**Лекция №5. Методы машинного обучения с учителем. Классификация.**

**Существует два основных алгоритма машинного обучения с учителем: регрессия и классификация. С регрессией мы уже познакомились на прошлом занятии.**

***Классификация*** – это задача прогнозирования категориальных переменных на основе данных. Задача классификации заключается в назначении объекта одному из заранее заданных классов на основе его характеристик.



Рисунок 1 – Классификация

Цель классификации состоит в том, чтобы спрогнозировать метку класса, которая представляет собой выбор из заранее определенного списка возможных вариантов.

Целевое значение категориальное – метка дискретных категорий или классов. Например, в задаче классификации изображений целевая переменная может указывать на категорию объекта на изображении (например, "кошка", "собака", "автомобиль").

Классификация делится на бинарную классификацию, которая является частным случаем разделения на два класса, и мультиклассовую классификацию, когда в классификации участвует более двух классов. Бинарную классификацию можно представить как попытку ответить на поставленный вопрос в формате «да/нет». Классификация электронных писем на спам и не спам является примером бинарной классификации. В данной задаче бинарной классификации ответ «да/нет» дается на вопрос «является ли это электронное письмо спамом?»

В бинарной классификации мы часто говорим о том, что один класс является положительным классом, а другой класс является отрицательным классом. При этом «положительный» означает здесь не наличие выгоды (ценности), а объект исследования. Таким образом, при поиске спама, положительным классом может быть класс «спам». Вопрос о том, какой из этих двух классов будет положительным, часто субъективен и зависит от предметной области исследования.

С другой стороны, определение вида растения на основе его размеров и формы листьев является примером мультиклассовой классификации. Еще один пример – прогнозирование языка веб-сайта. Классами здесь будет заранее определенный список возможных языков.

Выделяют также классификацию текста, изображений, видеозаписей, звуков и тд.

Классификация текста – это определение категории или темы текста. Например, классификация новостных статей по темам (политика, спорт, наука), определение тональности текста (позитивная, негативная, нейтральная) (рисунок 2).

Классификация изображений – это распознавание объектов на изображениях. Примеры включают классификацию животных на фотографиях, автомобилей по их маркам и моделям, распознавание лиц и др (рисунки 3, 4).



Рисунок 2 – Классификация текста

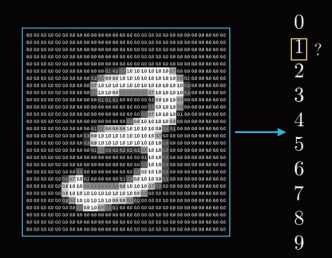


Рисунок 3 – Классификация изображений цифр

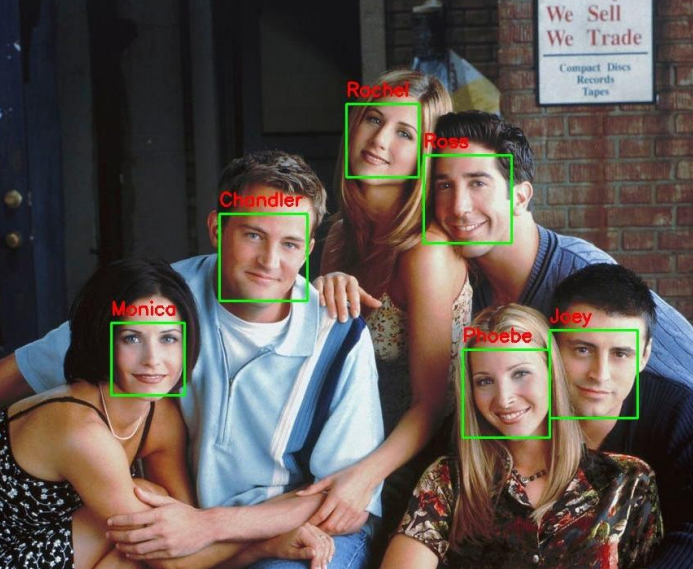


Рисунок 4 – Классификация изображений людей

Самый простой способ отличить классификацию от регрессии – спросить, заложена ли в полученном ответе определенная непрерывность (преемственность). Если полученные результаты непрерывно связаны друг с другом, то решаемая задача является задачей регрессии. Возьмем прогнозирование годового дохода. Здесь ясно видна непрерывность ответа. Разница между годовым доходом в 40000$ или 40001$ не существенна, хотя речь идет о разных денежных суммах. Если наш алгоритм предсказывает 39999$ или 40001$, в то время как он должен предсказать 40000$ (реальное значение годового дохода), мы не будем настаивать на том, что разница существенна. Наоборот, в задаче распознавании языка веб-сайта (задаче классификации) ответы четко определены. Контент сайта может быть написан либо на одном конкретном языке, либо на другом. Между языками нет непрерывной связи, не существует языка, находящегося между английским и французским.

