Байесовское обучение.

Ссылка на видеокурс от МИАН

Формула Байеса

Фиксируем вероятностное пространство $\langle \Omega, \mathcal{A}, \mathbb{P} \rangle$.

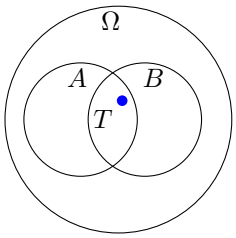
Пусть имеются события $A, B \in \mathcal{A}$. Их вероятность $\mathbb{P}(A) > 0, \mathbb{P}(B) > 0$.

Условная вероятность $\mathbb{P}(A|B)$ — вероятность события A при условии события B.

Определение 1. Условная вероятность A при условии B равняется частному от деления вероятности событий A и B на вероятность B:

$$\mathbb{P}(A|B) \stackrel{\text{def}}{=} \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{1}$$

Существует простая геометрическая интерпретация данного определения.



Положим $S(\Omega) = 1$. Тогда из определения геометрической вероятности $\mathbb{P}(B) = \frac{S(B)}{S(\Omega)} = S(B), \mathbb{P}(A \cap B) = \frac{S(A \cap B)}{S(\Omega)} = S(A \cap B)$. Известно, что точка T упала куда-то в область B. Зная это, мы определяем вероятность того, что мы попали также и в область B. Иными словами, находим долю пересечения $A \cap B$ от B:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{S(A \cap B)}{S(B)}.$$

Аналогично можем записать условную вероятность B при условии A:

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}. \tag{2}$$

Определение 2. Из (2) выражаем

$$\mathbb{P}(A \cap B) = \mathbb{P}(B|A) \cdot \mathbb{P}(A) \tag{3}$$

и подставляем в (1):

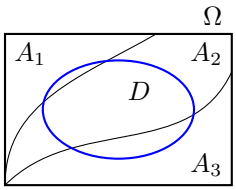
$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A) \cdot \mathbb{P}(A)}{\mathbb{P}(B)}.$$

Это формула называется **формулой Байеса**.

Формула Байеса широко применяется в приложениях. С помощью нее вычисление условной вероятности $P(A|B)$ можно свести к вычислению условной вероятности $P(B|A)$, что иногда бывает намного проще.

Формула полной вероятности

Для приложений необходимо обобщение формулы Байеса.



Разобьем множество Ω следующим образом:

$$\Omega = \bigcup_k^N A_k, \text{ где } A_k \in \mathcal{A},$$

т.е. множества A_k составляют разбиение множества всех элементарных исходов Ω. Причем предполагаем, что $A_k \cap A_m = \emptyset, k \neq m$ и $\mathbb{P}(A_k) > 0, k = 1, 2, \ldots, N$.

$\{A_k\}$ — полная группа непересекающихся (взаимоисключающих) событий, которые также называют **гипотезами**. Это те события, которые мы хотим спрогнозировать. В каждом опыте может быть реализовано только одно из этих событий, нас интересует, какое именно и с какой вероятностью. Например, сможет ли любимая команда победить в сегодняшенем футбольном матче, либо она проиграет — это два возможных исхода.

Нам известны вероятности каждого из таких событий $\mathbb{P}(A_k), k = 1, 2, \ldots, N$ — это **априорные оценки** для гипотез. Такие оценки рассчитываются из каких-то общих принципов, из статистики и т.д. К примеру, в предыдущие игры с сегодняшним противником наша команда чаще выигрывала, чем проигрывала, поэтому априорная вероятность проигрыша ниже.

Известно, что произошло некоторое событие $D \in A$, которое как-то может повлиять на наши оценки. Внутри этого события был реализован элементарный исход. D — это некая новая для нас информация, которая может поменять исход игры.

Теперь нас интересует условная вероятность той или иной гипотезы при условии события D. Для каждого фиксированного k необходимо вычислить $\mathbb{P}(A_k|D)$ — **апостериорную оценку**. Тогда мы сможем ответить на вопрос, как поступившая информация повлияла на вероятность наших исходов. К примеру, перед матчем мы узнали, что лучший игрок команды-противника травмирован, выступать он не сможет — эта информация увеличивает вероятность выигрыша команды, за которую мы болеем.

