Введение в Data Science
Занятие 6. Кластеризация

Николай Анохин Михаил Фирулик

9 апреля 2014 г.

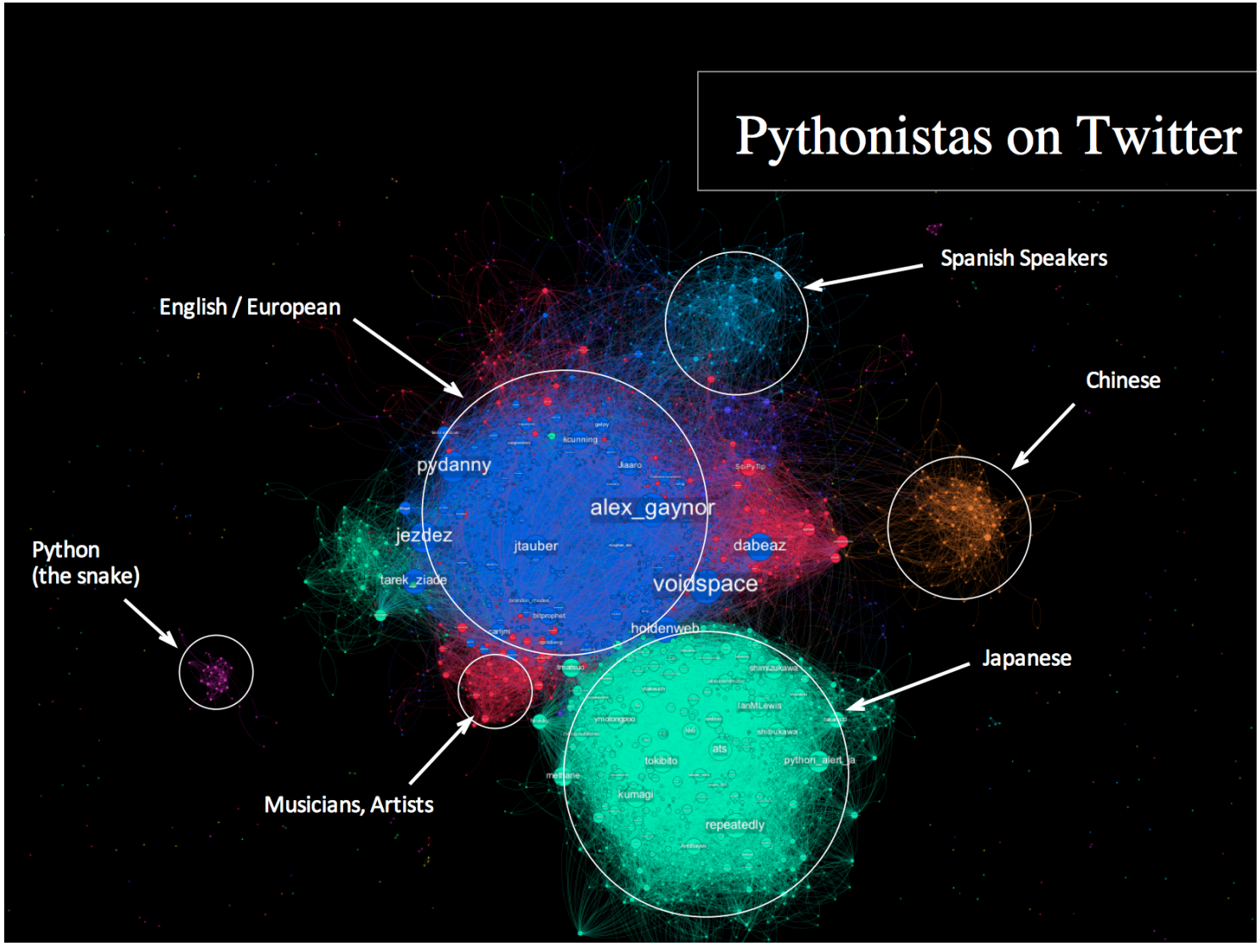
План занятия

