МГТУ
им. Н.Э. Баумана
ТЕХНОПАРК
Mail.Ru Group

ПРОГРАММИРОВАНИЕ CUDA C/C++, АНАЛИЗ ИЗОБРАЖЕНИЙ И DEEP LEARNING

Лекция №5

Спасёнов Алексей

Введение в Машинное обучение

Часть первая

1. Основный понятия
2. Основные типы задач
3. Примеры прикладных задач
4. Линейные модели

Рекомендуемая литература

1. Christopher M.Bishop. Pattern recognition and Machine Learning
2. Kevin P. Murphy. Machine Learning. A Probabilistic Perspective

Техностpим Mail.ru Group:

1. Введение в анализ данных
2. Data Mining
3. Методы обработки больших объёмов данных

Машинное обучение (Machine Learning)

Обширный подраздел прикладной математики, находящийся на стыке математической статистики, оптимизации, искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться по эмпирическим (прецедентным) данным.

Анализ данных (Data Mining)

Процесс извлечения знаний из различных источников данных, таких как базы данных, текст, картинки, видео и т.д. Полученные знания должны быть достоверными, полезными и интерпретируемыми.

Анализ данных (Data Mining)

Процесс построения модели, хорошо описывающей закономерности, которые порождают данные.

Подходы к построению моделей:

1) Статистический
2) На основании машинного обучения
3) Вычислительный

Введение в Машинное обучение

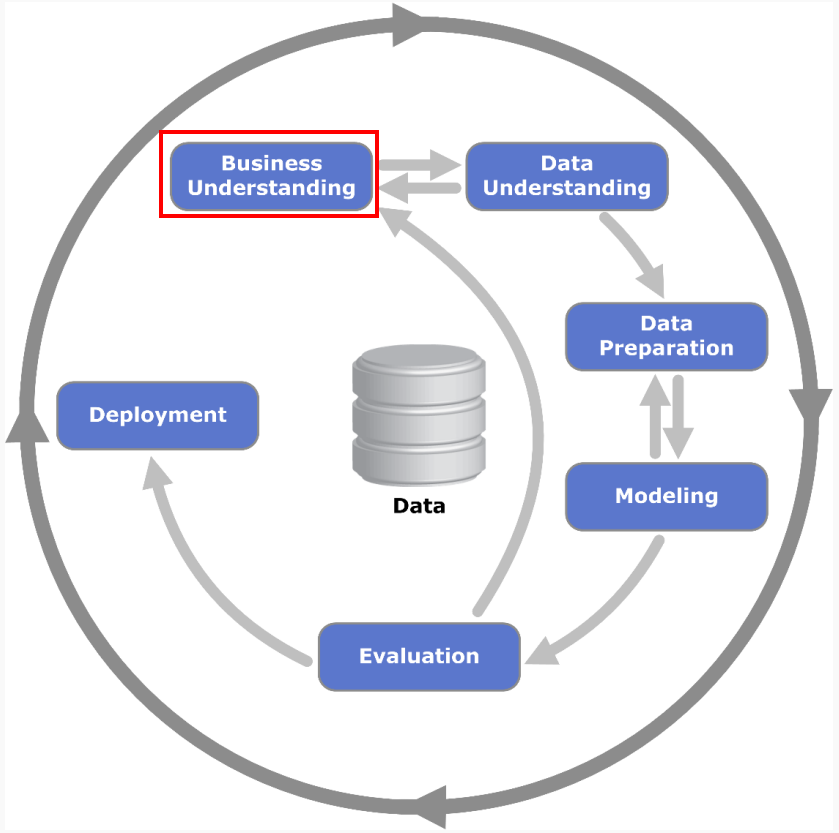
Cross Industry Standard Process for Data Mining

Введение в Машинное обучение

Постановка задачи

1) Распознаванию марки и модели автомобилей по изображениям
2) Информационный поиск, анализ текстов
3) Медицинская диагностика

…



Введение в Машинное обучение

Признаки (Features)

