Лекция 4
Data Science

Автоматическое формирование знаний

Data Mining – процесс обнаружения в «сырых» данных ранее неизвестных нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Этапы автоматического формирования знаний

Шаг 1. Подготовка исходного набора данных.

Создание набора данных, возможно, из различных источников, выбор обучающей выборки.

Шаг 2. Предобработка данных.

Данные могут содержать грамматические ошибки, аномальные значения и т.д.

Этапы автоматического формирования знаний

Шаг 3. Трансформация, нормализация данных.

Необходим для методов, которые работают с исходными данными определенного вида. Например, нейронные сети работают только с числовыми данными.

Шаг 4. Применение методов формирования знаний.

Применяются различные методы формирования знаний: статистические, нейронные сети и т.д.

К задачам формирования знаний относятся:

- прогнозирование;
- идентификация функций;
- классификация и кластеризация;
- фазификация нечетких переменных.

Для решения этих задач используются методы прикладной статистики.

Этапы автоматического формирования знаний

Шаг 5. Постобработка данных.

Интерпретация результатов и применение полученных знаний в бизнес-приложениях.

Метод деревьев решений (деревьев классификации).

Позволяет предсказывать принадлежность наблюдений или объектов к тому или иному классу в зависимости от соответствующих значений атрибутов, характеризующих эти наблюдения.

Метод деревьев решений (деревьев классификации).

Деревья решений обеспечивают автоматическое построение продукционных правил «если, …, то …» по имеющейся статистике, на основании которых в дальнейшем выносится решение о принадлежности наблюдения или объекта к тому или иному классу.

Пусть имеется совокупность n объектов, представленных множеством $T = \{t_1, t_2, \ldots t_n\}$, где каждый элемент этого множества описывается одним и тем же набором признаков (атрибутов) с именами C_i, $i=1,\ldots m$.

Каждый атрибут может принимать k_i значений - x_{ip}, p=1,…, k_i, измеряемых в произвольной шкале.

Пример.

Рассмотрим статистику по клиентам некоторого банка.

Тогда клиенты – это множество T.

Каждый клиент характеризуется набором характеристик: полом, возрастом, целью кредитования, совокупным доходом и т.п.

Это атрибуты C_1, C_2, C_3 и т.д.

Атрибут C_1 может принимать 2 значения: М и Ж, т.е. x_{11}=М, x_{12}=Ж и т.д.

Пусть имеется множество классов K_j , j=0,…J.

При этом каждый объект множества T (каждый клиент банка был отнесен к некоторому классу объектов K_j и это отражено в статистике.

Например, в случае с клиентами банка это могут быть два класса:

K_1 («заемщик вовремя обслуживает кредит, с такими характеристиками кредит можно выдавать»),

K_2 («заемщик неудовлетворительно обслуживает кредит, с такими характеристиками кредит нельзя выдавать»).

Требуется построить классифицирующие (продукционные) правила, позволяющие выявить закономерности между значениями атрибутов каждого объекта множества T и классом K_j, к которому объект относится.

Классифицирующее правило имеет вид:

«*если признаки объекта* t_i $(i=1,\ldots,n)$ *принимают значения*

$C_1 = x_{1p}$ *и* $C_2 = x_{2p}$ *и …* *и* $C_n = x_{1n}$,

то t_i *относится к классу* K_j»