**Лабораторная работа № 9**

**Тема**: Оценка обученности моделей по кривым обучения.

**Цель работы** *–* научитьсяприменять для классификации алгоритм RandomForest и анализировать кривые обучения.

1. Для датасета из 5ой работы у вас уже есть 3 обученные модели классификаторов: дерево решений, метод ближайших соседей и логистическая регрессия.
2. Добавьте к существующим моделям классификатор RandomForest
3. Постройте для всех моделей кривые обучения.
4. Проанализируйте степень обученности моделей по графикам кривых обучения.

**Вопросы:**

1. Что означает термин «ансамбли» в машинном обучении.

Ансамблем называется алгоритм, который состоит из нескольких алгоритмов машинного обучения, а процесс построения ансамбля называется ансамблированием.

Ансамбль алгоритмов используется с целью получения лучшей эффективности прогнозирования, чем можно было бы получить от каждого обучающего алгоритма по отдельности.

1. Назовите наиболее известные модели построения ансамблей (асамблирования).

**стекинг, бэггинг, бустинг**

1. Расскажите принцип работы алгоритма RandomForest.

Для понимания прочитайте пример, будет проще

Предположим, вы решили узнать у своих знакомых об их опыте путешествий. Вы, вероятно, получите рекомендации от каждого друга и составите из них список возможных локаций. Затем вы попросите своих знакомых проголосовать, то есть выбрать лучший вариант для поездки из составленного вами перечня. Место, набравшее наибольшее количество голосов, станет вашим окончательным выбором для путешествия.

Вышеупомянутый процесс принятия решения состоит из двух частей.

Первая заключается в опросе друзей об их индивидуальном опыте и получении рекомендации на основе тех мест, которые посетил конкретный друг. В этой части используется алгоритм дерева решений. Каждый участник выбирает только один вариант среди знакомых ему локаций.

Второй частью является процедура голосования для определения лучшего места, проведенная после сбора всех рекомендаций. Голосование означает выбор наиболее оптимального места из предоставленных на основе опыта ваших друзей. Весь этот процесс (первая и вторая части) от сбора рекомендаций до голосования за наиболее подходящий вариант представляет собой алгоритм случайного леса.

Технически Random forest — это метод (основанный на подходе «разделяй и властвуй»), использующий ансамбль деревьев решений, созданных на случайно разделенном датасете. Набор таких деревьев-классификаторов образует лес. Каждое отдельное дерево решений генерируется с использованием метрик отбора показателей, таких как критерий прироста информации, отношение прироста и индекс Джини для каждого признака.

Любое такое дерево создается на основе независимой случайной выборки. В задаче классификации каждое дерево голосует, и в качестве окончательного результата выбирается самый популярный класс. В случае регрессии конечным результатом считается среднее значение всех выходных данных ансамбля. Метод случайного леса является более простым и эффективным по сравнению с другими алгоритмами нелинейной классификации.

Как работает случайный лес?

Алгоритм состоит из четырех этапов:

Создайте случайные выборки из заданного набора данных.

Для каждой выборки постройте дерево решений и получите результат предсказания, используя данное дерево.

Проведите голосование за каждый полученный прогноз.

Выберите предсказание с наибольшим количеством голосов в качестве окончательного результата.

Классификация объектов проводится путём голосования: каждое дерево комитета относит классифицируемый объект к одному из классов, а побеждает класс, за который проголосовало наибольшее число деревьев.

1. Что можно сказать о модели по ее кривым обучения?

Графики из лабы, их смысле прописан в 6 пункте README.docx, фактически эффективность обучения в зависимости от количеств тестовых данных, время на обучение.