Задачи машинного обучения.

Ссылка на видеокурс от МИАН

Машинное обучение и искусственный интеллект

Разберемся с терминологией: **машинное обучение** и **искусственный интеллект**. Это практически одно и то же. **Искусственный интеллект** (ИИ, Artificial Intelligence, AI) – технологии (аппаратные или программные комплексы, в основном именно программные), которые могут решать (или пытаются решить) задачи (особенно нечеткие, плохо формализованные, обычно NP-полные) творческими, эвристическими методами. Под творческими понимают методы, которые копируют возможности людей. **Машинное обучение** (МО, Machine Learning, ML, МЛ) – технология построения программ, которая включает в себя обучение на примерах (обучение может быть как с учителем, так и без). Обучение с учителем дает возможность предоставить программе примеры с решениями и строится на этих примерах. В обучении без учителя программе предоставляется большое количество многомерной информации, которую она должна обработать. Вообще говоря, МО – более техничный, современный, прагматичный термин, нежели чем ИИ.

Важнейшими отличительными чертами алгоритмов МО и систем ИИ являются

- обучаемость (самообучаемость) — система способна работать с каждым разом все лучше, она развивается, так же как и человек
- обобщение — предполагается, что система способна обобщить поступающую ей информацию, т.е. применить полученные в процессе обучения знания на новых данных
- адаптация — способность работы в той среде, которая не была предопределена
- самоорганизация — при обобщении появляются новые структуры, которых изначально не было
- нечеткость и неточность — вообще говоря не предполагается, что система работает с четкими, конкретными данными, из этого следует неточность ответа
- компонента случайности — чтобы справиться с нечеткостью, при решении вносится некая рандомизация, из-за чего алгоритм МО не дает гарантии решить задачу.

Этим алгоритмы МО отличаются от стандартных алгоритмов, с которыми мы привыкли работать.

Задачи и методы машинного обучения

Рассмотрим прикладные задачи математики, которые решаются с помощью МО.

1. **Задача оптимизации.**

   Задача исходит из того, что имеется некоторое непустое произвольное конечное/бесконечное множество $U \neq \emptyset$ – это множество нашего выбора.

   Также имеется некоторая функция, которая задана на этом множестве:

   $$F : U \to \mathbb{R}.$$

   Мы допускаем, что и множество U, и функция F **не известны точно и четко не определены**. Множество может быть фиксированным, а может быть случайным. А функция может содержать ошибку, например, при подстановки одного и того же элемента мы будем получать разные числовые (на самом деле не обязательно числовые, но для простоты взяли числовые) значения.

Требуется найти такой элемент $\tilde{u} \in U$, что

$$F(\tilde{u}) = \min_{u \in U} F(u).$$

Аналогично для максимизации.

Корректна ли такая задача? Реализуем ли экстремум? Даже в самых простых случаях поставленная задача некорректна. Экстремум может не достигаться, функция может быть не ограничена снизу (или сверху) и т.д. В МО поставленная задача **не решается в точности**. Сама задача решается творческим методом, а формулировка является неким ориентиром, не обязательно достижимым.

2. **Задача классификации объектов.**

   Имеется некоторое непустое произвольное конечное/бесконечное множество $V \neq \varnothing$. Задано некоторое натуральное $K > 1$ – количество занумерованных классов.

   Требуется построить функцию
   $$\Phi \to \{1, 2, \ldots, K\}.$$

   Φ – функция классификации, которой на вход подается объект $v \in V$. Результатом является номер класса, к которому v принадлежит. Чаще всего в качестве результата выступает вектор $(p_1, p_2, \ldots, p_K)$, где $p_i \in [0, 1]$ – **вероятность** того, что объект v принадлежит классу i. Не обязательно сумма элементов этого вектора равняется 1. Выбираем класс, который имеет максимальную вероятность (таких классов может быть несколько). Наверное, именно к этому классу (классам) принадлежит v, если точность обучения высока.

Задача 1 и задача 2 сводимы друг к другу. Задача классификации сводится к задаче оптимизации, когда ищется максимальный элемент полученного вектора вероятностей. Задача оптимизации сводится к задаче классификации, когда множество U разбивается на 2 класса: элементы, на которых достигается экстремум, на которых не достигается экстремум. Так классическая нейронная сеть решает задачу 2, но мы можем считать, что решается задача 1. Верно и обратное.

Реальные задачи, которые решаются с помощью алгоритмов МО, конкретизируют две постановки.

Методы машинного обучения, которые будут рассмотрены:

1. **Методы оптимизации**
   (a) Метод имитации отжига;
   (b) Метод роя частиц;
   (c) Генетический алгоритм.
2. **Методы классификации**
   (a) Классические нейронные сети (многослойные НС, персептрон) — с обучающей выборкой;
   (b) Сети Кохонена, самоорганизующиеся карты Кохонена (SOM) — без обучающей выборки;
   (c) Сеть Хопфилда.
3. **Байесовские методы**
4. **Решающие деревья**

Машинное обучение с точки зрения стохастического анализа

Есть поток полезной информации, который будет подаваться в систему МО. Система усваивает входную информацию (обучается) оптимальным образом. На выходе получаем оптимальное решение. Если на вход системы подается новая информация, то полученное решение уточняется. Цикл обучения будет повторяться до тех пор, пока решение нас не устроит.

