**Лекция №4. Методы машинного обучения с учителем. Регрессия.**

Обучение с учителем (supervised learning) является одним из основных подходов в области машинного обучения. Этот подход основан на использовании размеченных данных, где каждый пример данных имеет соответствующую метку или правильный ответ. В процессе обучения модель пытается научиться предсказывать правильные ответы на основе предоставленных данных и меток. Машинное обучение с учителем часто требует вмешательства человека, чтобы получить обучающий набор данных, но потом оно автоматизирует и часто ускоряет решение трудоемких или неосуществимых задач. Обучение с учителем широко применяется в различных областях, таких как распознавание образов, классификация текстов, прогнозирование временных рядов и многое другое.



Рисунок 1 – Обучение с учителем

Рассмотрим основные принципы, обобщающую способность, переобучение и недообучение, преимущества и ограничения, а также алгоритмы применения обучения с учителем.

**Принципы обучения с учителем**

***1. Наличие учителя***

Основной принцип обучения с учителем заключается в наличии учителя, который предоставляет модели правильные ответы или метки для обучающих данных. Учитель может быть человеком, который размечает данные вручную, или может быть автоматическим алгоритмом, который предоставляет метки на основе заранее известных правильных ответов.

***2. Обучающие данные***

Для обучения с учителем необходимо иметь обучающие данные, которые состоят из пар входных данных и соответствующих правильных ответов или меток. Эти данные используются для обучения модели и настройки ее параметров таким образом, чтобы она могла предсказывать правильные ответы для новых данных.

***3. Функция потерь***

Для оценки качества предсказаний модели используется функция потерь. Функция потерь сравнивает предсказанные значения модели с правильными ответами и вычисляет разницу между ними. Цель обучения с учителем – минимизировать эту разницу, чтобы модель могла предсказывать правильные ответы с наименьшей ошибкой.

***4. Обучение и обновление параметров***

В процессе обучения модели с учителем ее параметры постепенно обновляются таким образом, чтобы минимизировать функцию потерь. Это достигается путем применения оптимизационных алгоритмов, которые настраивают параметры модели на основе градиента функции потерь. Чем больше обучающих данных и чем более разнообразны они, тем лучше модель сможет обучиться и предсказывать правильные ответы для новых данных.

***5. Оценка и тестирование***

После завершения обучения модели необходимо оценить ее качество и способность предсказывать правильные ответы для новых данных. Для этого используются отложенные тестовые данные, которые модель не видела во время обучения. Метрики, такие как точность, полнота, F-мера и другие, используются для оценки качества модели и сравнения ее с другими моделями или алгоритмами.

**Обобщающая способность, переобучение и недообучение**

В машинном обучении с учителем нам нужно построить модель на обучающих данных, а затем получить точные прогнозы для новых, еще не встречавшихся нам данных, которые имеют те же самые характеристики, что и использованный нами обучающий набор. Если модель может выдавать точные прогнозы на ранее не встречавшихся данных, мы говорим, что модель обладает способностью обобщать результат на тестовые данные. Нам необходимо построить модель, которая будет обладать максимальной обобщающей способностью. Обычно мы строим модель таким образом, чтобы она давала точные прогнозы на обучающем наборе. Если обучающий и тестовый наборы имеют много общего между собой, можно ожидать, что модель будет точной и на тестовом наборе. Однако в некоторых случаях этого не происходит. Например, если мы строим очень сложные модели, необходимо помнить, что на обучающей выборке можно получить произвольную правильность.

Чем сложнее модель, тем лучше она будет работать на обучающих данных. Однако, если наша модель становится слишком сложной, мы начинаем уделять слишком много внимания каждой отдельной точке данных в нашем обучающем наборе, и эта модель не будет хорошо обобщать результат на новые данные. Существует оптимальная точка, которая позволяет получить наилучшую обобщающую способность. Собственно, это и есть модель, которую нам нужно найти. Компромисс между переобучением и недообучением показан на рисунке 2.



Рисунок 2 – Компромисс между сложностью модели и правильностью на обучающей и тестовой выборках

Важно отметить, что сложность модели тесно связана с изменчивостью входных данных, содержащихся в вашем обучающем наборе: чем больше разнообразие точек данных в вашем наборе, тем более сложную модель можно использовать, не беспокоясь о переобучении. Обычно больший объем данных дает большее разнообразие, таким образом, большие наборы данных позволяют строить более сложные модели. Однако простое дублирование одних и тех же точек данных или сбор очень похожих данных здесь не поможет.

**Преимущества обучения с учителем**

1. Ясность и точность. Обучение с учителем предоставляет точные метки или ответы для каждого обучающего примера. Это позволяет алгоритму ясно понимать, какие признаки связаны с определенными метками и какие связи между ними существуют.

2. Высокая точность предсказаний. Благодаря наличию точных меток обучающих данных, модель, обученная с учителем, может достичь высокой точности предсказаний на новых, ранее не виденных данных.

3. Возможность обучения сложных моделей. Обучение с учителем позволяет использовать сложные алгоритмы и модели, такие как нейронные сети, которые могут обрабатывать большие объемы данных и выявлять сложные зависимости между признаками и метками.

**Ограничения обучения с учителем**

1. Зависимость от качества обучающих данных. Качество и представительность обучающих данных имеет решающее значение для успешного обучения с учителем. Если данные содержат ошибки, шум или недостаточно разнообразны, модель может давать неправильные или недостаточно точные предсказания.

