ЛЕКЦИЯ 2

Анализ моделей машинного обучения для выявления киберугроз

КЛАССИФИКАЦИЯ

Алгоритмы машинного обучения:

- Наивный байесовский классификатор (Naïve Bayes classifier);
- Логистическая регрессия (Logistic regression);
- Машина опорных векторов (Support vector machine);
- Метод k-ближайших соседей (k-nearest neighbors);
- Дерево решений (Decision tree);
- Случайный лес (Random forest);
- XGBoost.

НАИВНЫЙ БАЙЕС

- Наивный Байес – один из самых простых и часто применяемых алгоритмов машинного обучения для классификации текстов, использующий вероятностный подход, основанный на теореме Байеса с сильными предположениями о независимости данных. Наивный Байес рассматривает каждый признак независимо от других признаков и оценивает вероятность влияния каждого из них на итоговый результат. В контексте классификации текстов он обучается на документах каждого класса и вычисляет условную вероятность того, что документ d относится к классу c

$$P(c \mid d) = \frac{P(c) \times P(d \mid c)}{P(d)}$$

где $d = \{x_1, x_2, \ldots, x_n\}$, x_i – вес i^{th} слова в документе d, и c – класс документа.

МАШИНА ОПОРНЫХ ВЕКТОРОВ

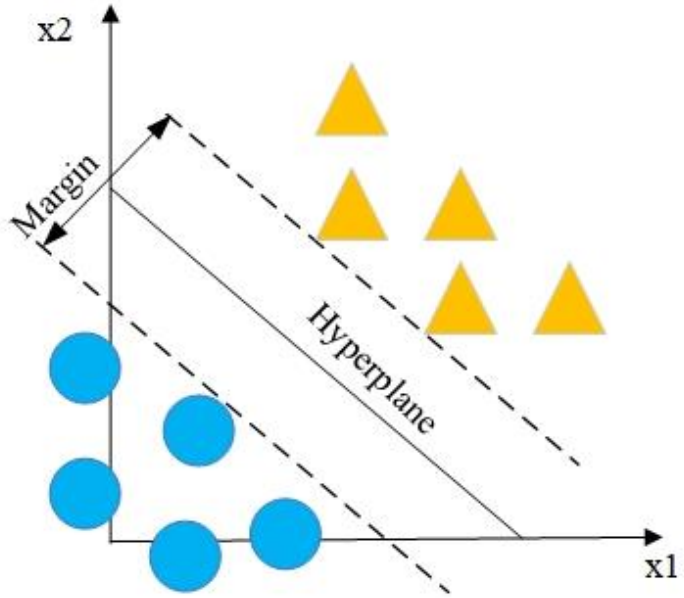
- Машина опорных векторов (Support vector machine) – еще один популярный алгоритм машинного обучения. Алгоритм использует пространство признаков, разделяемое гиперплоскостью, расположенной на максимальном расстоянии от ближайших точек двух классов обучающих данных. Чем шире граница, тем меньше ошибка классификатора, и достигается более эффективное разделение данных.
- Уравнение гиперплоскости записывается в следующем виде:

$$y_i(\vec{w} \times \vec{x} + b) \geq 0$$

где $\vec{x} = (x_1, x_2, ..., x_n)$ – вектор признаков; $\vec{w} = (w_1, w_2, ..., w_n)$ – вектор весов; y_i – выходное значение; b – сдвиг. Если значение больше или равно нулю, оно принадлежит к положительному классу. В противном случае относится к отрицательному классу.

МАШИНА ОПОРНЫХ ВЕКТОРОВ

- Разделяющая гиперплоскость Машины опорных векторов применяется преимущественно для двухклассовой классификации. Тем не менее, она без проблем адаптируется и для многоклассовой классификации с использованием метода «один против всех».

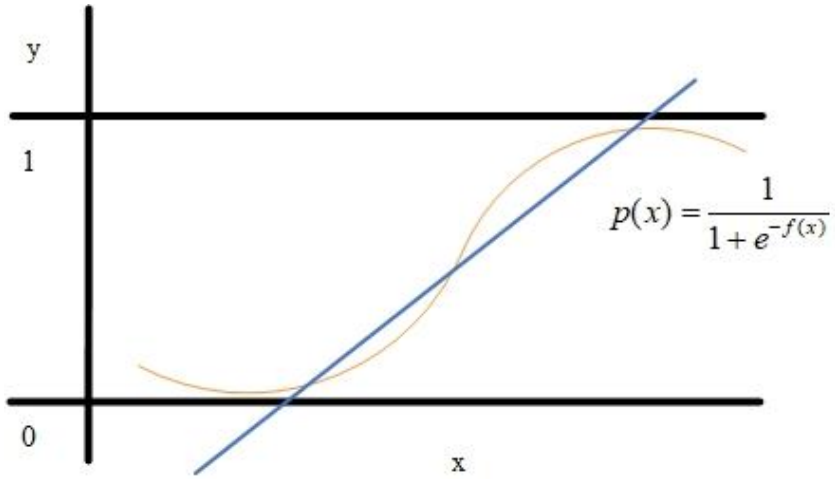


ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

- Логистическая регрессия предсказывает результат с использованием логистической функции

