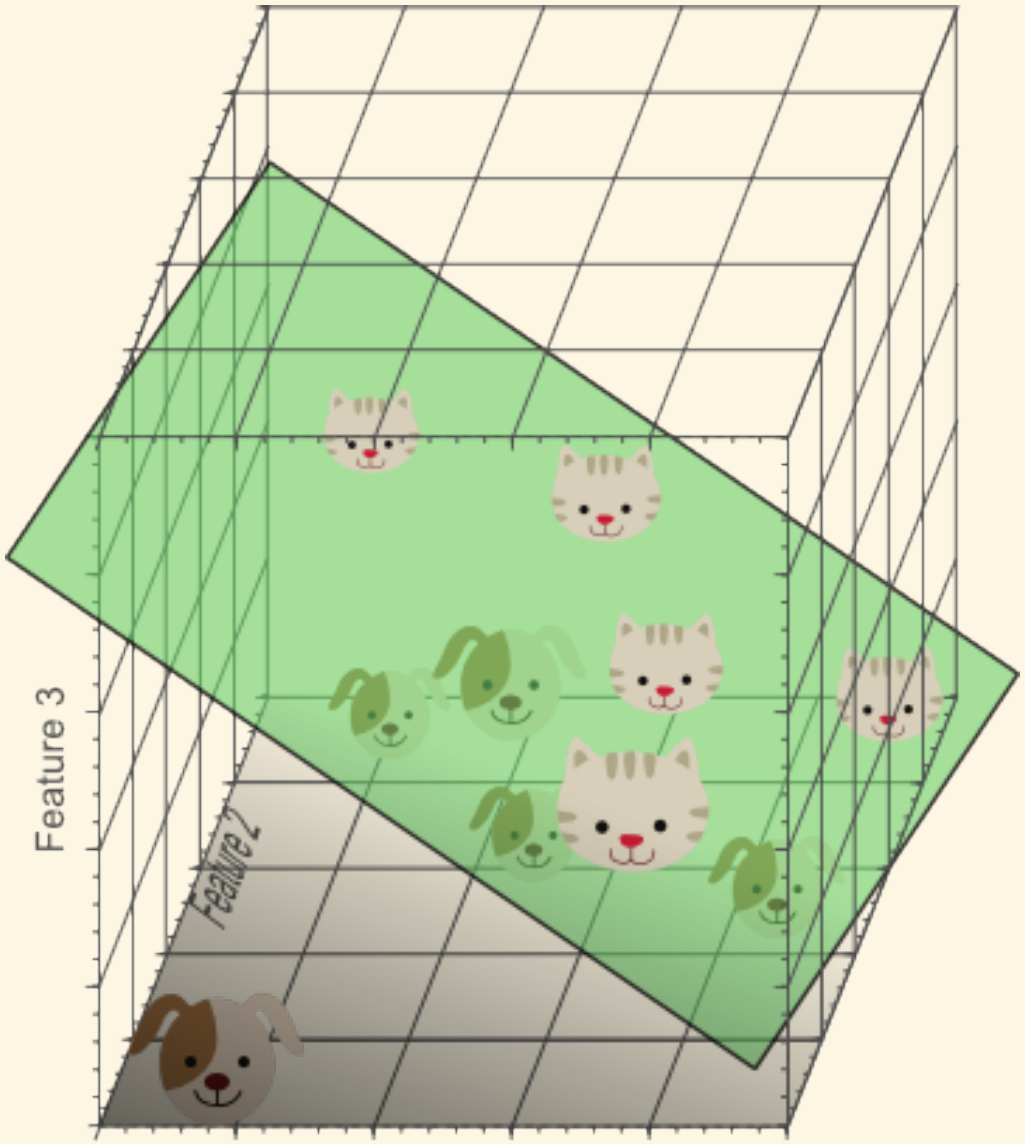
Проклятие размерности на кошках



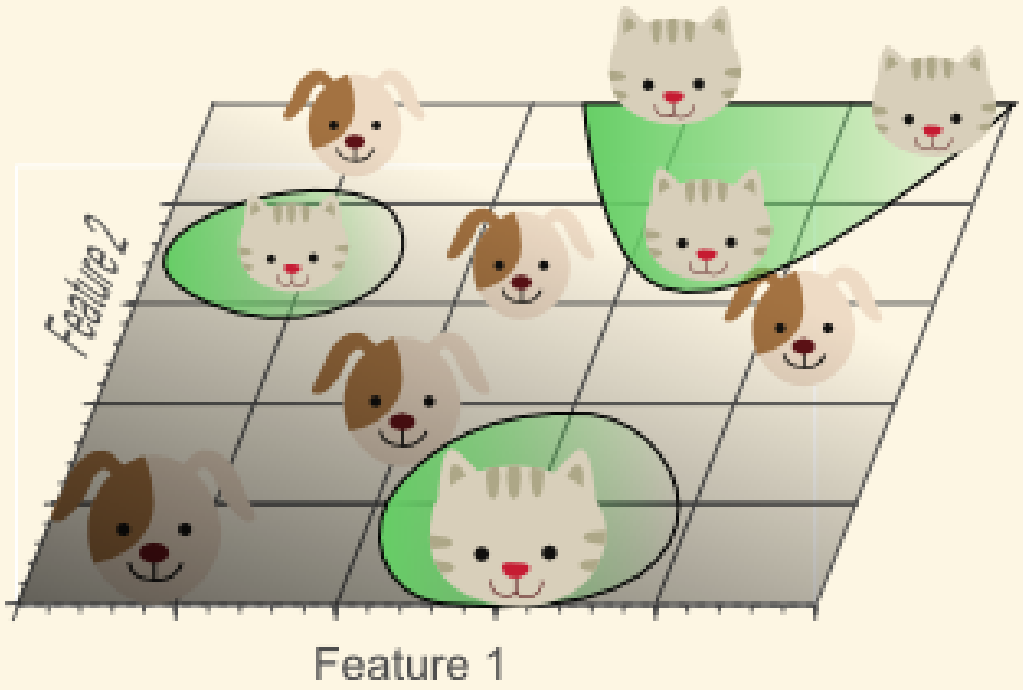
Проклятие размерности на кошках



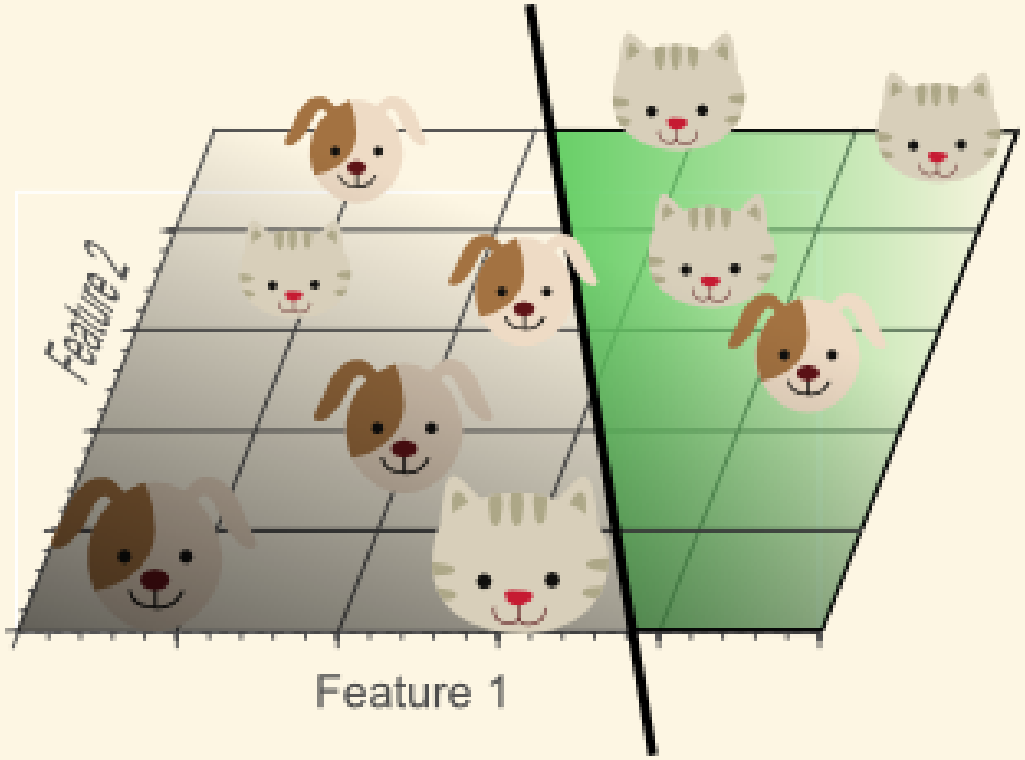
Проклятие размерности на кошках



Переобучение



Переобучение



Снижение размерности

Снижение размерности

- Матрица терминов-документов очень большая и редкая
- Близкие по смыслу слова не обязательно встречаются в одних и тех же документах:
  - синонимия
  - полисемия
  - шум
- Нужно сократить размерность матрицы (сделать меньше столбцов).