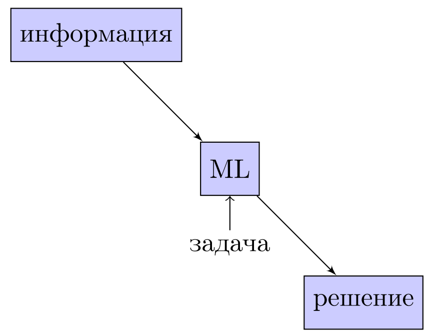

Схема обучения.

Особенность обучения в том, что на каждой его итерации подается **новая** информация. Если из раза в раз мы будем подавать одно и то же, то система обучаться не будет.

Нас интересует вопрос **полезности** информации. Вопрос о полезности/неполезности информации **субъективен**, так как он зависит от места, от времени и от субъекта, который воспринимает информацию. С точки зрения теории вероятностей, информация зависит от вероятности.

Советский ученый Александр Александрович Харкевич предложил меру оценки полезности информации. Имеется задача и все доступные средства для ее решения. Пусть $P \in [0, 1]$ – вероятность достигнуть поставленной цели. Добавляется некая новая информация I, которая меняет вероятность решить задачу P_I в зависимости от P. Полезность информации оценивается как $M = P_I - P$. Иными словами, мы определяем, насколько поступившая информация I изменила вероятность P. В этой постановке заложен субъект, который получает информацию – это конкретная задача. Для нас важно с помощью поступившей информации повысить вероятность её решения.

Формализуем понятие информации с точки зрения стохастического анализа.

Фиксируем вероятностное пространство $\langle \Omega, \mathcal{A}, \mathbb{P} \rangle$. Пусть $\mathcal{A}$ – σ-алгебра, которая описывает все возможные события, которые может выдать наша модель (“богатая” σ-алгебра). Рассматривается случайная величина $\xi(w)$, которая **измерима** относительно σ-алгебры $\mathcal{A}$. При таких условиях можно уверенно полагать, что мы можем располагать всей информацией, потому что ξ умеет разбирать все события из $\mathcal{A}$.

Теперь зафиксируем другое вероятностное пространство $\langle \Omega, \mathcal{A}_1, \mathbb{P} \rangle$, где $\mathcal{A}_1$ – “бедная” σ-алгебра, иными словами, $\mathcal{A}_1 \subset \mathcal{A}$. Для наглядности приведем небольшой пример.

Пример 1. Мы решили сыграть в “двадцать одно”. Тогда Ω — 36 карт. Пусть $\mathcal{A}$ — множество из всех подмножеств Ω, а $\xi(\omega)$ — случайная величина: “валет” – 2 очка, “дама” – 3 очка, “шестёрка” – 6 очков и т.д. Ясно, что ξ измерима относительно $\mathcal{A}$. Иными словами, мы делаем любые утверждения на основании любых событий, которые возможны при вытаскивании одной из 36 карт. Можем спокойно играть.

Теперь, пусть $\mathcal{A}_1 = \{\Omega, \emptyset, \text{Кр}, \text{Чр}\}$, т.е. на основании вытащенной карты мы можем делать вывод лишь о том, какого цвета масть. Естественно, $\mathcal{A}_1$ — σ-алгебра, $\mathcal{A}_1 \subset \mathcal{A}$, но $\xi(\omega)$ не измерима относительно $\mathcal{A}_1$. Мы не сможем полноценно решить, брать ли нам следующую карту, либо закрыть ход, имея в распоряжении только $\mathcal{A}_1$.

Далее поступила некая новая информация. Допустим, что теперь помимо цвета мы можем различать картинки. Понятно, что информация, которой мы располагаем, расширилась, и, очевидно, $\mathcal{A}_1 \subset \mathcal{A}_2 \subset \mathcal{A}$. В σ-алгебру $\mathcal{A}_1$ добавились новые события, но пока информации для игры в “двадцать одно” недостаточно, $\xi(\omega)$ также не измерима относительно $\mathcal{A}_2$. Однако мы можем постепенно расширять σ-алгебру по мере поступления свежей информации.

$\{\mathcal{A}\}_k = \mathcal{A}_0 \subset \mathcal{A}_1 \subset \mathcal{A}_2 \subset \dots \subset \mathcal{A}$ — фильтрация, или поток σ-алгебр. Вообще говоря, такая последовательность может быть бесконечной. Здесь $\mathcal{A}_0 = \{\Omega, \emptyset\}$ — самый “бедный” вариант σ-алгебры, мы не можем делать никаких самых простых предположений.

С ростом σ-алгебры растёт возможность субъекта воспринимать информацию, иначе говоря, распознавать события.