$$p(x) = \frac{1}{1 + e^{-f(x)}}$$

где $f(x) = w_0 + w_1 x_1 + \ldots + w_r x_r$ – линейная функция классификатора, $\vec{x} = (x_1, x_2, \ldots, x_n)$ – вектор признаков; $\vec{w} = (w_1, w_2, \ldots, w_n)$ – вектор весов. Логистическая функция $p(x)$ имеет вид сигмоида (рисунок 1.5) со значениями вероятности от 0 до 1. Документ d принадлежит к первому классу, если значение $p(x)$ близко к нулю. В противном случае его помещают во второй класс.



- В случае многоклассовой классификации используется подход «один против одного» (OvO), чтобы идентифицировать конкретный класс. В этом подходе датасеты с несколькими классами разбивается на несколько задач двоичной классификации, где каждый двоичный классификатор обучается на экземплярах, принадлежащих одному классу, и экземплярах, принадлежащих другому классу. Также используется метод «один против всех» (OvA), где множество двоичных классификаторов обучаются отличать экземпляры одного класса от всех других экземпляров. Преимущество OvO перед OvA заключается в том, что датасеты всех отдельных классификаторов сбалансированы, когда сбалансирован весь мультиклассовый датасет.

МЕТОД K-БЛИЖАЙШИХ СОСЕДЕЙ

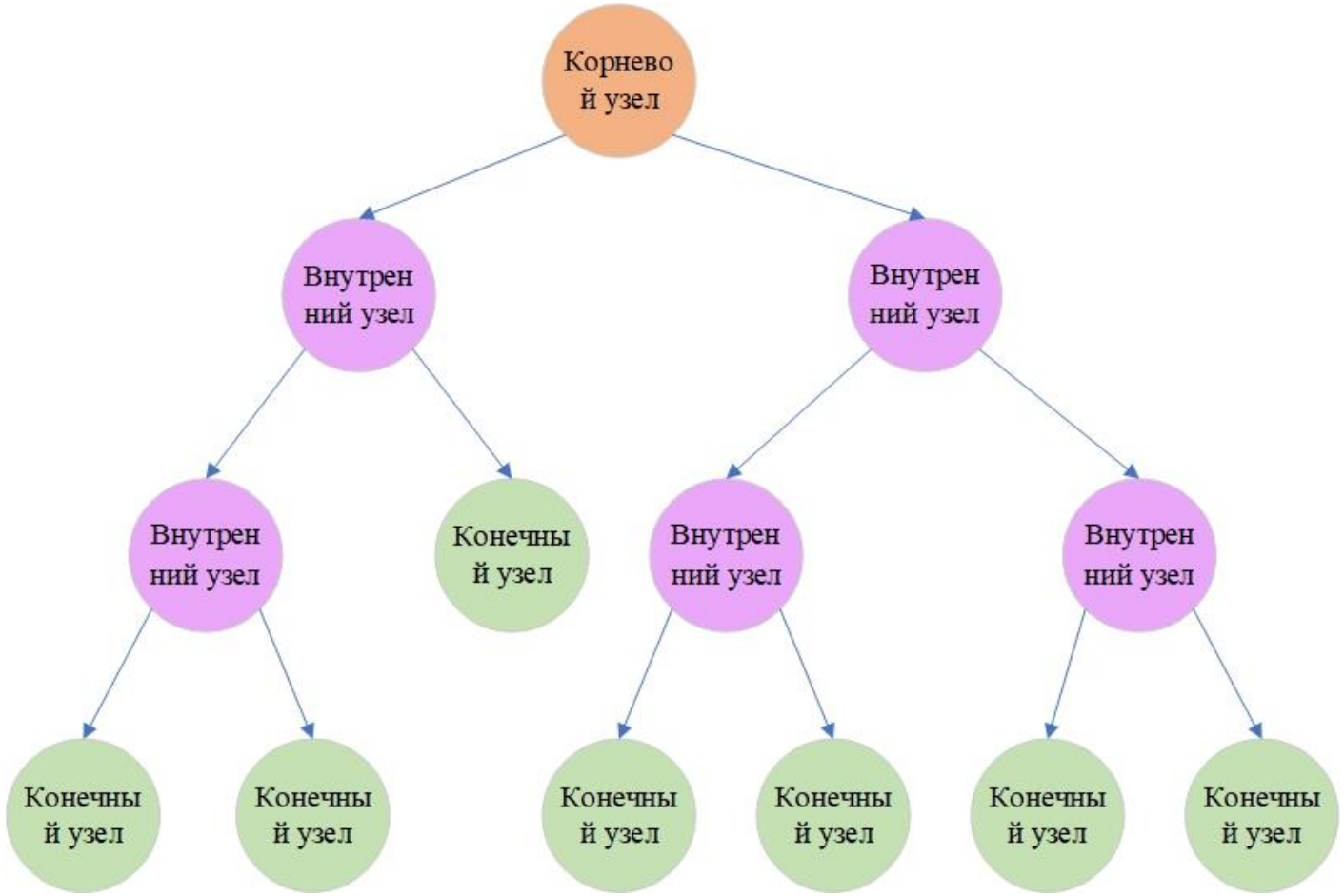
- Метод k-ближайших соседей – один из самых простых алгоритмов классификации данных. Он вычисляет расстояния между векторами и присваивает точки классу своих k ближайших соседних точек. В данном алгоритме вычисляется расстояние каждого объекта Для каждого объекта из тестовой выборки до всех объектов из обучающей выборки в пространстве признаков. Этот алгоритм обычно классифицирует документы с помощью наиболее широко используемой меры расстояния, называемой евклидовым расстоянием, которая определяется как $d(x,y)=\sqrt{\sum_{i=1}^{N}(a_{ix}-a_{iy})^2}$