2. Затраты на разметку данных. Обучение с учителем требует наличия размеченных данных, что может быть трудоемким и затратным процессом. Разметка данных требует участия экспертов или специалистов, что может быть дорого и занимать много времени.

3. Недостаточная обобщающая способность. Модель, обученная с учителем, может быть хорошо настроена на обучающие данные, но не всегда обладать достаточной обобщающей способностью для предсказания новых данных. Это может привести к переобучению модели, когда она слишком точно подстраивается под обучающие данные и не может обобщить свои знания на новые ситуации.

**Существует два основных алгоритма машинного обучения с учителем: регрессия и классификация.**

***Регрессией*** называют задачу прогнозирования количественных переменных на основе данных. В основе регрессионного анализа лежит построение функциональной зависимости между одной или несколькими независимыми переменными и одной зависимой переменной.

Цель регрессии состоит в том, чтобы спрогнозировать непрерывное число или число с плавающей точкой, если использовать термины программирования, или вещественное число, если говорить языком математических терминов.

Целевое значение численное – конкретное числовое значение. Модель стремится предсказать точное численное значение. Примеры численных целевых переменных: цена товара, температура, доход, количество продаж и так далее.

Примеры: прогнозирование объема урожая зерна на ферме в зависимости от объема предыдущего урожая, погоды и количество сотрудников, работающих на ферме; предсказание стоимости дома на основе его площади, количества комнат и расположения; прогнозирование годового дохода человека в зависимости от его образования, возраста и места жительства.

Важным методом регрессионного анализа является линейная регрессия, где зависимая переменная представляется в виде линейной комбинации независимых переменных.

***Линейная регрессия*** – это алгоритм обучения с учителем, который используется для предсказания непрерывной зависимой переменной на основе одной или нескольких независимых переменных. Он строит линейную модель, которая наилучшим образом соответствует данным.



Рисунок 3 – Линейная регрессия

Линейная регрессия является одним из наиболее широко известных методов моделирования, так как он является одним из первых элитных методов регрессионного анализа, которые были использованы людьми во время обучения прогнозному моделированию.

Обратите внимание, что в множественной линейной регрессии используется более одной независимой переменной, а в простой линейной регрессии – только одна независимая переменная. Таким образом, линейную регрессию лучше всего использовать только тогда, когда существует линейная связь между независимой и зависимой переменной.

**Пример:** компания может использовать линейную регрессию для измерения эффективности маркетинговых кампаний, ценообразования и рекламных акций по продажам продукта. Предположим, компания, продающая спортивный инвентарь, хочет понять, принесли ли ей существенную отдачу средства, вложенные в маркетинг и продвижение своей продукции. Линейная регрессия – лучший статистический метод для интерпретации результатов. Лучшее в линейной регрессии то, что она также помогает анализировать неясное влияние каждой маркетинговой и брендинговой деятельности, но при этом контролировать потенциал составной части для регулирования продаж. Если компания проводит две и более рекламных кампании одновременно; как если бы один на телевидении и два на радио, то линейная регрессия может легко проанализировать независимое, а также комбинированное влияние этих рекламных объявлений вместе.

**Как работает линейная регрессия?**

Возьмем небольшой набор данных. Предположим, что это группа коттеджей, расположенных в одном районе. На оси Х обозначена их площадь, а на оси Y – рыночная стоимость. Чтобы увидеть, как стоимость дома зависит от его площади, построим регрессию (рисунок 4).



Рисунок 4 – Набор из 7 точек

Это будет простая линейная регрессия с одной переменной. Изменится площадь дома – изменится и стоимость. Для вычисления используем стандартное уравнение регрессии: f (x) = b + m⋅x, где m – это наклон линии, а b – ее сдвиг по оси Y. То есть изменение коэффициентов m и b будет влиять на расположение прямой:

– если изменить m – прямая наклонится сильнее влево или вправо;

– если изменить b – прямая сместится вверх или вниз по оси Y.

Провести прямую линию через все точки на графике не получится, если они расположены в хаотичном порядке. Поэтому с помощью линейной регрессии определяется оптимальный вариант расположения этой прямой (рисунок 5). Некоторые точки все равно останутся на расстоянии, но оно должно быть минимальным. Расчет этого минимального расстояния от прямой до каждой точки называется функцией потерь.



Рисунок 5 – Дистанция ошибки

Для оценки точности регрессии используют разные метрики, например MSE (от англ. *mean squared error* – средняя квадратическая ошибка). Чем ниже MSE, тем лучше модель.



Рисунок 6 – Несколько случайных кандидатов

В первом случае MSE будет равна 0,17, во втором – 0,08, а в третьем – 0,02. Получается, что третья прямая лучше всего показывает зависимость цены дома от его площади.

**Преимущества линейной регрессии**

1. Линейная регрессия работает исключительно хорошо для линейно разделимых данных.

2. Легче внедрять, интерпретировать и эффективнее обучать.

3. Он довольно хорошо справляется с переоснащением, используя методы размерного уменьшения, регуляризацию и перекрестную проверку.

4. Еще одним преимуществом является экстраполяция за пределы определенного набора данных.

**Недостатки линейной регрессии**

1. Предположение о линейности между зависимыми и независимыми переменными.

2. Он часто весьма склонен к шуму и переобучению.

3. Линейная регрессия весьма чувствительна к выбросам.

4. Склонен к мультиколлинеарности.