Конспект вводной лекции

Краткая история дисциплины:

В 1959 году Артур Самуэль (Arthur Samuel), исследователь в области искусственного интеллекта и изобретатель первой самообучающейся компьютерной программы игры в шашки, ввел в научный обиход термин «машинное обучение». Самуэль определил машинное обучение как процесс, в результате которого компьютеры способны показать поведение, которое в них не было явно запрограммировано.

Ниже представлены некоторые исторические факты, связанные с историей развития методов машинного обучения:

1936 AT&T Bell Labs создает синтезатор речи.

1946 Представлен общественности компьютер ЭНИАК — сверхсекретный проект армии США.

1950 Алан Тьюринг (Alan Turing) создает Тьюринг тест для оценки интеллекта компьютера.

1952 Артур Самуэль, пионер в области искусственного интеллекта создает первую шашечную программу для IBM 701. В 1955 году Самуэль добавляет в программу способность к самообучению.

1958 Фрэнк Розенблатт (Frank Rosenblatt) придумал Персептрон — первую искусственную нейронную сеть и создал первый нейрокомпьютер «Марк-1» . Издание New York Times назвало Персептрон: «эмбрионом электронного компьютера, который в будущем сможет ходить, говорить, видеть, писать, воспроизводить себя и осознавать свое существование».

1959 Марвин Минский (Marvin Minsky) стал одним из сооснователей лаборатории Массачусетского технологического института. Профессор Минский создал первую обучающуюся машину SNARC со случайно связанной нейросетью.

1963 Ларри Робертс (Larry Roberts) сформулировал тезисы компьютерного зрения в своей диссертации в MIT.

1967 Написан метрический алгоритм классификации (Метод k ближайших соседей). Алгоритм позволил компьютерам использовать простые шаблоны распознавания.

1979 В Стэнфордском университете Ханс Моравек (Hans Moravec) перестроил стэнфордскую тележку. Оснастил ее стереоскопическим зрением. Стэнфордская тележка — долгосрочный исследовательский проект, который проводился в университете в течении 20 лет с 1960 по 1980 гг.

1981 Gerald Dejong представляет концепцию, основанную на обучении (Explanation Based Learning).

1985 Терри Сейновски (Terry Sejnowski) создает NetTalk искусственную нейронную сеть.

1986 Дэвидом Румельхартом (David Rumelhart) и Робби Вильямсом был заново открыт и популяризирован алгоритм обратного распространения ошибки. Этот алгоритм также был получен другими учеными независимо друг от друга. Впервые он был предложен Полом Вербосом (Paul Werbos) в 1974 году.

1997 Компьютер Deep Blue обыграл чемпиона мира по шахматам Гарри Каспарова.

2006 Джеффри Хинтон (Geoffrey Hinton), ученый в области искусственных нейросетей, ввел в обиход термин «Глубинное обучение» (Deep learning).

2011 Эндрю Нг (Andrew Ng) и Джефф Дин (Jeff Dean) основали Google Brain.

2011 Суперкомпьютер IBM Watson, оснащенный системой искусственного интеллекта, одержал победу в телевикторине Jeopardy!. Его соперниками были маститые игроки Брэд Раттер (Bred Ratter) и Кен Дженнингс (Ken Jennings).

2012 В Google X Lab разработали алгоритм, позволяющий идентифицировать видиоролики, содержащие котов.

2012 Google запускает облачный сервис Google Prediction API для машинного обучения, помогающий анализировать неструктурированные данные.

2014 В Facebook изобрели программный алгоритм DeepFace для распознавания лиц. Точность алгоритма составила 97%.

2015 Amazon запустила собственную платформу машинного обучения — Amazon Machine Learning.

2015 Microsoft создает платформу Distributed Learning Machine Toolkit, который предназначена для децентрализованного машинного обучения.

2016 Программа AlphaGo, разработанная гугловской компанией DeepMind, выиграла в четырех играх из пяти у чемпиона мира по игре в го корейца Ли Седоля (Lee Se-dol)

Краткая теория

Машинное обучение – это математическая дисциплина, находящаяся на стыке прикладной статистики, теории информации, вычислительной математики, дискретного анализа и численных методов оптимизации.

