Оригинал: Timeline for Data Science Competence | by Benjamin Obi Tayo Ph.D. | Towards AI

Вторая ссылка

https://pub.towardsai.net/timeline-for-data-science-competence-1b724e7977e0

Автор: https://benjaminobi.medium.com/

Бенджамин Оби Тайо

Преподаватель физики и инженерных дисциплин в Университете Эдмонта, Оклахома. Интересы: Data Science, машинное обучение, ИИ, Python, R, биофизика.

Формат	перевод
Автор	Цокто Жигмытов
Раздел	Программирование
SEO-Title	
Ключи	
Курс (название, ссылка)	Профессия Data Scientist
Внимание! Название — макс. 90 зн/пр., лид — макс. 150 зн/пр.	

Карта развития дата-сайентиста: с чего начать, к чему идти и сколько времени потребуется

На каком уровне находитесь вы и далеко ли до следующей ступеньки?

Каждый, кто заинтересовался наукой о данных, задаётся вопросом: а сколько времени понадобится, чтобы её изучить? Мы составили примерный график профессионального развития дата-сайентиста по трём уровням — базовый, средний и продвинутый. Чтобы было проще сравнивать с требованиями вакансий, привели их к принятым в ИТ терминам: стажёр (intern), джун (junior, младший), мидл (middle, средний) и сеньор (senior, старший).

Уровни дата-сайентиста рассмотрим на примере языка Python. Но вообще в Data Science используют и другие языки и платформы — R, Julia, SAS, MATLAB.

Дополнительно: Чтобы не запутаться в терминах, прочитайте нашу статью про Python-минимум, необходимый для первоначального погружения в Data Science. И будьте осторожны — дальше много чек-листов и перечислений :)

Уровень 1. От стажёра к джуну

Главное на этом уровне — научиться работать с датасетами в виде CSV-файлов, обрабатывать и визуализировать данные, понимать, что такое линейная регрессия.

Основы обработки данных

В первую очередь придётся манипулировать данными, чистить, структурировать и приводить их к единой размерности или шкале. От новичка ждут уверенной работы с библиотеками Pandas и NumPy и некоторых специальных навыков:

- Импорт и экспорт данных в CSV-формате.
- Очистка, предварительная подготовка, систематизация данных для анализа или построения модели.
- Работа с пропущенными значениями в датасете.
- Понимание принципов замены недостающих данных (импутации) и их реализация. Например, замена с помощью средних или медиан.
- Работа с категориальными признаками.
- Разделение датасета на обучающую и тестовую части.
- Нормировка данных с помощью нормализации и стандартизации.

- Уменьшение объёма данных с помощью техник снижения размерности. Например, метода главных компонент.

Визуализация данных

Новичок должен знать основные принципы хорошей визуализации и инструменты — в том числе Python-библиотеки matplotlib и seaborn (для R — ggplot2).

Какие компоненты нужны для правильной визуализации данных:

- **Данные**. Прежде чем решить, как именно визуализировать данные, надо понять, к какому типу они относятся: категориальные, численные, дискретные, непрерывные, временной ряд.
- **Геометрия**. То есть какой график вам подойдёт: диаграмма рассеяния, столбиковая диаграмма, линейный график, гистограмма, диаграмма плотности, «ящик с усами», тепловая карта.
- **Координаты**. Нужно определить, какая из переменных будет отражена на оси x, а какая — на оси y. Это важно, особенно если у вас многомерный датасет с несколькими признаками.
- **Шкала**. Решите, какую шкалу будете использовать: линейную, логарифмическую или другие.
- **Текст**. Всё, что касается подписей, надписей, легенд, размера шрифта и так далее.
- **Этика**. Убедитесь, что ваша визуализация излагает данные правдиво. Иными словами, что вы не вводите в заблуждение свою аудиторию, когда очищаете, обобщаете, преобразовываете и визуализируете данные.

