ML for Security – лаба 1

Возможные варианты заданий

1. Базовый контест (делают все хоть как-то)
   1.1. Обучить модель без каких-либо дополнительных условий, которая должна наилучшим образом отработать на тестовой выборке преподавателя с точки зрения F-меры
2. Вторичные контесты
   2.1. Максимально логичным образом обучить качественную модель, использующую для работы только 10 признаков из всего исходного множества.
   2.2. Обучить модель, обеспечивающую вероятность пропуска бота на уровне не выше 0.03, и имеющую насколько возможно низкую вероятность ложного обнаружения.
3. Углубленное аналитическое исследование по имеющимся данным
   3.1. Тем или иным способом выполнить сравнительное исследование значимости различных признаков применительно к произвольному фиксированному классификатору.
   3.2. Синтезировать 3 или более собственных признаков на основе имеющихся и показать, что они имеют какие-либо преимущества перед хотя бы какими либо из базовых признаков.
   3.3. Выбрать один базовый тип ML-модели на свой вкус (SVM, дерево решений, случайный лес, градиентный бустинг и пр.) и провести ROC-анализ в зависимости от её гиперпараметров.
   3.4. Провести исследование влияние параметров обучения на недо- и переобученность модели.

New! Важные примечания

- Выполнить нужно обязательно п. 1.1.
- Дополнительно можно выполнить любые задания из оставшихся. Больше трёх заданий при сдаче позднее начала ноября сдавать бессмысленно.
- Все модели, полученные в пп. 1.1, 2.1, 2.2 должны быть сохранены в виде файлов и отправлены вместе с файлами кода.

Формат функции классификации

```
def classify(model_file_name: str, data_file_name: str) -> np.array:

# model_file_name – имя файла, из которого будут загружены параметры обученной ML-модели (например, ‘model.txt’)

# data_file_name – имя CSV-файла с данными для анализа, которые нужно классифицировать. Файл имеет ту же структуру, что и файл обучающей выборки. Каждая строка файла (кроме заголовочной) содержит признаки одного пользователя.

# return – numpy-вектор с результатами классификации размерности (K, ), где K - количество содержательных строк в dataFileName
```

Описание набора признаков

Признаки, взятые непосредственно из профиля пользователя

Название признака	Описание
statuses_count	Количество твитов пользователя
followers_count	Количество подписчиков пользователя
friends_count	Количество подписок пользователя
favourites_count	Количество твитов в разделе понравившихся

listed_count	Количество твитов в разделе закрепленных твитов
is_default_profile	Использует ли профиль изображение по умолчанию
is_verified	Верифицирован ли данный аккаунт
is_profile_use_background_image	Использует ли аккаунт фоновое изображение для своей страницы

Подсчитанные признаки

Название признака	Описание
user_age	Время существования аккаунта в днях
tweet_freq	Отношения числа твиттов к времени существования аккаунта
followers_growth_rate	Отношение числа подписчиков к времени существования аккаунта
friends_growth_rate	Отношение числа подписок к времени существования аккаунта
favourites_growth_rate	Отношение числа понравившихся твиттов к времени существования аккаунта
listed_growth_rate	Отношение числа твиттов из раздела закрепленных к времени существования аккаунта
followers_friends_ratio	Отношение числа подписок к числу подписчиков
screen_name_length	Длина ника пользователя
num_digits_in_screen_name	Количество цифр в нике пользователя
length_of_name	Длина имени пользователя
num_digits_in_name	Количество цифр в имени пользователя
description_length	Длина описания аккаунта пользователя

Теоретические основы лабораторной работы

Боты в социальных сетях

Будем называть социальной сетью некий электронный ресурс, который предоставляет для пользователей следующие возможности:

1) возможность создавать учётную запись на этом ресурсе (аккаунт);
2) вносить на этот ресурс свои личные данные;
3) взаимодействовать и группироваться по интересам с другими людьми;
4) генерировать некую информацию и делиться ею с другими пользователями;
5) потреблять информацию, которую сгенерировали другие пользователи.