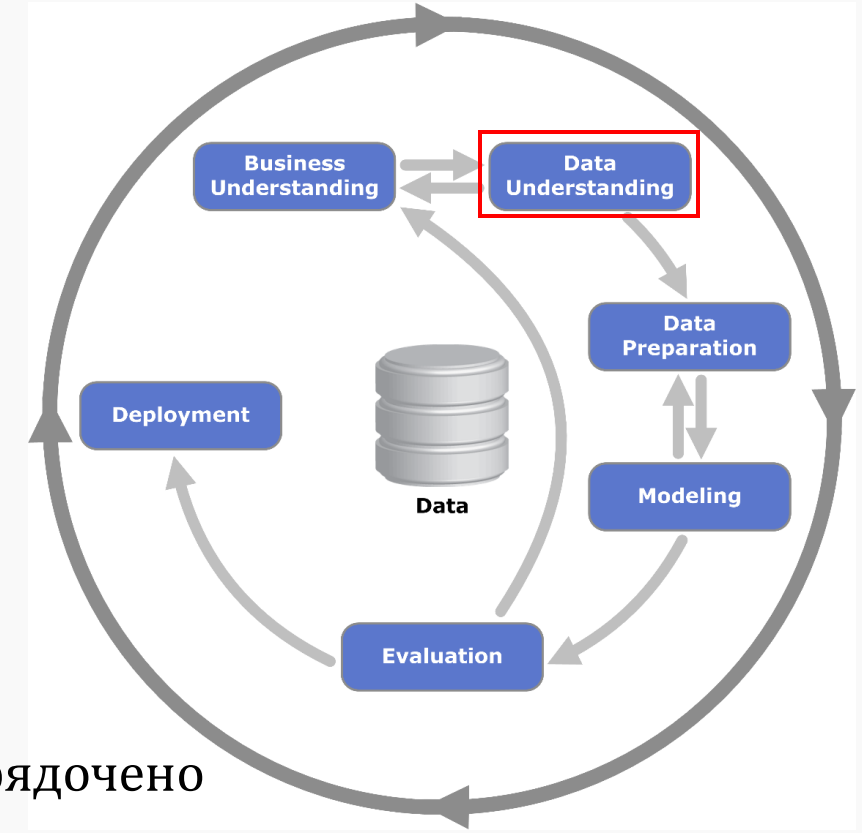
D – множество объектов (Data set)

$d \in D$ – обучающий объект

$\phi_i : D \to F_j$ - признак

Виды признаков:

1) Бинарные	Binary	$F_j = \{true, false\}$
2) Номинальные	Categorical	F_j – конечно
3) Порядковые	Ordinal	F_j – конечно упорядочено
4) Количественные	Numerical	$F_j = \mathbb{R}$



Введение в Машинное обучение

Подготовка данных (Data Preparation)

1)Удаление шума

2)Заполнение отсутствующих значений

3)Трансформация значений

4)Выбор факторов

5)Использование априорных знаний

Создание модели (Modeling)

Модель

Семейство параметрических функций вида

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

Алгоритм обучения

Выбор наилучших параметров Θ

$$A(X, Y): (X \times Y)^N \rightarrow \Theta$$

В итоге:

$$h^*(x) = h(x, \Theta^*)$$

Введение в Машинное обучение

Основные типы задач

1) Обучение с учителем (supervised learning)

Каждый прецедент представляет собой пару «объект, ответ». Требуется найти функциональную зависимость ответов от описаний объектов и построить алгоритм, принимающий на входе описание объекта и выдающий на выходе ответ.

2) Обучение без учителя (unsupervised learning)

Ответы не задаются, и требуется искать зависимости между объектами

3) Частичное обучение (semi-supervised learning)

Каждый прецедент представляет собой пару «объект, ответ», но ответы известны только на части прецедентов.

4) Обучение с подкреплением (reinforcement learning)

Роль объектов играют пары «ситуация, принятое решение», ответами являются значения функционала качества, характеризующего правильность принятых решений (реакцию среды).

…

Обучение с учителем (обучения по прецедентам)

Задачи классификации (classification)

- $F_j = \{true, false\}$ – классификация на 2 класса
- $F_j = \{1, \dots, M\}$ – классификация на M непересекающихся классов
- $F_j = \{0,1\}^M$ - классификация на M классов, которые могут пересекаться

Задача восстановления регрессии (regression)

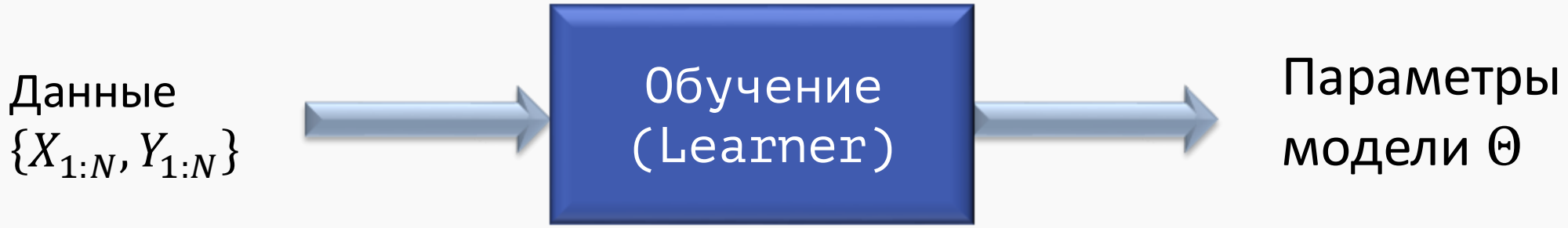
- $F_j = \mathbb{R}$ или $F_j = \mathbb{R}^M$ (ответом является действительное число или числовой вектор)