Обучение с учителем: предсказание непрерывных переменных

Главное: стажёру придётся изучить методы регрессии, стать почти на «ты» с библиотеками scikit-learn и caret, чтобы строить модели линейной регрессии. Но чтобы стать полноценным джуниором, стажёр должен знать и уметь ещё кучу всего (осторожно — там сложные слова, но есть подсказки):

- Проводить простой регрессионный анализ с помощью NumPy или Pylab.
- Использовать библиотеку scikit-learn, чтобы решать задачи с множественной регрессией.
- Понимать методы регуляризации: метод LASSO, метод упругой сети, метод регуляризации Тихонова.

- Знать непараметрические методы регрессии: метод k-ближайших соседей и метод опорных векторов.
- Понимать метрики оценок моделей регрессии: среднеквадратичная ошибка, средняя абсолютная ошибка и коэффициент детерминации R-квадрат.
- Сравнивать разные модели регрессии.

А как вы хотели — сделать Терминатора непросто :)

Уровень 2. От джуна к мидлу

Прочно закрепив на практике все те неприличные слова из блока для джуна, можно штурмовать более продвинутые техники и методы: предсказание дискретных переменных в обучении с учителем (supervised learning), оценку и настройку моделей, а также сбор разных алгоритмов в единые ансамбли методов. Вы уже поняли, что сейчас опять начнётся ковровое бомбометание дата-сайентистскими терминами? Не вздумайте употреблять их в публичных местах — а то бабушки начнут креститься, как будто увидели сатаниста или парня с татуировками по всему телу :)

Обучение с учителем: предсказание дискретных переменных

Начните с алгоритмов бинарной классификации — вот какие из них надо знать мидлу:

- перцептрон.
- логистическая регрессия;
- метод опорных векторов;
- решающие деревья и случайный лес;
- k-ближайших соседей;
- наивный байесовский классификатор.

> **Дополнительно:** Небольшая статья о том, как создать простую модель машинного обучения. Формируем и делим датасет, обучаем модель Random Forest, предсказываем дискретную переменную и вот это всё.

Мастхэв — на хорошем уровне работать с библиотекой scikit-learn (она уже тут мелькала), которая помогает строить модели. Также придётся решать задачи на нелинейную классификацию с помощью метода опорных векторов; владеть несколькими метриками для оценки алгоритмов классификации — точность, погрешность, чувствительность, матрица ошибок, F-мера, ROC-кривая.

Оценка моделей и оптимизация гиперпараметров

Чтобы правильно оценивать и настраивать модели, специалисту нужно:

- Соединять трансформеры и модули оценки (estimators) в конвейеры машинного обучения (machine learning pipelines). К Оптимусу Прайму и Бамблби они отношения не имеют — пока.
- Использовать кросс-валидацию для оценки модели.
- Устранять ошибки в алгоритмах классификации с помощью кривых обучения и валидации.
- Выявлять проблемы смещения и дисперсии при помощи кривых обучения.
- Работать с переобучением и недообучением, используя кривые валидации.
- Настраивать модель машинного обучения и оптимизировать гиперпараметры с помощью поиска по решётке.
- Читать и правильно интерпретировать матрицу ошибок.
- Строить и правильно толковать ROC-кривую.

Сочетание разных моделей в ансамбле методов

- Использовать ансамбль методов с различными классификаторами.
- Комбинировать разные алгоритмы классификации.
- Знать, как оценить и настроить ансамбль моделей классификации.

