Structural approach to the deep learning method

NEC–2019, 30 September – 4 October, 2019 Budva, Montenegro

# Технологии

## Проблемы машинного обучения

* Технологии машинного обучения и нейросетей переоценены.
* Методы машинного обучения уничтожают культуру аналитического мышления.
* Для достижения результата в проектах, подразумевающих анализ данных, важнее знание предмета, нежели глубокие знания ML.
* Профессия Data Scientist’а сильно переоценена, универсальных специалистов больше не будет.

## Технологии переоценены

* Большинство задач, которые сейчас пытаются решать с помощью современных методов анализа данных и нейросетей – решаются уже давно.
* Задачи не новые. Их решают аналитики, которые разбираются в предметной области.
* Зачастую алгоритмы машинного обучения в таких системах уже заложены.
* Сделать тут что-то принципиально новое и реально применимое крайне сложно.
* Яблоки, упавшие с дерева, уже собраны.

## Культура аналитического мышления

* Нужно глубоко разобраться в предметной области.
  + какие данные нужны;
  + нужны ли какие-либо предсказательные алгоритмы;
  + есть ли возможность верифицировать предсказание.
* Требуется аналитический подход.
* Требуется культура работы с данными.
* Требуется умение ставить гипотезы.

## Уничтожение культуры аналитического мышления

Большинство современных Data Scientist’ов = дети на спорткаре

* считают себя уникальными;
* водить не умеют;
* едут быстро только потому, что сильное железо.

## Важнее знание предмета

Data Scientists:

* почти не задают никаких вопросов;
* данные и так обо всем расскажут;
* забирают какие-то данные;
* говорят, что построили какую-то модель.

Результат не проверяем.

## Универсальных специалистов больше не будет

Data Scientist

* не может быть универсалом;
* должен быть экспертом в предметной области.

## Универсальных специалистов больше не будет

Хайп закончился.

# Структура проекта

## Как устроен проект анализа данных

1. Требования к проекту
2. Данные проекта
3. Разработка и внедрение проекта

## Требования

* Мы ничего не знаем о том, какие у нас есть данные.
* Мы должны вникнуть в постановку задачи.
* Мы должны понять, какой результат требуется получить от проекта.
* Мы должны решить, каким методом задача будет решаться.
* Мы должны задать требования к данным.

## Данные

* Поиск данных для решения задачи:
  + мы узнаем, какие источники нам доступны;
  + мы формируем выборку, с которой в дальнейшем будем работать.
* Исследование данных:
  + исследовать центральное положение и вариабельность;
  + выявить корреляции между признаками;
  + построить графики распределения.
* Подготовка данных.

## Разработка и внедрение

* Разработка модели.
* Программная реализация модели.
* Прогонка обучающей выборки.
* Проверка на тестовой выборке.
* Верификация результата.
* Цикл (можно начинать все сначала).

# Требования

## Понимание задачи

* Фундамент всей работы.
* Необходимо четко определить цель исследования.
* Что является проблемой?
* По каким метрикам будет оцениваться успешность?

## Выбор аналитического подхода

* Выбор подхода зависит от того, какой тип ответа нужно получить в итоге:
  + если нужен ответ вида да/нет, подойдёт байесовский классификатор;
  + если нужен ответ в виде численного признака, то подойдут регрессионные модели;
  + если нужно определить вероятности определённых исходов, необходимо использовать предиктивную модель;
  + если нужно выявить связи, используется дескриптивный подход.

## Требования к данным

* Какие данные позволят дать искомый ответ?
* Требования к данным:
  + контент;
  + форматы данных;
  + источники данных.

# Данные

## Сбор данных

* Мы выполняем сбор данных из имеющихся источников.
* Убеждаемся, что источники:
  + доступны;
  + надёжны;
  + могут быть использованы для получения искомых данных в требуемом качестве.
* Необходимо понять, получили ли мы те данные, какие хотели.
* Пересмотр требований к данным.
* Принятие решения о необходимости дополнительных данных.
* Нахождение замены недостающим данным.

## Анализ данных

* Репрезентативны ли собранные данные относительно поставленной задачи?
* Описательная статистика применяется ко всем переменным, которые будут использоваться в выбранной модели:
  + исследуется центральное положение (среднее, медиана, мода);
  + ищутся выбросы и выполняется оценка вариабельности (дисперсия, стандартное отклонение);
  + строятся гистограммы распределения переменных;
  + применяются другие инструменты визуализации (например, ящики с усами).

## Анализ данных

* Вычисляются корреляции между переменными.
* Если найдутся значительные корреляции между переменными, некоторые переменные могут быть отброшены, как избыточные.

## Подготовка данных

Сбор и анализ данных + подготовка данных = 70%–90% времени проекта.

## Подготовка данных

* Мы перерабатываем данные в такую форму, чтобы с ними было удобно работать:
  + удаляем дубликаты;
  + обрабатываем отсутствующие или неверные данные;
  + проверяем и исправляем ошибки форматирования.
* Мы конструируем набор факторов, с которым на следующих этапах будет работать машинное обучение:
  + извлечение признаков;
  + отбор признаков.
* Ошибки на этом этапе могут оказаться критическими.
  + Избыточное количество признаков = модель переобучена.
  + Недостаточное количество признаков = модель недообучена.

# Разработка и внедрение

## Построение модели

Когда тип модели определён и имеется обучающая выборка, мы разрабатываем модель и проверяем её на наборе признаков.

## Применение модели

* Вычисления чередуются с настройкой модели.
* Отвечает ли построенная модель исходной задаче?
* Вычисление модели имеет две фазы:
  + проводятся диагностические измерения, которые помогают понять, работает ли модель, так как задумано;
  + проводится проверка статистической значимости гипотезы. Она необходима, чтобы убедиться, что данные в модели правильно используются и интерпретируются и полученный результат выходит за пределы статистической погрешности.

## Внедрение

* Внедрение проводится поэтапно:
  + ограниченная группа пользователей;
  + тестовое окружение.
* Система обратной связи.

Wer’s nicht glaubt, bezahlt einen Taler