Задача кластеризации

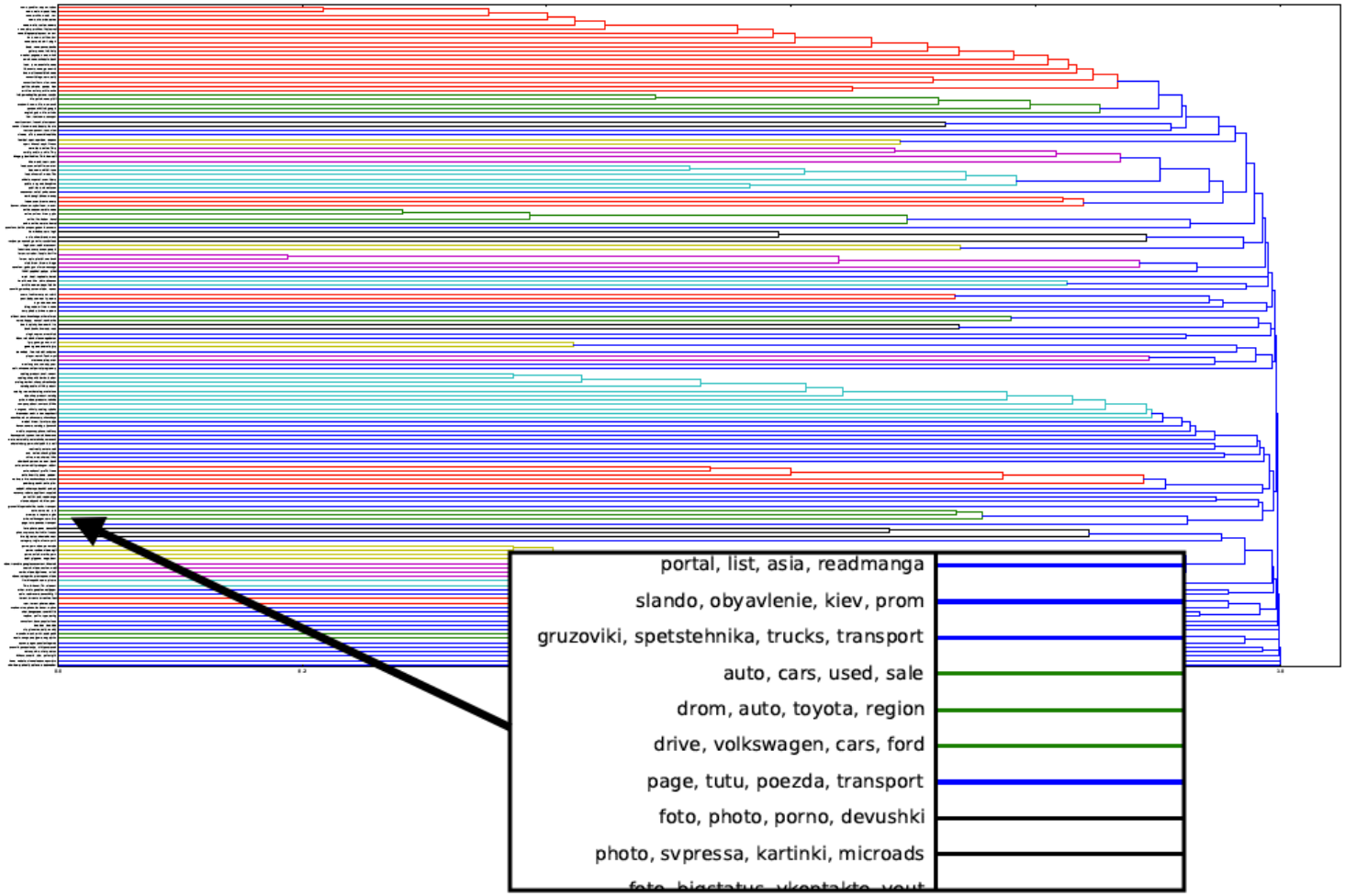
Функции расстояния

Критерии качества кластеризации

Python-программисты (by Gilad Lotan)



