# **6. Алгоритмы моделей ML**

Рассмотрим популярные алгоритмы моделей ML:

*Дерево решений.* Древовидный граф используется для поддержания принятий решений, который является достаточно простым и лёгким для понимания. В процессе построения такого дерева происходит рекурсивное деление множества на подмножества, используя правила решений в узлах.

*Байесовская классификация*. Этот метод берёт своё начало с теоремы Байеса. Обычно наивный байесовский классификатор причисляют к семейству простых вероятностных классификаторов.

*Метод наименьших квадратов.* Одним из вариантов реализации линейной регрессии является метод наименьших квадратов, который в отличии от линейной функции сводит к минимуму погрешности, создавая метрики ошибок.

*Метод опорных векторов.* Алгоритм строит гиперплоскость N-1-мерное таким образом, чтобы все объекты были разделены по двум группам.

*Логистическая регрессия.* Этот метод находит зависимость между категориальной зависимой переменной и одной или несколькими независимыми.

*Алгоритмы кластеризации.* Самые популярные алгоритмы, для задач кластеризации: на основе центра тяжести треугольника, на базе подключения, сокращение размерности, нейронные сети и вероятностные.

*Метод ансамблей.* Метод основывается на алгоритмах, которые генерируют классификаторы и разделяют объекты усредняя их или посредством итогов голосования.

Алгоритмы метода: бустинг, бэггинг, корректирование ошибок выходного кодирования. Метод ансамблей в отличии от моделей прогнозирования является лучшим инструментом.

Следующие описанные ниже методы используются непосредственно при анализе данных дипломной работы:

1. Линейная регрессия

Эта модель ML, которая была перенята из статистического анализа и предназначена выявлять зависимости между одним переменным и другим или несколькими. Линейная регрессия относится к моделям обучения с учителем, так как можно увидеть взаимосвязь между входным значением и результатом некоторого определенного числа примеров.

Для того чтобы на основе модели линейной регрессии сделать прогнозы можно построить график и определить функцию – функцию гипотезы. Точность показателей функции может быть охарактеризована функцией потерь или функцией стоимости. Эта функция считает разницу между обучающей выборкой и функцией гипотезы при определенных значениях x.

Обучение модели с помощью линейной регрессии – это, другими словами, минимизация функции потерь.

Для того чтобы усовершенствовать модели и избежать их усложнения, необходимо использовать регуляризацию данных.

Существует несколько типов регуляризации, основные из которых L1, L2, ElasticNet.

L1 регуляризация или регуляризация – Lasso основывается на подсчете абсолютного значения коэффициентов модели и добавлении этого значения в качестве штрафа, таким образом, к функции потерь прибавляется значение, отвечающее за штраф сложной модели.

(2)

Первое слагаемое является значением ошибки модели по метрике, а второе – штрафом, значение – сила регуляризации, чем больше это значение, тем сильнее регуляризация модели.

L2 регуляризация или регуляризация – Ridge также основана на добавлении штрафа, но в этом случае штраф рассчитывается как сумма квадратов весов модели, умноженных на гиперпараметр регуляризации. L1 регуляризация обнуляет большие значения весов, а Ridge регуляризация только приближает эти значения к нулю.

(3)

Как и в регуляризации – Lasso, первое слагаемое является значением ошибки модели по метрике, а второе – штрафом, значение – сила регуляризации, чем больше это значение, тем сильнее регуляризация модели. L2 регуляризация штрафует значения более гладко – это позволяет эффективнее использовать такой тип регуляризации в моделях, нуждающихся в более гладких производных.

Регуляризация ElasticNet основана на регуляризации L1 и L2. В этом типе регуляризации по мимо параметра используется и параметр , которое показывает, насколько комбинированной должна быть регуляризация Lasso и Ridge. Причем при параметре используется чисто Lasso, а при – чисто Ridge.

1. Экстремальный градиентный бустинг

Бустинг – такой метод ML, который уменьшает количество ошибок путём обучения нескольких моделей, что повышает точность всей системы в целом. Слабые модели, которые при изменении мельчайших деталей не способны работать корректно, с помощью бустинга преобразуются в одну сильную модель обучения.

Существует несколько видов бустинга, основными из которых являются адаптивный бустинг, градиентный бустинг и экстремальный градиентный бустинг.

Градиентный бустинг улучшает модель, сразу генерируя точные результаты, и не акцентирует внимание на исправлении ошибок. Экстремальный градиентный бустинг, эффективно работающий с большими объемами данных, всеми способами совершенствует градиентный бустинг, оптимизируя скорость вычислений и масштаб модели, что осуществляется путем распараллеливания, распределений вычислений и внешних вычислений.

Бустинг достаточно эффективен и прост в реализации, но главным недостатком метода является его чувствительность к выбросам данных.