С ростом числа пользователей и информации, которую они генерируют каждый день, появилась потребность в автоматизации многих действий. Таким образом, появились аккаунты, которые управляются не реальными людьми, а программами. Такие аккаунты стали называться ботами.

Боты, как правило, используют API (Application Program Interface), который предоставляет социальная сеть, и могут совершать некий набор заранее запрограммированных действий:

1) публикация контента;
2) переписка с другими пользователями;
3) просмотр контента, который был создан другими аккаунтами;
4) написание комментариев;
5) нажатия кнопок «мне нравится» и подобных.

Следует отметить, что не всегда бот может быть создан с враждебными для других пользователей целями. Боты могут использоваться для автоматизации многих рутинных действий. Существуют боты, которые осуществляют рассылку полезной информации, агрегируют данные с разных источников и в удобном виде демонстрируют информацию пользователям социальной сети и т.д.

Но в том числе боты используется злоумышленниками. В таком случае по цели создания и характеру выполняемых действий ботов можно разделить на несколько групп:

1) боевые боты;
2) дезинформаторы;
3) спамеры;
4) технические боты;
5) тролли.

Боевыми ботами можно назвать ботов, которые устраивают фишинговые рассылки, выгружают данные пользователей, совершают действия, из-за которых страница реального человека может быть заблокирована и т.д.

Дезинформаторы характеризуются тем, что они больше всех остальных пытаются имитировать поведение реальных пользователей. Первоочередная задача таких ботов – распространение дезинформации.

Спамеры занимаются рассылкой бесполезной информации, которая просто засоряет информационный фон и мешает пользователям получать нужную информацию.

Под техническими ботами понимаются боты, которые используются для выполнения большого количества однообразных действий. Например, для написания простых комментариев и накрутки просмотров, лайков или репостов. Делается это с целью создать видимость активности для других пользователей или страниц в социальной сети, так как в этом случае они будут восприниматься с бóльшим доверием. Например, можно создать бота - дезинформатора и с помощью технических ботов сильно повысить вероятность того, что люди поверят в дезинформацию, которую он распространяет.

Под троллями понимаются боты, которые пишут оскорбительные комментарии, разжигают дискуссию, повышают уровень агрессии, поддерживают некий выбранный, как правило, негативный информационный фон. Такие боты с одной стороны могут быть использованы для усиления влияния ботов-дезинформаторов, а с другой стороны, их можно использовать в тандеме с техническими ботами, чтобы активность на данной странице повышалась также за счёт реальных пользователей, привлечённых ботами-троллями.

Также необходимо отметить, что чаще всего боты используются большими группами, что повышает эффективность решения поставленных задач. Такие группы ботов называют ботнетами. Для координации действий ботнетов обычно используется один реальный аккаунт, который управляется злоумешленником. Такой аккаунт называют ботмастером. Ботмастер управляет ботнетом путём пересылки команд через канал управления.

Общей чертой всех ботов является следующая особенность: все они в той или иной степени пытаются имитировать действия реальных пользователей социальной сети.

Постановка задачи обнаружения ботов; признаки ботов

С точки зрения машинного обучения задача определения ботов в социальных сетях – это задача классификации на 2 класса, где один из классов – это боты, а другой – реальные пользователи.

Для большинства социальных сетей все признаки можно разделить на статические и поведенческие. Поведенческие характеристики связаны с анализом действий исследуемого аккаунта:

- участие в искусственном продвижении материалов;
- скорость комментирования;
- комментарии разных аккаунтов с одного IP за короткий промежуток времени;
- содержание комментариев и т.д.

Под статическими характеристиками признаками понимают особенности оформления аккаунтов, такие как:

- корректность написания имени;
- наличие публикаций;
- наличие и содержание фотографий;
- дата создания аккаунта;
- число подписок;
- число подписчиков и т.д.