Слова и топики



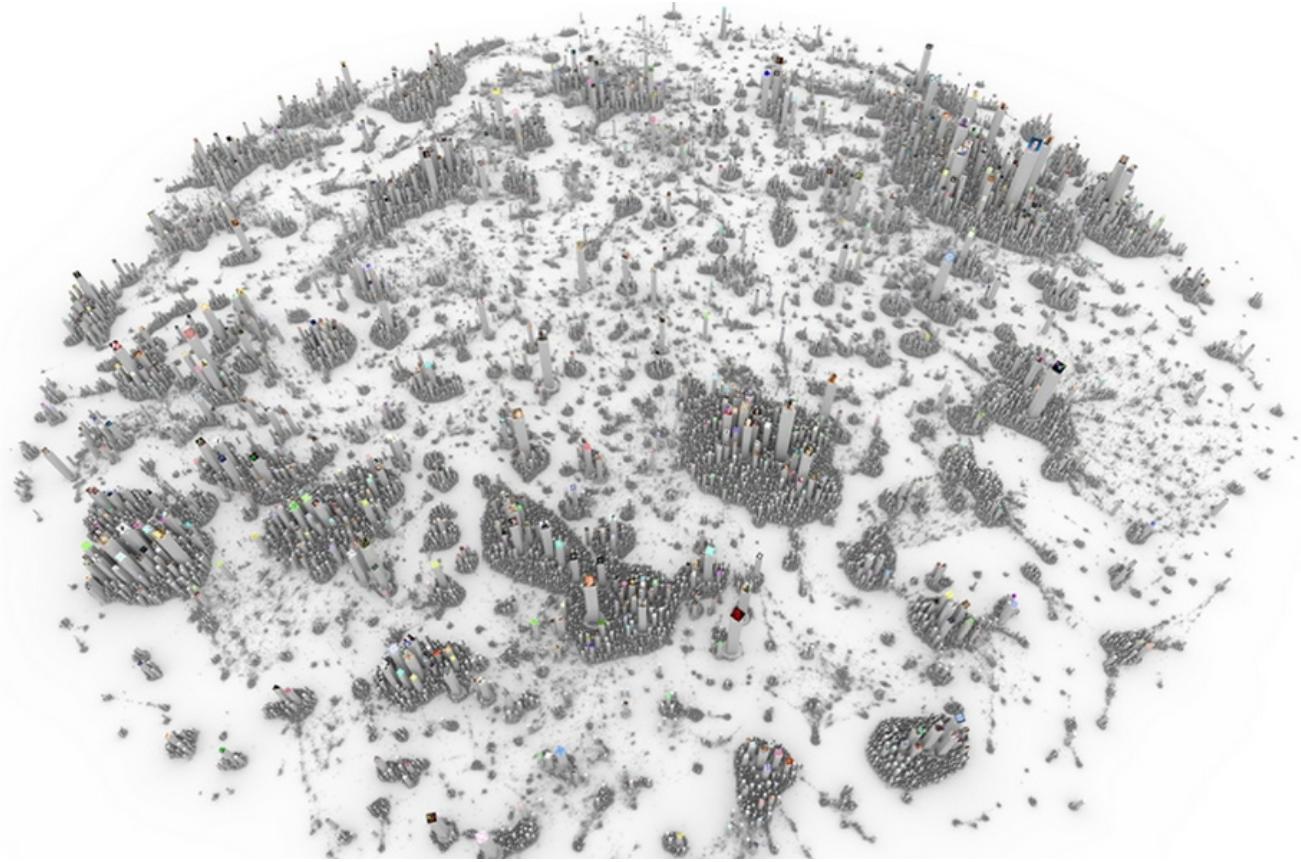
Интуиция

Кластерный анализ

Разбиение множества обучающих объектов на непересекающиеся подмножества (кластеры) с использованием некоторой функции расстояния между объектами так, чтобы любые два объекта, лежащие в одном кластере были схожи, а любые два объекта, лежащие в разных кластерах существенно различались.

Зачем

- расширение обучающей выборки
- ручная разметка объектов
- конструирование признаков
- суммаризация данных
- визуализация данных



Формально

Дано:

- Обучающая выборка $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$, где $\mathbf{x}_j$ – объект, принадлежащий некоторому множеству $\mathbf{X}$
- На $\mathbf{X}$ определена функция расстояния (схожести)

Найти: Разбиение $f : \mathbf{X} \to C$, где $C = \{1, \ldots, K\}$ – множество идентификаторов K кластеров.