где $d(x,y)$ расстояние между двумя документами; a_{ix} и a_{iy} веса i терма в документе x и y, соответственно; N номер уникального слова в списке документов. Метод k-ближайших соседей в процессе обучения запоминает векторы признаков и их метки классов. Там, где метки классов неизвестны, определяется расстояние между новым наблюдением и ранее запомненными векторами. Затем выбирается k ближайших векторов, и новый объект относится к классу, к которому принадлежит большинство. Выбор значения параметра k неоднозначен и требует экспериментальных подходов. При его увеличении улучшается точность классификации, но границы между классами становятся менее четкими. Данный метод показывает хорошие результаты классификации, однако основным его недостатком является высокая вычислительная трудоемкость при увеличении размера обучающей выборки.

ДЕРЕВО РЕШЕНИЙ

Дерево решений– метод обучения с учителем, который использует набор правил для принятия решений подобно тому, как человек принимает решения. В данном методе данные разделяются на подмножества в зависимости от определенных признаков, отвечая на определенные вопросы до тех пор, пока все точки данных не будут принадлежать определенному классу. Таким образом, образуется древовидная структура с добавлением узла для каждого вопроса. Первый узел является корневым узлом (root node). При классификации документов на первом этапе выбирается слово, и все документы, содержащие его, помещаются в одну сторону, а документы, не содержащие его, помещаются в другую сторону. В результате образуются два датасета. После этого в этих датасетах выбирается новое слово, и все предыдущие шаги повторяются. Так продолжается до тех пор, пока весь датасет не будет разделен и присвоен конечным узлам. Если в конечном узле все точки данных однозначно соответствуют одному и тому же классу, то класс узла точно определен. В случае смешанных узлов алгоритм присваивает данному узлу класс с наибольшим числом точек данных, относящихся к нему.