Уровень 3. От мидла к синьору

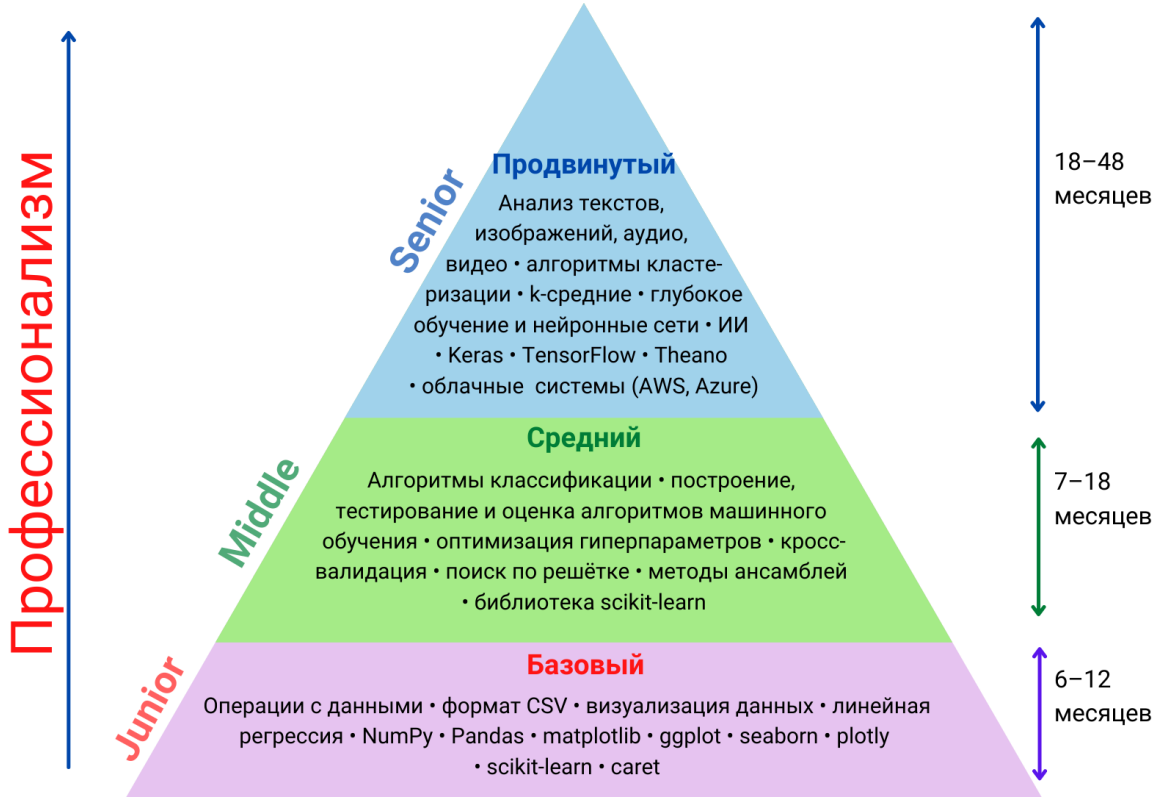
На этом уровне дата-сайентист углубляется в конкретную специализацию — и разбег по требованиям может быть очень большим. Однако каждому благородному дону, то есть синьору, точно придётся работать со сложными датасетами: текстом, изображениями, аудио (голос) и видео. Поэтому к навыкам среднего уровня добавится вот что:

- алгоритм кластеризации (обучение без учителя);
- k-средние;
- глубокое обучение;
- нейронные сети;
- библиотеки Keras, TensorFlow, Theano;
- основы разработки в облачных сервисах: AWS, Azure.

> **Дополнительно:** Здесь не повредит понимание различий между искусственным интеллектом, машинным обучением и глубоким обучением. У нас как раз есть статья на эту тему.

Дорожная карта развития навыков Data Science

Итак, чтобы стать специалистом базового уровня, понадобится от 6 до 12 месяцев. Вырасти с базового уровня до среднего можно за 7–18 месяцев. Продвинутый уровень потребует ещё от 18 до 48 месяцев.



Конечно, это приблизительные сроки. Многое зависит от бэкграунда: тем, кто неплохо прокачан в физике, математике, естественных и компьютерных науках, работал инженером или финансистом, будет гораздо проще. Но в первую очередь важны усилия и время, которые вы вкладываете в изучение Data Science, — в общем, никакой магии. Просто берём и делаем.

> На курсе «Профессии Data Scientist» мы даём не только базовые знания, но и часть навыков из среднего и продвинутого уровней. В итоге у вас появятся портфолио проектов, стаж не менее года, заряженные единомышленники и компетентные наставники. Приходите!