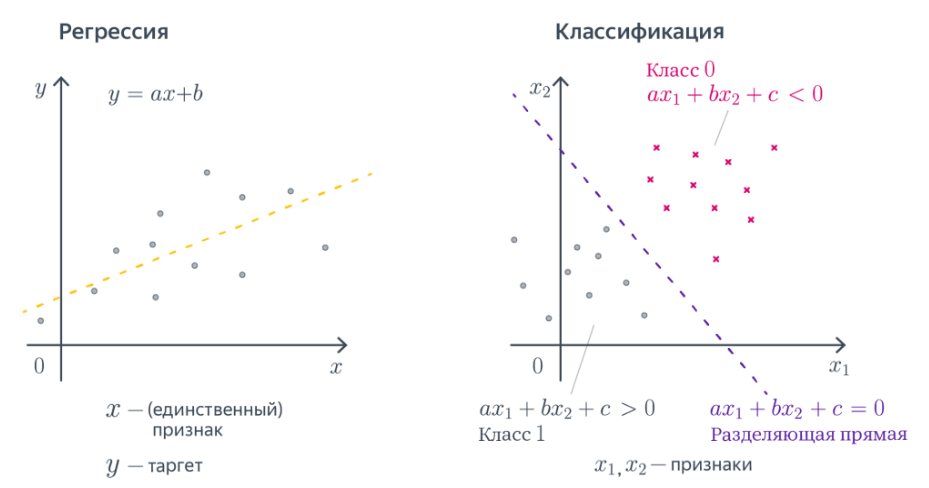


Рисунок 5 – Сравнение регрессии и классификации

Важными методами классификации являются логистическая регрессия, метод k-ближайших соседей, деревья решений, случайный лес и нейронные сети.

**Метод k-ближайших соседей (K-Nearest Neighbors)**

Метод k-ближайших соседей (KNN, где k – одно целое число, выбираемое заранее) – метод, который работает с помощью поиска кратчайшей дистанции между тестируемым объектом и ближайшими к нему классифицированным объектами из обучающего набора. Классифицируемый объект будет относится к тому классу, к которому принадлежит ближайший объект набора.

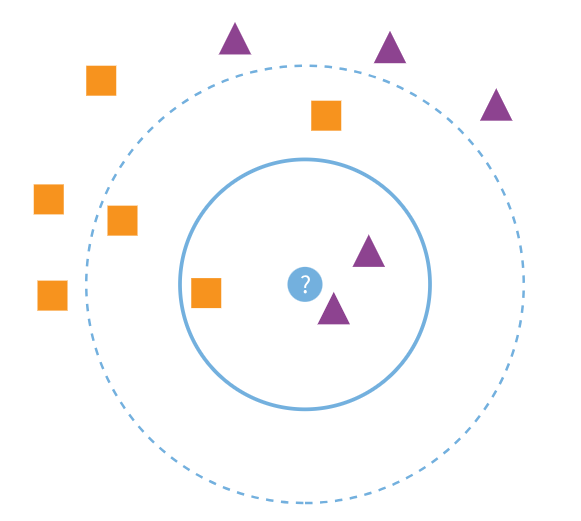


Рисунок 6 – Метод k-ближайших соседей

Метод k-ближайших соседей, возможно, является самым простым методом машинного обучения.

Метод был впервые разработан Эвелином Фиксом и Джозефом Лоусоном Ходжесом в 1951 году, и позднее развит Томасом Ковером.

