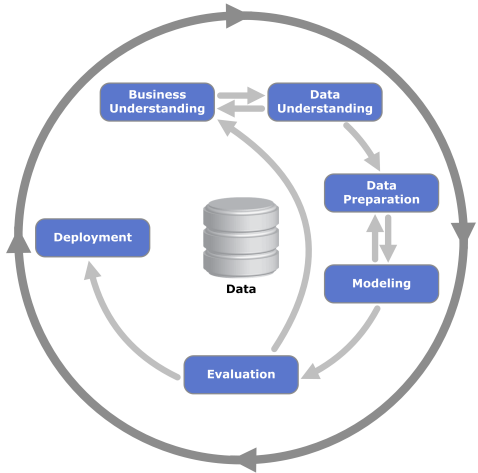
Процессы Data Mining

Николай Анохин

CRISP-DM

Cross Industry Standard Process for Data Mining



Игра в гольф[1]

Business understanding

- понимание задачи с точки зрения бизнеса
- сбор требований и ограничений
- постановка задачи в терминах Data Mining

$\mathcal{D}$ – множество, содержащее все рассматриваемые в задаче объекты

$f : \mathcal{D} \to \mathcal{Y}$ – целевая функция

Цель – с использованием данных о конечном множестве объектов из $\mathcal{D}$ (data set) научиться предсказывать значения целевой функции для любых объектов из $\mathcal{D}$

Задача **с учителем** – для "известных" объектов дано значение целевой функции, иначе – задача **без учителя**.

[1] Induction of Decision Trees / R. Quinlan

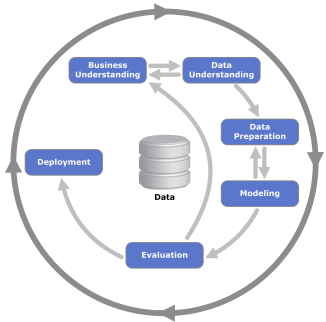
Игра в гольф

Data understanding

- первичный сбор данных
- ознакомление с данными и понимание их специфики

Data preparation

- формирование финального набора данных



Признаки

$\mathcal{D}$ – множество, содержащее все рассматриваемые в задаче объекты
$d \in \mathcal{D}$ – объект, $\phi_j : \mathcal{D} \to F_j$ – признак

Виды признаков

- Бинарные/Binary
  $F_j = \{true, false\}$
- Номинальные/Categorical
  F_j – конечное
- Порядковые/Ordinal
  F_j – конечное, упорядоченное
- Количественные/Numerical
  $F_j = \mathbb{R}$

Признаковое представление объекта d

$$\mathbf{x} = (\phi_1(d), \ldots, \phi_m(d)) \in \mathcal{X}$$

Игра в гольф: признаки

Outlook	Temperature	Humidity	Wind	Play
Sunny	85	85	false	no
Sunny	80	90	true	no
Overcast	83	86	false	yes
Rainy	70	96	false	yes
Rainy	68	80	false	yes
Rainy	65	70	true	no
Overcast	64	65	true	yes
Sunny	72	95	false	no
Sunny	69	70	false	yes
Rainy	75	80	false	yes
Sunny	75	70	true	yes
Overcast	72	90	true	yes
Overcast	81	75	false	yes
Rainy	71	91	true	no

Моделирование

- перебор различных моделей
- настройка параметров моделей

Модель

признаковое описание объекта d:

$$\mathbf{x} = (\phi_1(d), \ldots, \phi_m(d)) \in \mathcal{X}$$

значение целевой функции для объекта d: $f(d) = y \in \mathcal{Y}$

модель – семейство функций вида

$$H = \{h(\mathbf{x}, \theta) : \mathcal{X} \times \Theta \to \mathcal{Y}\},$$

где $\theta \in \Theta$ – неизвестный вектор параметров

Виды моделей

Качество вина

признаковое описание: $\mathbf{x} \in \mathbb{R}^1$

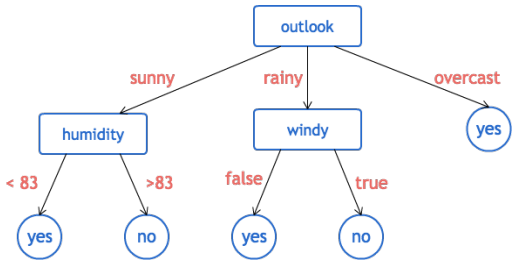
целевая переменная: $y = 1$, если вино хорошее, $y = 0$ иначе