Функция расстояния

Def

Функция $d(\mathbf{x}, \mathbf{y}) : \mathbf{X} \times \mathbf{X} \to R$ является функцией расстояния, определенной на пространстве $\mathbf{X}$ тогда и только тогда, когда $\forall \mathbf{x} \in \mathbf{X},\ \forall \mathbf{y} \in \mathbf{X},\ \forall \mathbf{z} \in \mathbf{X}$ выполнено:

1. $d(\mathbf{x}, \mathbf{y}) \geq 0$
2. $d(\mathbf{x}, \mathbf{y}) = 0 \Leftrightarrow \mathbf{x} = \mathbf{y}$
3. $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$
4. $d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{y}, \mathbf{z})$

Расстояния 1

- Минковского

$$d_r(\mathbf{x}, \mathbf{y}) = \left[\sum_{j=1}^{N} |x_j - y_j|^r \right]^{\frac{1}{r}}$$

- Евклидово $r = 2$

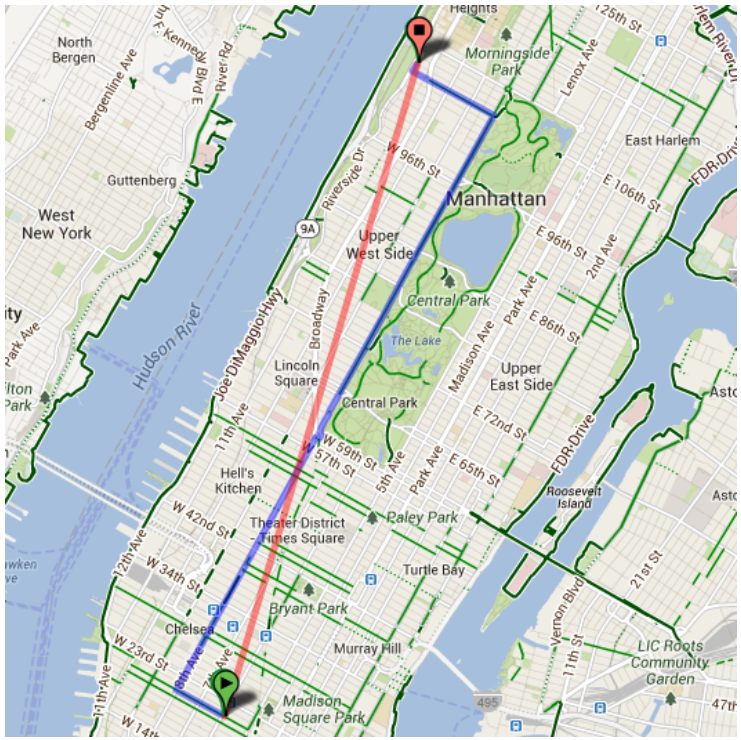
$$d_E(\mathbf{x}, \mathbf{y}) = d_2(\mathbf{x}, \mathbf{y})$$

- Манхэттэн $r = 1$

$$d_M(\mathbf{x}, \mathbf{y}) = d_1(\mathbf{x}, \mathbf{y})$$

- $r = \infty$

$$d_\infty(\mathbf{x}, \mathbf{y}) = \max_j |x_j - y_j|$$



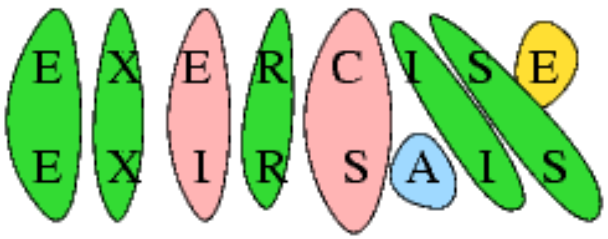
Проблема

Функции расстояния чувствительны к преобразоаниям данных

Решение

- Преобразовать обучающую выборку так, чтобы признаки имели нулевое среднее и единичную дисперсию – инвариантность к растяжению и сдвигу (stanartize)
- Преобразовать обучающую выборку так, чтобы оси совпадали с главными компонентами матрицы ковариации – инвариантность относительно поворотов (PCA)

Расстояния 2



- Жаккар

$$d_J(\mathbf{x}, \mathbf{y}) = 1 - \frac{|\mathbf{x} \cap \mathbf{y}|}{|\mathbf{x} \cup \mathbf{y}|}$$

- Косинус

$$d_c(\mathbf{x}, \mathbf{y}) = \arccos \frac{\mathbf{x}\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}$$

- Правки
  d_e – *наименьшее количество удалений и вставок, приводящее* $\mathbf{x}$ *к* $\mathbf{y}$.
- Хэмминг
  d_H – *количество различных компонент в* $\mathbf{x}$ *и* $\mathbf{y}$.

