Data Science для решения задач информационной безопасности

Вводная лекция. Цели и задачи курса

Павел Владимирович Слипенчук

3 сентября 2019

10 сентября 2019

Москва, МГТУ им.Бауманка,
каф.ИУ-8, КИБ

План лекции

1. Цели. Требования на входе. Результат
2. Темы спецкурса
3. Как попасть на спецкурс?
4. Вопросы к собеседованию
5. Рекомендуемая литература

Цели. Требования на входе. Результат.

Цели студента

- Основы Data Science – это часть «джентльменского минимума» образованного профессионала в сфере IT/ИБ
- Понимание границ задач ИБ, для которых методы машинного обучения применимы и возможны
- Применение на практике знаний по математике
- Опыт построения программных систем
- Рекомендация на практику/трудоустройство в сфере анализа данных / Data Science / Data Engineering
- **Нормальные** темы для НИРС / дипломов

Цели педагога

- Рекрутинг
- Развитие речи
- Формирование «ML & DS культурной среды безопасников»
- Самообразование[1]
- снять лапшу с ушей! :)
- It is fun!

[1] Всегда попадаются 2-3 умных студента, которые задают интересные и полезные для педагога и его саморазвития вопросы.

Требования на входе

1. Linux
2. Студент должен иметь уверенный навык программирования (в предпочтении Python).
3. Знание Python или желание быстро и самостоятельно освоить этот язык программирования
4. Знание комбинаторики, теор.вера, мат.стата, дискретной математики
5. Алгоритмы и структуры данных

Результат на выходе

1. Целостное понимание что такое Machine Learning, Data Science, Data Engineering, Feature Extraction. Задачи ИБ, которые эффективно решаются данными методиками.
2. Научитесь работать в Jupyter
3. Базовые знания, умения и навыки в сфере Data Science, достаточные для самостоятельного вхождения в профессию (задан правильный «скелет» – вектор дальнейшего развития)
4. Лапша с ушей будет снята!
5. Снятие страха и загадочности у UEBA подхода. UEBA – это просто!
6. Понимание бизнес-процессов на стыке Data Science и ИБ.

Темы спецкурса

1. Экспертные системы, машинное обучение, Data Science, Data Engineering.
2. Проблемы ИБ решаемые экспертными системами.
3. Проблемы ИБ НЕ решаемые экспертными системами ;)
4. Джентльменский минимум по ML: гиперплоскость, регрессия, функция штрафа, полнота и точность, VaR, SSI, ROC-кривая и AUC, понятие ансамбля, бутстрепинг, бустинг, бэггинг.
5. Random Forest как наиболее простой и эффективный алгоритм. Его преимущества для «задач с противником».
6. Feature Extraction в ИБ задачах

7 UEBA: keyboard dynamics

8 UEBA: predilactions

9 UEBA: mouse track analysis

10 UEBA: алгоритм детектирования социальной инженерии

11 RSA[2]

12 Brand Protection и Anti Piracy.

13 Бизнес-процессы в DS

14 Как НЕ работают нейронные сети :)

[2] Имеется в виду система фрод-мониторинга RSA, а не асимметричный алгоритм шифрования :)

Как попасть на спецкурс?

Собеседования

03 и 10 сентября, во вторник, в 10:15 в НОЦ ИБ, класс 6 будет проведено собеседование.

По результатам будут отобраны ребята на спецкурс.

Замечание

Собеседование – это не экзамен. Вы такие, какие есть. Имеет смысл освежить в памяти ряд вопросов (см. «Вопросы к собеседованию»).

Спецкурс читается каждый год. Не пройдёте сейчас – пройдёте в следующем году

Пунктуальность

3 сентября для 5-х и 6-х курсов и 1-х, 2-х курсов магистратуры.

10 сентября для 2-4 курсов и для студентов не из МГТУ им.Баумана.

!

Запись проходит с 10:15 до 10:30 в НОЦ ИБ, класс 6. После 10:30 заявки не принимаются.

До 12:00 собеседования тет-а-тет.

На кого рассчитан спецкурс

На студентов 3-4 курсов, уже «набившие руку» в программировании, имеющие базовые представления в алгоритмах и структур данных, знающие комбинаторику, теор.вер., матстат. на инженерном уровне.

Студенты 1-го курса не допускаются. Студентам 2-го курса рекомендуется подождать (особо желающих берём)

Студентам 5-го и 6-го курса (1й, 2й курсы магистратуры) рекомендуется посещение только если спецкурс связан с работой. Иначе у вас не будет времени на нормальное усвоение материала.

Студенты не из МГТУ им.Баумана

Если вы не учитесь в МГТУ им.Баумана, вам нужен пропуск.

Для этого напишите в телеграмм **Александре**: @solinenarany

Приходите на собеседование 10 сентября. **Не опаздывайте**

При успешно пройденном собеседовании вам выпишут его на семестр.

Вопросы к собеседованию

Комбинаторика

1. Биномиальный коэффициент. Основные комбинаторные схемы: сочетание, размещение, перестановка, размещение с повторением, сочетание с повторением.
2. Формула включения/исключения
3. Подстановка
4. «Урновые схемы»
5. Отображения. Подсчёт количества отображений

Теория вероятности

1. Вероятность как статистическая величина и как априорная величина
2. Закон сложения и закон умножения
3. Условные вероятности. Формула Байеса
4. Ошибка первого и второго рода
5. Связь теории вероятности и комбинаторики

Математическая статистика

1. Математическое ожидание, медиана, мода
2. Момент случайной величины
3. Квантиль, перцентиль
4. Нормальное распределение
5. Корреляция
6. ЦПТ
7. Доверительный интервал
8. Метод Монте-Карло
9. Связь математической статистики и теории вероятности

Алгоритмы и структуры данных

1. массив, очередь, список
2. хеш-таблица, словарь
3. дерево.
4. дерево поиска. Примеры.
5. алгоритм Дейкстры, алгоритм Краскала
6. P и NP задачи. NP-complete задачи. Проблема $P \stackrel{?}{=} NP$.
7. Задача коммивояжёра. Алгоритм Литтла.
8. Решение задачи математическое и решение для нужд бизнеса.

Рекомендуемая литература

Python

Книги

- Для начинающих: **Марк Саммерфилд. Python на практике**
- Для продвинутых: **Лучано Рамальо. Python: к вершинам мастерства.**

Пишите много кода и придумывайте интересные задачи. Например это:

- «Русскоязычный чат-бот Boltoon: создаем виртуального собеседника» `https://habr.com/ru/post/340190/`
- «Хэш-стеганография с использованием vkapi» `https://habr.com/ru/post/351370/`

Комбинаторика

Методичка Жуковых.

В канале по хештегу библиотека есть.

Теория вероятности и математическая статистика.

Курс будет *прикладным*, поэтому для его понимания нужно хорошо знать основные определения и их смысл.

Есть много хороших книг, например «Феллер В. Введение в теорию вероятностей и ее приложения».

Если вы захотите стать профессиональным Data Scientist-ом, то, конечно теор.вер и мат.стат нужно будет выучить «на зубок».

Однако для подготовки к спецкурсу – это излишне.

Смотри вопросы из секции «Вопросы к собеседованию».