модель:

$$\begin{cases} p(\mathbf{x}|\text{good}) \sim \mathcal{N}(\mathbf{x}|\mu_g, \sigma_g), \quad p(\text{good}) = \frac{1}{2} \\ p(\mathbf{x}|\text{bad}) \sim \mathcal{N}(\mathbf{x}|\mu_b, \sigma_b), \quad p(\text{bad}) = \frac{1}{2} \end{cases} \quad + \quad y = \mathcal{I}(p(\text{good}|\mathbf{x}) > p(\text{bad}|\mathbf{x}))$$

параметры: $\theta = (\mu_g, \sigma_g, \mu_b, \sigma_b)$

Дерево решений



Outlook	Temperature	Humidity	Wind	Play
Sunny	85	85	false	no
Sunny	80	90	true	no
Overcast	83	86	false	yes
Rainy	70	96	false	yes
Rainy	68	80	false	yes
Rainy	65	70	true	no
Overcast	64	65	true	yes
Sunny	72	95	false	no
Sunny	69	70	false	yes
Rainy	75	80	false	yes
Sunny	75	70	true	yes
Overcast	72	90	true	yes
Overcast	81	75	false	yes
Rainy	71	91	true	no

Обучение модели

- дана обучающая выборка (data set) $X = \{\mathbf{x}_1, \ldots, \mathbf{x_N}\}$
- для каждого из объектов обучающей выборки дано значение целевой функции $Y = \{y_1, \ldots, y_N\}$ (если задача с учителем)

Алгоритм обучения

Выбор наилучших параметров θ^* с использованием обучающей выборки

$$A(X, Y) : (\mathcal{X} \times \mathcal{Y})^N \to \Theta$$

В итоге:

$$h^*(\mathbf{x}) = h(\mathbf{x}, \theta^*)$$

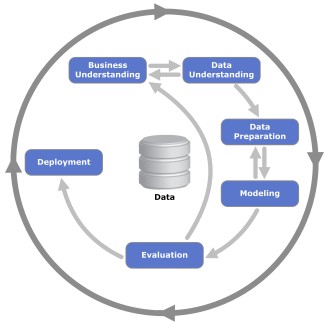
Пример 2. Игра в гольф

Evaluation

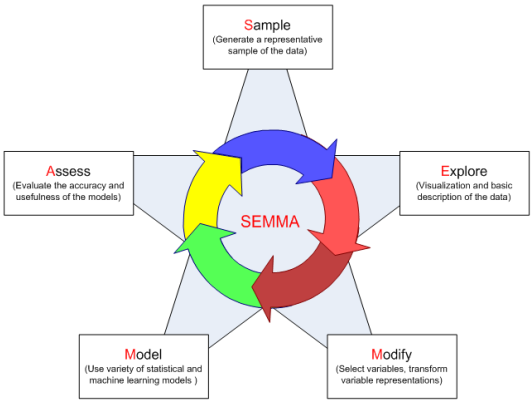
- тщательная проверка качества модели
- подробное рассмотрение шагов, предпринятых при построении
- поиск бизнес-требований, которые не удоволетворены

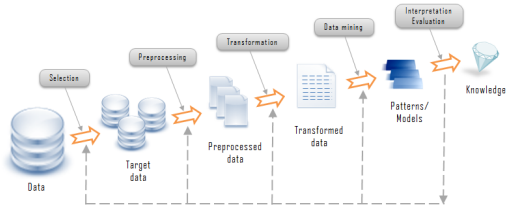
Deployment

- презентация модели клиенту
- развертывание и использование модели



Другие процессы: SEMMA[2], KDD[3]





[2] http://timkienthuc.blogspot.ru/2012/04/crm-and-data-mining-day-08.html
[3] http://www.rithme.eu/