Статистические методы обучения.
Обучение с подкреплением

Статистические методы обучения

Основными понятиями при использовании статистических методов обучения продолжают оставаться данные и гипотезы, но данные рассматриваются как свидетельства, то есть конкретизации случайных переменных, описывающих проблемную область, а гипотезы представляют собой вероятностные теории того, как функционирует проблемная область.

Рассмотрим простой пример. На кондитерской фабрике выпускаются леденцы двух разновидностей – вишневые и лимонные, которые заворачиваются в одинаковые фантики и упаковываются в очень большие внешне неразличимые пакеты, относящиеся к следующим пяти типам:

h_1: 100% вишневых леденцов,

h_2: 75% вишневых + 25% лимонных леденцов,

h_3: 50% вишневых + 50% лимонных леденцов,

h_4: 25% вишневых + 75% лимонных леденцов,

h_5: 100% лимонных леденцов.

Работник ОТК должен определить, к какому типу относится предоставленный на контроль пакет, которому соответствует случайная переменная H, принимающая значение от h_1 до h_5. По мере развертывания конфет регистрируются данные о них $D_1, D_2, \ldots D_n$, где D_i – случайная переменная, принимающая значение из множества {cherry, lime}.

Работник ОТК должен предсказать к какой разновидности относится следующая выбираемая конфета.

В баесовском обучении исходя из полученных данных вычисляется вероятность каждой гипотезы и делается предсказание. Пусть переменная D представляет все данные с наблюдаемым значением d, тогда вероятность каждой гипотезы может быть определена с помощью правила Байеса:

$P(h_i|d) = \alpha\ P(d|h_i)\ P(h_i)$

Пусть необходимо сделать предсказание в отношении неизвестного количества X. В таком случае применяется следующее уравнение:

$$P(X|d) = \sum_i P(X \mid d, hi) P(hi \mid d) = \sum_i P(X \mid hi) P(hi \mid d)$$

где предполагается, что каждая гипотеза определяет распределение вероятностей по X. Это уравнение показывает, что предсказания представляют собой взвешенные средние по предсказаниям отдельных гипотез. Сами гипотезы, по сути, являются “посредниками” между фактическими данными и предсказаниями.

Основными количественными показателями в байесовском подходе являются распределение априорных вероятностей гипотезы $P(h_i)$ и правдоподобие данных согласно каждой гипотезе $P(d|h_i)$.

Предположим, что изготовитель объявил о наличии распределения априорных вероятностей по значениям $h_1, \ldots, h_5$, которое задано вектором {0.1, 0.2, 0.4, 0.2, 0.1). Правдоподобие данных рассчитывается в соответствии с предположением, что наблюдения являются независимыми и одинаково распределенными, поэтому соблюдается следующее уравнение:

$$P(d|h_i) = \prod_j P(dj \mid hi)$$

Например, если в действительности пакет содержит только лимонные леденцы (h_5) и все первые 10 конфет являются лимонными леденцами, то значение $P(d|h_3)$ равно 0.5^{10}, поскольку в пакете типа h_3 половина конфет – лимонные леденцы. Априори наиболее вероятным вариантом является гипотеза h_3 и остается таковой после развертывания 1 конфеты с лимонным леденцом. После развертывания 2 конфет с лимонными леденцами наиболее вероятной становится гипотеза h_4, а после обнаружения 3 или больше лимонных леденцов наиболее вероятной становится гипотеза h_5. байесовская вероятность того, что следующий леденец будет лимонным, согласно уравнению для $P(X|d)$ монотонно увеличивается до 1.

Данный пример показывает, что истинная гипотеза в конечном итоге будет доминировать над байесовским предсказанием.

При любом заданном распределении априорных вероятностей, которое не исключает с самого начала истинную гипотезу, апостериорная вероятность любой ложной гипотезы в конечном итоге полностью исчезает.

В реальных задачах обучения пространство гипотез обычно является очень большим или бесконечным, поэтому приходится вместо прямого вычисления суммы для $P(x|d)$ (или, в непрерывном случае, интегрирования) приходится прибегать к приближенным или упрощенным методам.

Упрощение может быть достигнуто путем предсказаний на основе единственной наиболее вероятной гипотезы, т.е. той гипотезы h_i, которая максимизирует значение $P(h_i|d)$. Такую гипотезу h_{map} называют максимально апостериорной. Предсказания, сделанные на основе такой гипотезы, являются приближенно байесовскими до такой степени, что $P(X|d) = P(X|h_{map})$. В примере $h_{map} = h5$ после обнаружения 3 лимонных леденцов подряд.

Упрощение может быть также достигнуто, например, путем принятия предположения о равномерном распределении априорных вероятностей по пространству гипотез. В этом случае обучение с помощью максимально апостериорной гипотезы сводится к выбору гипотезы h_i, которая максимизирует значение $P(d|h_i)$. Такая гипотеза **называется гипотезой с максимальным правдоподобием**. Это – приемлемый подход, применяемый в тех обстоятельствах, когда нет оснований априорно отдавать предпочтение одной гипотезе перед другой. Такой метод обучения становится хорошей аппроксимацией байесовского обучения и обучения с помощью максимально апостериорной гипотезы, когда набор данных имеет большие размеры, поскольку сами данные исправляют распределение априорных вероятностей по гипоте-

зам, но связан с возникновением определенных проблем при использовании небольших наборов данных.