Например, у нас есть набор данных о фруктах: вес, цвет, форма и т. д. Каждый фрукт относится к одному из трех классов: «яблоко», «груша», «апельсин». Алгоритм получает задание классифицировать новый фрукт. Он использует информацию из исходного набора и определяет класс нового фрукта путем нахождения k ближайших фруктов в пространстве признаков. Например, k = 5, тогда по признакам подбираются пять максимально похожих фруктов из исходного набора данных. Если большинство из них «яблоки», тогда и новый относится к классу «яблоко».

***Алгоритм:***

1. Вычислить расстояние до каждого из объектов обучающей выборки.

2. Отобрать k объектов обучающей выборки, расстояние до которых минимально

3. Класс классифицируемого объекта – это класс, наиболее часто встречающийся среди k ближайших соседей.

***Достоинства:***

1. Простота реализации.

2. Классификацию, проведенную алгоритмом, легко интерпретировать путем предъявления пользователю нескольких ближайших объектов.

***Недостатки:***

1. Необходимость хранения обучающей выборки целиком.

2. Поиск ближайшего соседа предполагает сравнение классифицируемого объекта со всеми объектами выборки.

***Когда следует использовать KNN?***

Предположим, вы хотели снять квартиру и недавно узнали, что соседка вашего друга может сдать свою квартиру в аренду через 2 недели. Поскольку квартира еще не размещена на сайте аренды, как вы могли бы попытаться оценить ее арендную стоимость?

Допустим, ваш друг платит 1200 долларов за аренду. Ваша арендная плата может быть примерно такой же, но квартиры не совсем одинаковые (площадь, качество мебели и т.д.), поэтому было бы неплохо иметь больше данных о других квартирах.

Расспросив других соседей и просмотрев квартиры в том же здании, которые были указаны на веб-сайте аренды, можно узнать, что арендная плата за четыре ближайшие соседние квартиры составляет 1200, 1210, 1210 и 1215 долларов. Эти квартиры находятся в том же квартале и на том же этаже, что и квартира вашего друга.

Арендная плата за другие квартиры, расположенные дальше, на том же этаже, но в другом блоке, составляет 1400, 1430, 1500 и 1470 долларов. Кажется, что они стоят дороже из-за того, что вечером больше солнечного света.

Учитывая близость квартиры, предполагаемая арендная плата составит около 1210 долларов. Таково общее представление о том, что делает метод k-ближайших соседей (KNN).

**Логистическая регрессия (Logistic Regression)**

Логистическая регрессия – это один из наиболее распространенных методов классификации, который используется для решения задачи бинарной (0, 1) классификации, когда нужно разделить данные на два класса. Она получила свое название благодаря тому, что использует логистическую функцию для прогнозирования вероятности принадлежности объекта к одному из классов. Логистическая регрессия является линейным классификатором и поэтому используется, когда в данных прослеживается какая-то линейная зависимость.

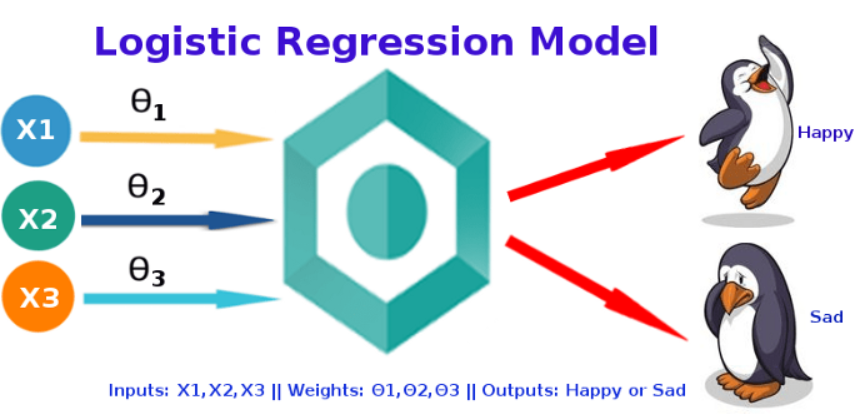


Рисунок 7 – Модель логистической регрессии

Если значение чего-либо равно либо больше 0.5, то объект классифицируется в большую сторону (к единице). Если значение меньше 0.5 – в меньшую (к нулю).

Логистическая регрессия использует линейную комбинацию входных признаков и соответствующих весов, которая описывает линейную гиперплоскость в пространстве признаков. Затем этот результат проходит через логистическую функцию, которая переводит линейную комбинацию в вероятность принадлежности объекта к одному из классов.