Проклятие размерности

Задача

Даны два случайных вектора $\mathbf{x}$ и $\mathbf{y}$ в пространстве размерности D. Как зависит математическое ожидание косинус-расстояния между $\mathbf{x}$ и $\mathbf{y}$ от размерности D?

$$d_c(\mathbf{x}, \mathbf{y}) = \arccos \frac{\sum_{j=1}^{D} x_j y_j}{\sum_{j=1}^{D} x_j^2 \sum_{j=1}^{D} y_j^2}$$

Наблюдения:

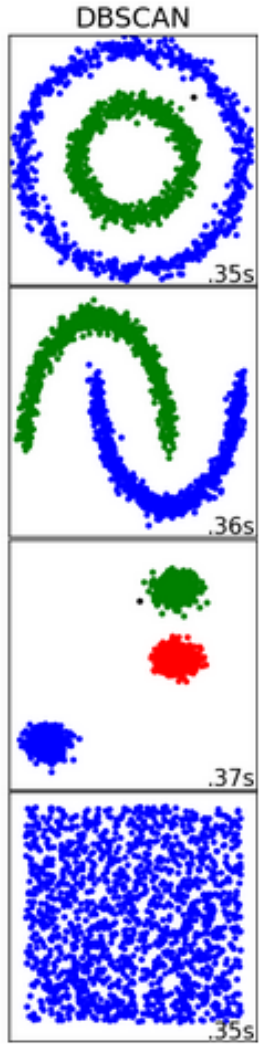
- числитель стремится к нулю
- знаменатель положительный

Вывод: $d_c(\mathbf{x}, \mathbf{y}) \to \frac{\pi}{2}$.

Выбор разбиения

Идея

Определить критерий качества кластеризации J и выбрать разбиение выборки на кластеры, которое которое соответствует оптимальному значению этого критерия.



Квардатичная ошибка

Среднее k-го кластера

$$\mathbf{m}_k = \frac{1}{n_k} \sum_{\mathbf{x}_i \in C_k} \mathbf{x}_i$$

Критерий

$$J_E = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} \|\mathbf{x}_i - \mathbf{m}_k\|^2 \to \min$$

Предпочтение кластерам близких размеров

$J_e = \text{large}$

$J_e = \text{small}$

Обобщение квадратичной ошибки

Критерий

$$
\begin{aligned}
J_E &= \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} \|\mathbf{x}_i - \mathbf{m}_k\|^2 = \\
&= \frac{1}{2} \sum_{k=1}^{K} n_k \left[\frac{1}{n_k^2} \sum_{\mathbf{x}_i \in C_k} \sum_{\mathbf{x}_j \in C_k} \|\mathbf{x}_i - \mathbf{x}_j\|^2 \right] = \\
&= \frac{1}{2} \sum_{k=1}^{K} n_k \left[\frac{1}{n_k^2} \sum_{\mathbf{x}_i \in C_k} \sum_{\mathbf{x}_j \in C_k} s(\mathbf{x}_i, \mathbf{x}_j) \right] = \frac{1}{2} \sum_{k=1}^{K} n_k \bar{s}_k
\end{aligned}
$$

Примеры $\bar{s}_i$

$$
\bar{s}_k = \min_{\mathbf{x}_i, \mathbf{x}_j} s(\mathbf{x}_i, \mathbf{x}_j); \quad \bar{s}_k = \max_{\mathbf{x}_i, \mathbf{x}_j} s(\mathbf{x}_i, \mathbf{x}_j)
$$

Матрицы разброса

Матрица разброса внутри кластеров

$$S_k = \sum_{\mathbf{x}_i \in C_k} (\mathbf{x}_i - \mathbf{m}_k)(\mathbf{x}_i - \mathbf{m}_k)^\top; \quad S_W = \sum_{k=1}^{K} S_k$$

Матрица разброса между кластерами

$$S_B = \sum_{k=1}^{K} n_k (\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^\top$$

Матрица разброса

$$S_T = \sum_{\mathbf{x}_i \in \mathbf{X}} (\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^\top$$

$$\boxed{S_T = S_W + S_B}$$

Критерии

- След

$$J_E = \text{tr} S_W = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_K} \|\mathbf{x}_i - \mathbf{m}_k\|^2 \to \min$$

- Детерминант (инвариант относительно растяжения)

$$J_d = \det S_W \to \min$$

- Инварианты относительно линейных преобразований

$$J_l = \text{tr} S_W^{-1} S_B = \sum_{j=1}^{d} \lambda_j \to \max, \quad J_f = \sum_{j=1}^{d} \frac{1}{1+\lambda_j} \to \min$$