Обучение с подкреплением

Задача обучения с подкреплением состоит в том, чтобы обеспечить использование наблюдаемых вознаграждений для определения в процессе обучения оптимальной стратегии для данной среды.

Предполагается, что среда является полностью наблюдаемой, поэтому информация о текущем состоянии поступает с результатами каждого восприятия. Обычно рассматривают три проекта агентов:

- Агент, действующий с учетом полезности, определяет с помощью обучения функцию полезности состояния и использует ее для выбора действий, которые максимизируют ожидаемую полезность результата;
- Агент, действующий по принципу Q-обучения, определяет с помощью обучения функцию “действие-значение”, или Q-функцию, получая сведения об ожидаемой полезности выполнения данного конкретного действия в данном конкретном состоянии.
- Рефлексный агент определяет с помощью обучения стратегию, которая непосредственно отображает состояния в действия.

Пассивное обучение с подкреплением

При таком виде обучения стратегия агента остается неизменной, а задача состоит в том, чтобы определить с помощью обучения полезности состояний (или пар “состояние-действие”). Для этого может также потребоваться определение с помощью обучения модели среды.

Задача пассивного обучения аналогична задаче оценки стратегии, которая является частью алгоритма итерации по стратегиям. Пассивный обучающийся агент не знает модели перехода $T(s,a,s')$, которая определяет вероятность достижения состояния s' из состояния s после выполнения действия

a; он также не знает функцию вознаграждения R(s), которая задает вознаграждение для каждого состояния.

Существует простой метод непосредственной оценки полезности, идея которого состоит в том, что полезностью данного конкретного состояния является ожидаемое суммарное вознаграждение, связанное с действиями, выполняемыми, начиная с этого состояния, а каждая попытка представляет собой выборку этого значения для каждого посещенного состояния.

Таким образом, в конце каждой последовательности алгоритм вычисляет наблюдаемое будущее вознаграждение для каждого состояния и обновляет соответствующим образом оценку полезности для этого состояния путем ведения текущего среднего значения для каждого состояния. В пределе, после выполнения бесконечного количества попыток, среднее по выборкам сходится к значению истинного ожидания.

Очевидно, что непосредственная оценка полезности представляет собой один из видов контролируемого обучения, в котором каждый пример задает состояние в качестве входных данных, а наблюдаемое будущее вознаграждение – в качестве выходных. Это означает, что данный метод позволяет свести обучение с подкреплением к стандартной задаче индуктивного обучения. Однако в этом методе не учитывается тот факт, что полезности состояний не являются независимыми. Дело в том, что полезность каждого состояния равна сумме его собственного вознаграждения и ожидаемой полезности его состояний-преемников. Данный метод можно рассматривать как поиск в пространстве гипотез, которое имеет размеры намного большие, чем необходимо, поскольку включает также много функций, которые нарушают уравнения Беллмана. По этой причине данный алгоритм часто сходится очень медленно.

Активное обучение с подкреплением

Пассивный обучающийся агент руководствуется постоянно заданной стратегией, которая определяет его поведение, а активный агент должен сам принимать решение о том, какие действия следует предпринять. Прежде все-

го, агенту потребуется определить с помощью обучения полную модель с вероятностями результатов для всех действий, а не просто модель для заданной стратегии. Затем необходимо принять в расчет тот факт, что агент должен осуществлять выбор из целого ряда действий. Полезности, которые ему потребуются для обучения, определяются оптимальной стратегией и подчиняются уравнениям Беллмана:

$$U(s) = R(s) + \gamma \max_s \sum_{S'} T(s,a,s') \; U(s')$$

U(s) – функция полезности состояния **s** (суммарное вознаграждение из состояния **s** с продолжением дальше);

R(s) – кратковременное вознаграждение за пребывание в состоянии **s**,

T(s,a,s') – вероятность достижения **s'** из состояния **s** после выполнения действия **a**.

Эти уравнения могут быть решены для получения функции полезности U с помощью алгоритмов итерации по значениям или итерации по стратегиям. Последняя задача состоит в определении того, что делать на каждом этапе. Получив функцию полезности U, оптимальную для модели, определяемой с помощью обучения, агент может извлечь информацию об оптимальном действии, составляя одношаговый прогноз для максимизации ожидаемой полезности. Еще один вариант состоит в том, что если используется итерация по стратегиям, то оптимальная стратегия уже известна, поэтому агент должен просто выполнить действие, рекомендуемое согласно оптимальной стратегии.

Однако может оказаться, что выбор оптимального действия приводит к неоптимальным результатам. Причиной этого может быть то, что модель, определяемая с помощью обучения, не является такой же, как истинная среда; поэтому то, что оптимально в модели, определяемой с помощью обучения, может оказаться неоптимальным в истинной среде. Поэтому агент должен искать компромисс между потреблением полученных результатов для максимизации своего вознаграждения (что отражается в его текущих оценках

полезностей) и исследованием среды для максимизации своего долговременного благосостояния.