Снижение размерности

- Матрица терминов-документов очень большая и редкая
- Близкие по смыслу слова не обязательно встречаются в одних и тех же документах:
  - синонимия
  - полисемия
  - шум
- Нужно сократить размерность матрицы (сделать меньше столбцов).

Стоп-слова

Простейший способ уменьшить число столбцов — просто удалить лишние слова:

- Статический список:
  без более бы был была были было быть в вам вас весь во вот все всего всех вы где да даже для ...
- Динамический список:
  - Слишком частотные (N самых частотных)
  - Слишком редкие (порог: не менее чем в F документов)
  - Слишком короткие (меньше M букв)

Оценка качества классификации

Accuracy

$$Accuracy = \frac{P}{N} \tag{4}$$

P количество документов, где классификатор принял правильное решение

N размер обучающей выборки

Таблица сопряженности

Таблица верных и неверных решений по документам данного класса:

Категория i		Экспертная оценка	
		Положительная	Отрицательная
Оценка системы	Положительная	TP	FP
	Отрицательная	FN	TN

TP правильно отнесла к классу

TN правильно не включила в класс

FP ошибочно отнесла к классу

FN ошибочно не включила в класс

Точность и полнота

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

Матрица неточностей

	0.91	0.96	0.94	0.75	1.00	0.83	0.85	0.97	1.00	0.86	1.00	0.79	1.00	0.75	1.00	1.00	0.96	0.90	0.81	0.89	0.94	0.98	0.86	0.89	0.94	0.92	0.96
0.80		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
0.95	1	94	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0
1.00	2	0	32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.29	3	0	0	6	0	0	3	2	0	1	0	0	0	0	0	0	1	1	0	0	1	0	1	3	0	2	0
1.00	4	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.50	5	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2	0	1	1
0.92	6	1	0	0	0	0	152	0	0	1	0	0	0	0	0	0	0	1	4	2	3	0	0	0	0	2	0
0.97	7	1	0	1	0	0	0	256	0	0	0	0	0	0	0	0	0	0	0	1	2	0	0	0	0	2	0
0.33	8	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0
0.97	9	0	0	0	0	0	0	0	0	69	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
0.82	10	0	0	0	0	0	2	0	0	0	18	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
0.87	11	0	0	0	0	0	0	0	0	0	0	34	0	4	0	0	0	0	0	0	0	0	0	1	0	0	0
1.00	12	0	0	0	0	0	0	0	0	0	0	0	37	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.57	13	0	0	0	0	0	0	0	0	0	0	9	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0
0.63	14	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	3	0	0	0	0	0	0	0	0	0
0.50	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	1	1	0	0	0	0	0	0
0.77	16	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	47	0	1	3	4	0	0	2	0	1	0
0.87	17	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	69	1	2	5	0	0	0	0	0	0
0.97	18	0	0	0	0	1	4	0	0	1	0	0	0	0	0	0	0	0	197	1	0	0	0	0	0	0	0
0.78	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	35	183	13	0	0	2	0	1	0
0.97	20	0	0	0	0	0	10	3	0	1	0	0	0	0	0	0	0	0	0	4	702	0	0	0	0	6	0
0.93	21	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	56	0	2	0	0	0
0.29	22	0	0	1	0	0	2	0	0	6	0	0	0	0	0	0	0	0	1	1	1	0	6	2	0	1	0
0.91	23	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	3	6	0	0	115	0	0	0
1.00	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	0	0
0.93	25	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2	4	5	0	0	0	1	196	0
0.98	26	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	78

F-MEPA

$$F_\alpha = \frac{(1+\alpha)PR}{\alpha P + R} \tag{7}$$

$$F_1 = \frac{2PR}{P+R} \tag{8}$$

Kappa

$$\kappa = \frac{P_{\text{observed}} - P_{\text{expected}}}{1 - P_{\text{expected}}} \quad (9)$$

	коты	собаки	
коты	20	5	25
собаки	10	15	25
	30	20	

Kappa

$$\kappa = \frac{P_{\text{observed}} - P_{\text{expected}}}{1 - P_{\text{expected}}} \quad (9)$$

	коты	собаки	
коты	20	5	25
собаки	10	15	25
	30	20	

$$P_{\text{observed}} = (20 + 15)/50 = 0.7 \quad (10)$$

$$P_{\text{expected}} = ((25 * 30)/50 + (25 * 20)/50))/50 = (15 + 10)/50 = 0.5 \quad (11)$$

$$\kappa = \frac{0.7 - 0.5}{1 - 0.5} = 0.4 \quad (12)$$