$\lambda_1, \ldots, \lambda_d$ – собственные числа $S_W^{-1} S_B$

Итеративный алгоритм

Критерий

$$J_k = \sum_{\mathbf{x}_i \in C_k} \|\mathbf{x}_i - \mathbf{m}_k\|^2, \quad \mathbf{m}_k = \frac{1}{n_k} \sum_{\mathbf{x}_i \in C_k} \mathbf{x}_i$$

Переносим $\hat{\mathbf{x}}$ из кластера l в кластер k

$$\mathbf{m}_k^* = \mathbf{m}_k + \frac{\hat{\mathbf{x}} - \mathbf{m}_k}{n_k + 1}, \quad J_k^* = J_k + \frac{n_k}{n_k + 1} \|\hat{\mathbf{x}} - \mathbf{m}_k\|^2$$

$$\mathbf{m}_l^* = \mathbf{m}_l - \frac{\hat{\mathbf{x}} - \mathbf{m}_l}{n_l - 1}, \quad J_l^* = J_l - \frac{n_l}{n_l - 1} \|\hat{\mathbf{x}} - \mathbf{m}_l\|^2$$

Перенос имеет смысл, если

$$\frac{n_l}{n_l - 1} \|\hat{\mathbf{x}} - \mathbf{m}_l\|^2 > \frac{n_k}{n_k + 1} \|\hat{\mathbf{x}} - \mathbf{m}_k\|^2$$

Итеративный алгоритм

```
cluster(X, K):
  инициализируем C_1, ..., C_K, n_1, ..., n_K
  do:
    случайно выбираем x̂ ∈ X
    if n_l > 1:
      k* = arg min_k J*
      перемещаем x̂ в k*
  until кластеры стабильны или превышено число итераций
```

+ Обучение online
— Локальная оптимизация
— Зависимость от порядка рассмотрения $\mathbf{x}$

Качество кластеризации

Задача

Пусть дана обучающая выборка, для которой правильная кластеризация C известна. С помощью выбранного алгоритма получена кластеризация K. Проверить, насколько K совпадает с C.

- Rand Index

  a – кол-во пар объектов, попавших в один кластер и в C, и в K

  b – кол-во пар объектов, попавших в разные кластеры и в C, и в K

$$RI = \frac{a + b}{C_2^N}$$

- Mutual Information

$$MI = \sum_{c \in C} \sum_{k \in K} p(c, k) \log \frac{p(c, k)}{p(k)p(c)}$$

Зоопарк алгоритмов

По способу формирования кластеров

- иерархические (hierarchical, agglomerative)
- поточечные (point assignment)

По типу пространства $\mathbf{X}$

- $\mathbf{X}$ – евклидово
- $\mathbf{X}$ – не евклидово

По требованиям к памяти

- обучающая выборка должна помещаться в основную память
- поддерживает обработку данных кусками

Задача модуля

Классифицировать пользователей социальных сетей с использованием реализованных в ДЗ алгоритмов классификации и регрессии

Презентация 12.04.2014

1. Описание задачи (1-2 слайда)
   - 1.1 Количество объектов в выборке
   - 1.2 Распределение целевой переменной
   - 1.3 Метод тестирования алгоритмов
2. Каждый участник группы представляет алгоритм (1-2 слайда)
   - 2.1 Использованные признаки: распределения, преобразования
   - 2.2 Алгоритм: реализация и выбор параметров
   - 2.3 Метрики качества: результаты
3. Итоговый выбор (1-2 слайда)
   - 3.1 Какой выбран алгоритм
   - 3.2 Труднсти

Итог: 6-12 слайдов на группу, 15 мин на доклад, 30 (20 + 10) баллов

ДЗ

Использование готового алгорима = половина баллов (если не оговорено обратное)

Общие замечания

1. Дедлайны
2. Интерфейс классификаторов: fit/predict/predict_proba

Пожелания

1. PEP-8
2. Коменты
3. Адекватный текст коммитов

Обсуждение

1. Процедура сабмита ДЗ
2. Размер группы

Практика: итеративный алгоритм

1. Скачать ветку `clust` из репозитория и убедиться, что все работает
2. Проверить, что предложенный алгоритм кластеризации не является глобальным
3. Реализовать n-restart алгоритма для обеспечения глобальности
4. Поэкспериментировать с раздичными линейными преобразованиями пространства
5. (*) Реализовать критерий оптимальности, инвариантный относительно преобразований