Теперь нас интересует вопрос, как вычислить вероятность события D. Графически можем видеть, что $D = D \cap (A_1 \cup A_2 \cup \cdots \cup A_N) = DA_1 \cup DA_2 \cup \cdots \cup DA_N$. Поскольку события A_k взаимоисключающие, то их пересечения с D также взаимоисключающие события. Исходя из аксиомы Колмогорова и формулы (3), получаем

$$\mathbb{P}(D) = \sum_k^N \mathbb{P}(D \cap A_k) = \sum_k^N \mathbb{P}(D|A_k) \cdot \mathbb{P}(A_k)$$

формулу полной вероятности.

Для каждой из гипотез формула Байеса принимает следующий вид:

$$\mathbb{P}(A_k|D) = \frac{\mathbb{P}(D|A_k) \cdot \mathbb{P}(A_k)}{\mathbb{P}(D)} = \frac{\mathbb{P}(D|A_k) \cdot \mathbb{P}(A_k)}{\sum\limits_k^N \mathbb{P}(D|A_k) \cdot \mathbb{P}(A_k)} = C \cdot \mathbb{P}(D|A_k) \cdot \mathbb{P}(A_k), \text{ где } C = \frac{1}{\sum\limits_k^N \mathbb{P}(D|A_k) \cdot \mathbb{P}(A_k)},$$

C — нормирующий множитель. Если необходимо выяснить, какая из гипотез наиболее вероятна (определяем лишь исход матча), то вычислять C не имеет смысла, так как для всех апостериорных оценок гипотез он одинаков.

Нам нужно выбирать ту гипотезу A_{k^*}, которая имеет наибольшую условную вероятность

$$\mathbb{P}(A_{k^*}|D) = \max_k \mathbb{P}(A_k|D).$$

Выбранная таким образом гипотеза называется MAP-гипотеза, где MAP есть аббревиатура «maximum a posteriori», или максимальная апостериорная гипотеза.

Пример 1. Вирусом заражен 1% населения Земли по случайной выборке. Имеется тест, который с точностью 95% определяет, болен ли человек или нет. Исходя из результата теста, пациент A болен. С какой истинной вероятностью человек болен?

Гипотезы:

$$\left.\begin{array}{l} A_1 - \text{ человек болен} \\ A_2 - \text{ человек здоров} \end{array}\right\} \text{ полная группа событий}$$

$$\left.\begin{array}{l} \mathbb{P}(A_1) = 0.01 \\ \mathbb{P}(A_2) = 0.99 \end{array}\right\} \text{ априорные оценки}$$

Событие $D = \{$ тест показал положительный результат $\}$.

Требуется вычислить $\mathbb{P}(A_1|D)$.

Нам известна следующая информация:

$$\begin{array}{l} \mathbb{P}(D|A_1) = 0.95 - \text{ тест положительный, если человек болен} \\ \mathbb{P}(D|A_2) = 0.05 - \text{ тест положительный, если человек здоров (ошибка)} \end{array}$$

Применим формулу Байеса:

$$\mathbb{P}(A_1|D) = \frac{\mathbb{P}(D|A_1)\mathbb{P}(A_1)}{\mathbb{P}(D|A_1)\mathbb{P}(A_1) + \mathbb{P}(D|A_2)\mathbb{P}(A_2)} = \frac{0.95 \cdot 0.01}{0.95 \cdot 0.01 + 0.05 \cdot 0.99} \approx 0.16.$$

Вероятность того, что пациент болен на самом деле – 16%.

Мораль: если вирусом болеет малое число людей, то при самом точном тесте мы не можем с достоверностью утверждать, что конкретный пациент болен. Чтобы с большей достоверностью выявить заболевание, нужно проделать повторные тесты. Если же болезнь имеет большое распространение, то достаточно проведения одного теста.

Наивный байесовский классификатор

Довольно часто D представляет собой совокупность признаков:

$$D = D_1 \cap D_2 \cap \cdots \cap D_L = D_1 D_2 \dots D_L,$$

где D_l — это элементарные события, каждое из которых может возникнуть или нет.

В ситуации, когда число L является большим, вычисление события D становится очень сложным, поскольку события D_l могут быть зависимыми событиями (как правило так и есть). Вероятность события D вычисляется так:

$$\mathbb{P}(D) = \mathbb{P}(D_1) \cdot \mathbb{P}(D_2|D_1) \cdot \mathbb{P}(D_3|D_1 D_2) \dots \mathbb{P}(D_L|D_1 D_2 \dots D_{L-1}). \tag{4}$$

В практике байесовского машинного обучения большую популярность получил так называемый **наивный байесовский классификатор**. Этот подход состоит в том, что мы считаем, что все события $\{D_l\}_{l=1}^M$ являются независимыми. Наивность классификатора заключается в том, что делая предположение о независимости признаков, мы понимаем, что в реальности это не так. В этом случае вероятность $P(D)$ может быть вычислена с помощью простой формулы

$$\mathbb{P}(D) = \mathbb{P}(D_1) \cdot \mathbb{P}(D_2) \dots \mathbb{P}(D_L).$$

Применение такого предположения, во-первых упрощает задачу, во-вторых улучшает работу классификатора. Улучшение происходит за счет того, что в формуле (4) условные вероятности получены из статистических наблюдений, которые, в свою очередь, зачастую получены с помощью применения множества различных предположений, а в нашем случае преположение одно — в этом заключена эффективность.