Вот ***несколько примеров***, когда мы можем использовать логистическую регрессию:

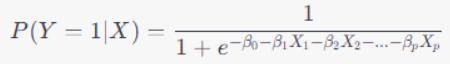
1. Мы хотим использовать кредитный рейтинг и банковский баланс, чтобы предсказать, не выполнит ли данный клиент дефолт по кредиту. (Переменная ответа = «По умолчанию» или «Нет по умолчанию»)

2. Мы хотим использовать среднее количество подборов за игру и среднее количество очков за игру, чтобы предсказать, будет ли данный баскетболист выбран в НБА (переменная ответа = «Выбран» или «Не выбран»).

3. Мы хотим использовать квадратные метры и количество ванных комнат, чтобы предсказать, будет ли дом в определенном городе продаваться по цене 200 тысяч долларов или выше (Переменная ответа = «Да» или «Нет»).

4. Обратите внимание, что переменная ответа в каждом из этих примеров может принимать только одно из двух значений. Сравните это с линейной регрессией, в которой переменная отклика принимает некоторое непрерывное значение.

По своей сути логистическая регрессия просто берет уравнение линейной регрессии и использует его как параметр сигмовидной функции. Математически это выражается следующим образом:

,

где:

Y – бинарный выходной результат (0 или 1);

X – вектор признаков, используемый для прогнозирования $Y$;

P(Y=1|X) – вероятность того, что $Y$ равно 1 при заданном $X$;

beta\_0, beta\_1, beta\_2, …, beta\_p – коэффициенты модели, которые нужно определить в ходе обучения, чтобы достичь наилучшего соответствия данных;

e – число Эйлера.

В целом, логистическая регрессия – это мощный инструмент для решения задач бинарной и многоклассовой классификации в Python. Она проста в использовании и предоставляет множество метрик для оценки качества работы модели.

***Разница между логистической регрессией и линейной регрессией***

В логистической регрессии по сравнению с линейной регрессией логистическая регрессия оценивает вероятность события, происходящего на основе независимых переменных, а линейная регрессия предсказывает значение зависимой переменной на основе независимой переменной.

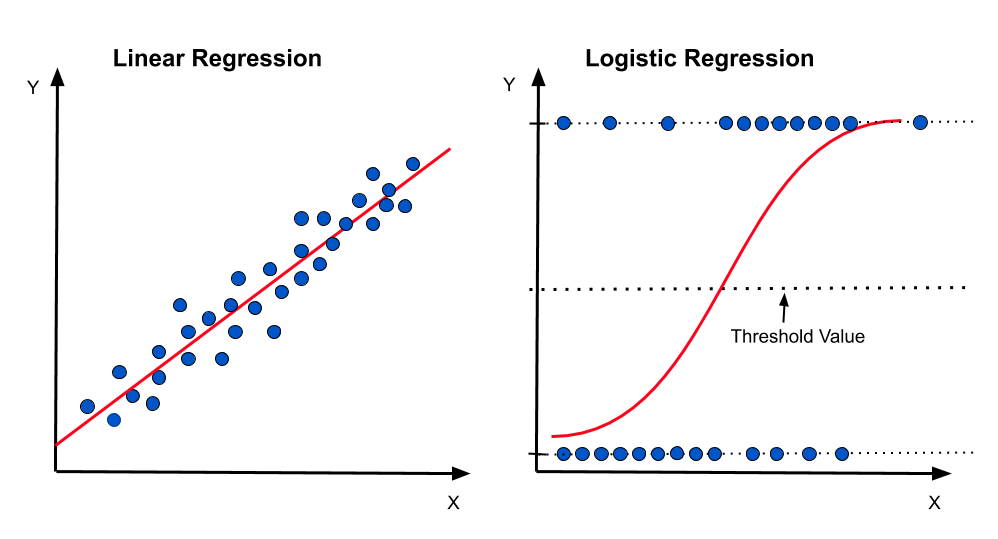


Рисунок 8 – Сравнение линейной и логистической регрессии

Поскольку логистическая регрессия оценивает вероятность, на выходе будет число от 0 до 1; зависимая переменная находится в двоичной форме. В случае линейной регрессии зависимая переменная (переменная отклика) является непрерывной.