ТЕХНОСФЕРА

Лекция 1
Задачи Data Mining

Николай Анохин

30 сентября 2014 г.

План занятия

Задача кластеризации

Статистический подход: смесь распределений

Задача кластеризации

В задачах кластеризации целевая переменная не задана. Цель – отыскать “скрытую структуру” данных.

Зачем вообще рассматривать задачи без целевой переменной?

1. разметка данных – дорогое удовольствие
2. можно сначала поделить, а потом разметить
3. возможность отслеживать эволюционные изменения
4. построение признаков
5. exploratory data analysis

Пример 1

Пример 2

Топ 1000 самых посещаемых доменов рунета

T-SNE + DBSCAN

Постановка задачи

Дано. N обучающих D-мерных объектов $\mathbf{x}_i \in \mathcal{X}$, образующих тренировочный набор данных (training data set) X.

Найти. Модель $h^*(\mathbf{x})$ из семейства параметрических функций $H = \{h(\mathbf{x}, \theta) : \mathcal{X} \times \Theta \to \mathbb{N}\}$, ставящую в соответствие произвольному $\mathbf{x} \in \mathcal{X}$ один из K кластеров так, чтобы объекты внутри одного кластера были похожи, а объекты из разных кластеров различались.

- Как определить похожесть объектов?
- Как оценить качество модели?
- Как выбрать K?

Многомерное нормальное распределение

Вопросы