К основным задачам данной дисциплины относятся: классификация, кластеризация, регрессия, уменьшение размерности и прогнозирование. Рассмотрим более детально каждый из описанных пунктов:

1. Классификация

Терминология:

Множество объектов – задаётся описанием признаков для каждого объекта.

Обучающая выборка – заданное конечное множество объектов, для которых задано их соответствие классам из множества классов.

Классифицировать объект – сопоставить объекту номер (или наименование) класса из заданного множества.

Результатом классификации объекта является номер или наименование класса, который будет получен в результате работы алгоритма.

Постановка задачи:

Имеется множество объектов, которые некоторым образом разделены на классы. Задано конечное множество объектов и множество классов. Для всех объектов из данного множества известно, к какому классу они относятся. Кроме того, существует множество объектов, для которых неизвестна классовая принадлежность. Требуется разработать алгоритм, позволяющий классифицировать неопределённые объекты и отнести их к каким-либо из заданных классов.

Типы классов:

- Непересекающиеся классы, когда объект может относиться только к одному классу.

- Пересекающиеся классы, когда объект может относиться одновременно к нескольким классам.
- Нечёткие классы, когда необходимо определить вероятность соответствия объекта каждому из классов.

Формальная постановка задачи:

Пусть X – множество описаний объектов, Y – конечное множество номеров (наименований) классов. Существует неизвестная искомая зависимость – отображение $y^*: X \to Y$, значения которой известный только для объектов конечной обучающей выборки $X^n = \{(x_1, y_1), \dots, (x_n, y_n)\}$. Требуется разработать алгоритм $a: X \to Y$, позволяющий классифицировать любой объект $x \in X$.

Пример задачи классификации:

Допустим, существует множество изображений, каждое из которых относится к одному из двух типов: на изображении присутствуют люди, на изображении нет людей. Необходимо для произвольного изображения определить, присутствует ли на нём человек.

Методы классификации: Байесовский классификатор, нейронная сеть, линейный разделитель, индукция правил, алгоритмическая композиция, сокращение размерности, выбор модели

2. Кластеризация

Кластерный анализ – задача разбиения заданного конечного множества объектов на непересекающиеся подмножества, называемые кластерами, таким образом, чтобы каждый кластер состоял из схожих объектов, и чтобы объекты из разных кластеров существенно отличались друг от друга.

Функция расстояния – некоторая функция, определяющая насколько «далеко» находятся друг от друга объекты. Очень легко понять, что это такое, если считать, что изучаемые объекты – это точки на плоскости и требуется разбить их на группы, в каждой из которых точки располагаются максимально близко друг к другу. В таком случае функцией расстояния будет являться евклидово расстояние между двумя точками на плоскости.

Обучающая выборка – конечное множество объектов, заданное либо при помощи признаков каждого объекта, которые позволяют определить расстояние между двумя объектами, либо при помощи матрицы расстояний, в которой для каждой пары объектов задано расстояние между ними.

Формальная постановка задачи:

Пусть X — множество объектов, Y — множество номеров (имён, меток) кластеров. Задана функция расстояния между объектами $p(x_i, x_j)$. Имеется конечная обучающая выборка объектов $X^n = \{x_1, \dots, x_n\} \subset X$. Требуется разбить выборку на непересекающиеся подмножества, называемые *кластерами*, так, чтобы каждый кластер состоял из объектов, близких по метрике p, а объекты разных кластеров существенно отличались. При этом каждому объекту $x_i \in X^n$ приписывается номер кластера y_i.

Алгоритм кластеризации — это функция $a: X \to Y$, которая любому объекту $x \in X$ ставит в соответствие номер кластера $y \in Y$. Множество Y может быть известно и задано заранее, тем не менее чаще всего ставится задача определить оптимальное число кластеров, с точки зрения того или иного *критерия качества* кластеризации.

Методы кластеризации: графовые алгоритмы, статистические алгоритмы, иерархическая кластеризация, нейронная сеть.