ДЕРЕВО РЕШЕНИЙ

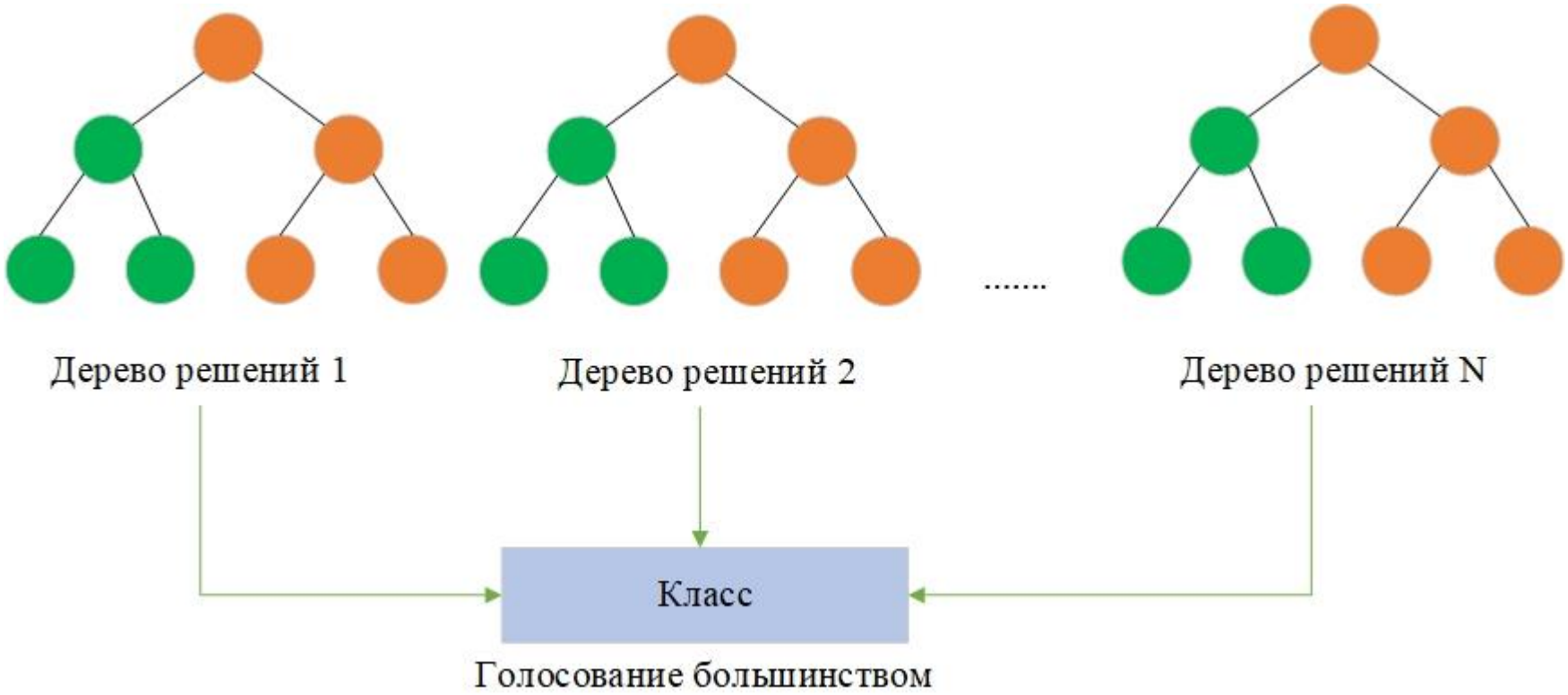


СЛУЧАЙНЫЙ ЛЕС

- Случайный лес– популярный алгоритм машинного обучения, основанный на концепции ансамблевого обучения. В данной концепции несколько классификаторов объединяются для улучшения производительности модели. Случайный лес состоит не из одного, а из множества деревьев решений. В задачах классификации каждый документ независимо кдассифицируется всеми деревьями. Класс документа определяется на основе наибольшего числа голосов среди всех деревьев.
- Алгоритм случайного леса имеет следующий ряд особенностей и преимуществ:

– Довольно быстро обучается.
– Эффективно обрабатывает датасеты с большим числом признаков.
– Выполняет предсказание данных с очень высокой точностью.
– Показывает хорошую эффективность даже при наличии большого числа пропусков данных.
– Хорошо обрабатываются как непрерывные, так и дискретные признаки.
– Обладает высокой масштабируемостью.

СЛУЧАЙНЫЙ ЛЕС



XGBOOST

- XGboost (eXtreme Gradient Boosting) – оптимизированный продвинутый алгоритм машинного обучения, использующий принцип бустинга. Он имеет хорошую производительность и решает большинство проблем регрессии и классификации. Использование ансамблевой техники подразумевает, что ошибки предыдущих шагов устраняются в новой модели. Отклонения прогнозов обученного ансамбля вычисляются на обучающем наборе на каждой итерации. Таким образом, оптимизация выполняется путем добавления новых древовидных прогнозов в ансамбль, уменьшая среднее отклонение модели. Эта процедура продолжается до тех пор, пока не будет достигнут требуемый уровень ошибки или критерий ранней остановки (максимальное количество деревьев или достижение заданной точности).

XGBoost

СПАСИБО ЗА ВНИМАНИЕ!!!