Поскольку признаки независимы (по допущению), то и их условные вероятности также независимы. Формула для MAP-гипотезы выглядит так:

$$\max_k \mathbb{P}(A_k|D) = \max_k \mathbb{P}(A_k) \cdot \mathbb{P}(D_1 D_2 \dots D_L|A_k) = \max_k \mathbb{P}(A_k)\mathbb{P}(D_1|A_k)\mathbb{P}(D_2|A_k)\dots\mathbb{P}(D_L|A_k)$$

Приведем пример использования наивного байесовского классификатора.

Пример 2. Имеется коллектив, в котором 60% женщин и 40% мужчин.

Для каждого члена коллектива рассматриваются следующие признаки:

- D_1 — наличие высшего образования;
- D_2 — рост больше 1.75м;
- D_3 — владение иностранным языком;
- D_4 — наличие водительских прав.

Наличие или отсутствие того или иного признака для каждого персонажа зафиксировано в большой сводной таблице:

№	D_1	D_2	D_3	D_4	
1	true	true	true	true	male
2	true	false	true	true	female
...					

Из таблицы получена статистика, представленная в следующей таблице:

	$\mathbb{P}(D_1)$	$\mathbb{P}(D_2)$	$\mathbb{P}(D_3)$	$\mathbb{P}(D_4)$
male	0.6	0.5	0.4	0.3
female	0.3	0.2	0.5	0.15

Процесс получения статистики из сводной таблицы называется **обучение**.

Некоторый человек имеет высшее образование, его рост 172 см, знает английский язык, не имеет водительских прав. Это мужчина или женщина?

Имеем две гипотезы:

- A_1 — это мужчина;
- A_2 — это женщина.

Воспользуемся наивным байесовским классификатором для оценки вероятностей:

$$\mathbb{P}(A_1|D) = C \cdot \mathbb{P}(A_1)\mathbb{P}(D_1|A_1)\mathbb{P}(D_2|A_1)\mathbb{P}(D_3|A_1)\mathbb{P}(D_4|A_1) = C \cdot 0.4 \cdot 0.6 \cdot (1-0.5) \cdot 0.4 \cdot (1-0.3) = C \cdot 0.0336;$$
$$\mathbb{P}(A_2|D) = C \cdot \mathbb{P}(A_2)\mathbb{P}(D_1|A_2)\mathbb{P}(D_2|A_2)\mathbb{P}(D_3|A_2)\mathbb{P}(D_4|A_2) = C \cdot 0.6 \cdot 0.3 \cdot (1-0.2) \cdot 0.5 \cdot (1-0.15) = C \cdot 0.0612.$$

Видно, что этот некто — женщина. Если мы хотим знать, с какой вероятностью, то необходимо посчитать нормирующий множитель:

$$\mathbb{P}(A_2|D) = \frac{0.0612}{0.0336 + 0.0612} \approx 0.65.$$

Замечание 1. В таблице-статистике одна из вероятностей, вообще говоря, может равняться нулю $P(D_l) = 0$. В этом случае ко всем элементам таблицы нужно прибавить 1 — это не сильно повлияет на статистику, а главное не занулит оценку классификатора.

С помощью представленного примера показана работоспособность байесовского обучения.

Приложения байесовского обучения

Байесовское обучение в приложениях позволяет решать задачи, связанные с принятием решений в условиях неопределенности, но при наличии дополнительной информации. Такая ситуация имеет место в задачах выбора наиболее вероятной гипотезы.

Принятие решений в условиях неопределенности — это всегда выбор одного решения из множества возможных решений при условиях, когда априори нет полной информации для выбора оптимального решения. С другой стороны, поскольку в реальности мы всегда имеем дополнительную информацию, которую можно использовать для более точного принятия решения, байесовское обучение позволяет использовать эту информацию оптимальным образом, на основе условных вероятностей, которые рассчитываются по статистическим данным.