Основные цели кластеризации:

- Понимание данных за счёт выявления кластерной структуры. Разбиение входных данных на группы схожих объектов может помочь упростить дальнейшую обработку данных и принятие решения.
- Сжатие данных. Если исходная выборка достаточно велика, то существует возможность сократить её, оставив только наиболее типичных представителей каждого кластера.
- Выявление нового. Выделяются нетипичные объекты, которые невозможно отнести ни к одному из заданных кластеров.

Примером может служить разбиение большого количества неупорядоченных файлов (например, при восстановлении большого числа файлов на повреждённом жестком диске, для которых невозможно определить их расширение) по группам схожести для дальнейшего анализа.

3. Регрессия

Регрессионный анализ – это статистический метод исследования влияния одной или нескольких независимых переменных на зависимую переменную. Используется как метод моделирования данных и исследования их свойств.

Регрессионная модель – это функция от независимой переменной и параметров с добавленной случайной переменной. Параметры модели

настраиваются таким образом, чтобы как можно точнее приближать данные. Критерием качества модели, как правило, считается среднеквадратическая ошибка (сумма квадратов разности значений модели и значений зависимой переменной для всех значений независимой переменной в качестве аргумента).

Формальное определение:

Регрессия — зависимость математического ожидания случайной величины от одной или нескольких других случайных величин (свободных переменных), то есть $E(y|x) = f(x)$. Под регрессионным анализом подразумевается нахождение такой функции f, которая описывает эту зависимость. Регрессия может быть представлена в виде суммы неслучайной и случайной составляющих.

$$y = f(x) + \nu,$$

где f— функция регрессионной зависимости, а ν— аддитивная случайная величина с нулевым матожиданием. Предположение о характере распределения этой величины называется гипотезой порождения данных. Обычно предполагается, что величина ν имеет гауссово распределение с нулевым средним и дисперсией σ_ν^2.

4. Уменьшение размерности

Дано конечное множество объектов, каждый из которых описан конечным числом признаков. Необходимо сформулировать алгоритм, позволяющий уменьшить количество описывающих признаков, не изменив при этом структуру множества объектов.

Основные два требования к методам понижения размерности заключаются в том, что количество новых признаков, описывающих объекты исходного множества должно быть меньше, чем количество исходных признаков, и при этом новые признаки должны содержать как можно больше информации из исходных признаков.

В качестве примера подобной задачи можно рассмотреть множество точек в трёхмерном пространстве (каждая из них задаётся тремя координатами), которые находятся при этом в одной плоскости. Тогда можно спроецировать эти точки на плоскость, к которой они относятся, и избавиться от третьей координаты. При этом свойства, необходимые, например, для задачи кластеризации, будут сохранены, но размер исходных данных уменьшится на треть.

5. Прогнозирование

Прогнозирование временных рядов – это множество методов, позволяющих определить следующее значение x_{t+1} временного ряда $X = x_1 x_2 \dots x_t$, в котором все значения принадлежат некоторому конечному алфавиту A.

Формальное определение задачи:

Пусть известна последовательность событий происходивших последовательно во времени $X = x_1 x_2 \dots x_t, x_i \in A$, которую будем называть временной серией. Тогда функцию $F(X, a) = P(x_{t+1} = a|X), a \in A$, которая сопоставляет каждому символу из алфавита A вероятность того, что именно он появится следующим после последовательности X, будем называть прогнозной функцией. Кроме того, определим прогнозное значение $y \in A$, такое что $F(X, y) = \max_{a \in A} F(X, a)$.

Ошибкой прогнозирования будем называть отклонение прогнозного значения от фактического. Задачей прогнозирования является нахождение такой функции $F(X, a)$, которая минимизирует ошибку прогнозирования.

Задача прогнозирования является одной из самых известных и распространённых задач машинного обучения, люди, далёкие от изучения данной дисциплины ежедневно сталкиваются с результатами работы подобных методов (прогноз погоды). Помимо этого, методы прогнозирования могут использоваться для предсказания природных явлений (землетрясения, цунами, высокая солнечная активность), предсказания курсов валют и значений различных экономических показателей, для прогнозирования сбоев в вычислительных системах и сетях и т.д.