Задача ранжирования (learning to rank)

- F_j - конечно упорядочено (ответы надо получить сразу на множестве объектов, после чего отсортировать их по значениям ответов)

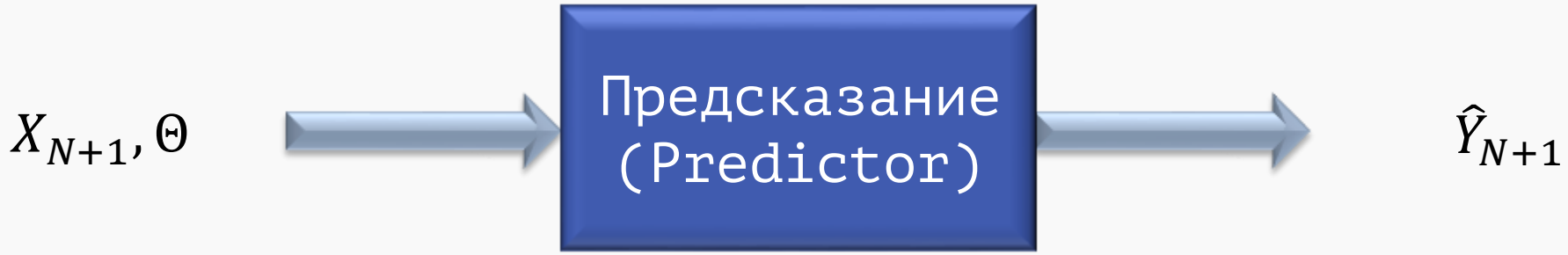
Введение в Машинное обучение

Этап обучения (train)



Необходимо учитывать представительность выборки

Этап применения (test)



Пример модели

Линейная модель (задача восстановления регрессии)

$$\hat{y}(X_i) = \Theta_1 + x_i\Theta_2$$

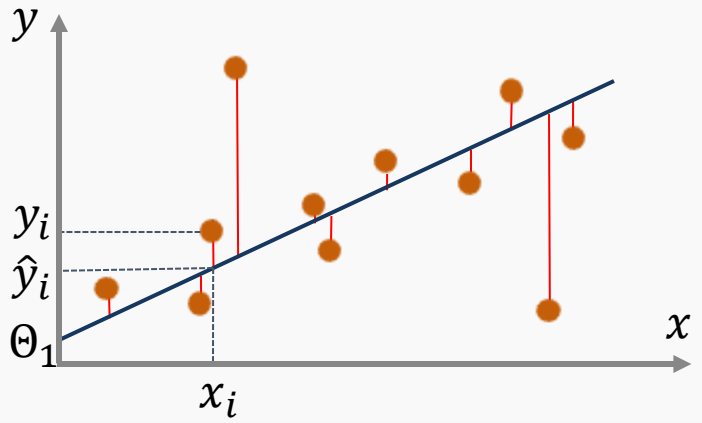
Целевая функция (Objective function, Energy, Loss)

Величина ошибки алгоритма на обучающей выборке

Пример для задачи регрессии

$$J(\Theta) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \Theta_1 - x_i\Theta_2)^2$$

Метод наименьших квадратов (Ordinary Least Squares)



Пример модели

Линейная модель

$$\hat{y}_i = \sum_{j=1}^{d} x_{ij}\Theta_j = 1 * \Theta_1 + x_{i2}\Theta_2 + x_{i3}\Theta_3 + \ \dots + x_{id}\Theta_d$$

В матричной форме:

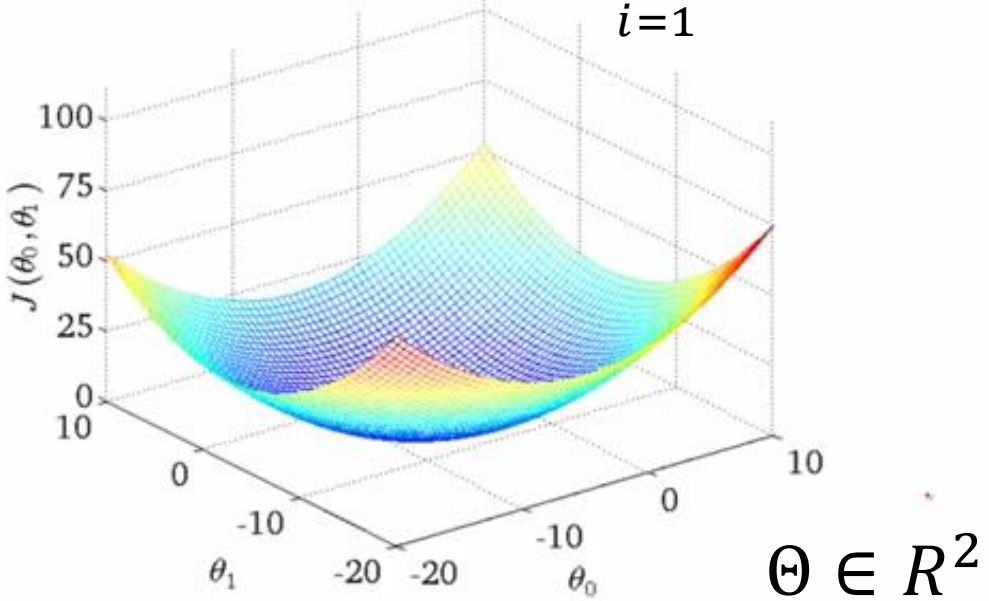
$$\hat{y} = X\Theta$$

$$\begin{bmatrix} \hat{y}_1 \\ \dots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{1d} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{nd} \end{bmatrix} \begin{bmatrix} \Theta_1 \\ \dots \\ \Theta_d \end{bmatrix}$$

Пример модели

Целевая функция

$$J(\Theta) = (y - X\Theta)^T(y - X\Theta) = \sum_{i=1}^{n}(y_i - X_i^T\Theta)^2$$



$\Theta \in R^2$

Пример модели

Линейная модель

$$J(\Theta) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \Theta_1 - x_i\Theta_2)^2$$

Поиск решения:

$$\frac{\partial J(\Theta)}{\partial \Theta} = \frac{\partial}{\partial \Theta}(y^T y - 2y^T x\Theta + \Theta^T x^T x\Theta) = 0$$

$$\Theta = (x^T x)^{-1} x^T y$$

Проблема переобучения

1) Обучающая выборка

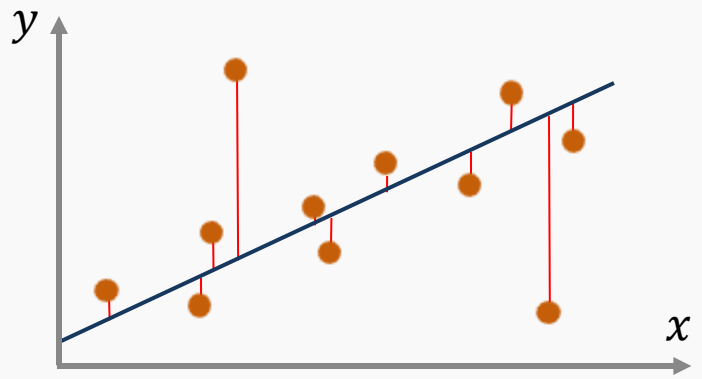
2) Контрольная выборка

Пример

Модель: $h(X, \theta) = \theta_0 + \theta_1 \cdot x + \cdots + \theta_n x^n$

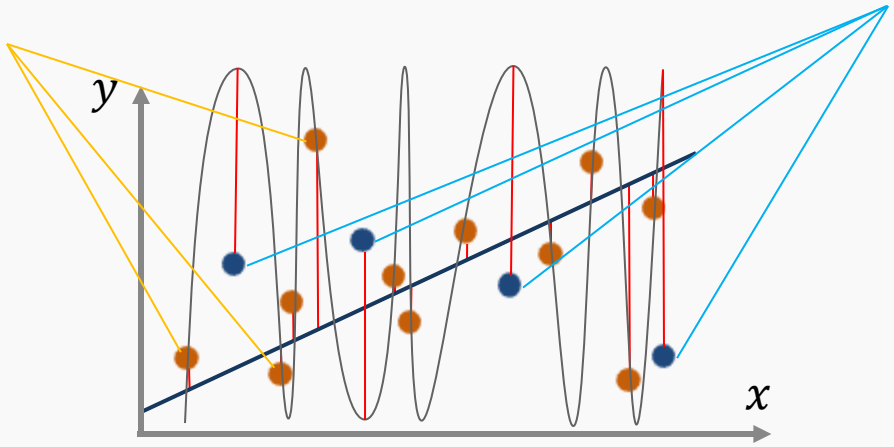
Целевая функция: $J(X, \Theta) = \sum_{i=0}^{n} (\theta_0 + \theta_1 \cdot x_i + \cdots + \theta_n x_i^n - y_i)^2$

Что будет, если увеличить n?



Проблема переобучения

Обучающая выборка
(Ошибка = 0)



Контрольная выборка

Контакты:
a.spasenov@corp.mail.ru
alex_spasenov (Skype)

Спасибо за внимание!