Изкуствен интелект - летен семестър, 2023/2024 учебна година

Тема 11: Бейсов класификатор. Класификация на текстове

Често в областта на машинното самообучение се решават задачи, свързани с определяне на най-добрата хипотеза от някакво пространство на хипотези H при зададени обучаващи данни D. Един от начините да се определи значението на често използвания термин „най-добра хипотеза“ е да се счита, че най-добра е *най-вероятната* хипотеза при зададени данни D и налични в момента знания за априорните вероятности на различните хипотези в H.

Теоремата на Бейс дава директен метод за изчисляване на вероятностите на отделните хипотези. По-точно, тази теорема позволява да изчислим вероятността на една хипотеза въз основа на нейната априорна вероятност, вероятността да се наблюдават съответните данни при наличието на хипотезата и вероятността на самите данни.

Означения

С *P(h)* означаваме началната вероятност на предположението, че хипотезата *h* е вярна, преди да наблюдаваме каквито и да е обучаващи данни. *P(h)* се нарича *априорна вероятност на h* и може да отразява всякакви основни знания, с които разполагаме, за шансовете на *h* да бъде коректна хипотеза. Ако нямаме никакви априорни знания за възможни хипотези, можем просто да определим една и съща априорна вероятност за всяка от тях.

С *P(D)* означаваме априорната вероятност за наблюдаване на обучаващите данни *D* (т.е. вероятността на *D* без да са налице никакви знания за това, коя от хипотезите е вярна).

P(D|h) означава вероятността да наблюдаваме данните *D* при условие, че хипотезата *h* е вярна. В общия случай *условната вероятност P(x|y)* означава вероятността да се случи някакво събитие *x* при условие, че вече се е случило събитието *y*.

В машинното самообучение интерес представлява вероятността *P(h|D)*, че хипотезата *h* е вярна при наблюдавани обучаващи данни *D*. Тази вероятност се нарича *апостериорна вероятност на h*, тъй като отразява степента на нашата убеденост, че *h* е вярна *след* наблюдаване на *D.* Апостериорната вероятност отразява влиянието на данните *D*, за разлика от априорната вероятност *P(h)*, която не зависи от никакви данни.

Теоремата на Бейс е основата на Бейсовите методи за самообучение, тъй като осигурява начин за изчисляване на апостериорната вероятност *P(h|D)* от известни априорни вероятности *P(h)* и *P(D)*, както и от *P(D|h)*:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Както се вижда, *P(h|D)* нараства с нарастване на *P(h)* и *P(D|h)* и намалява с увеличаването на *P(D)*, тъй като колкото по-вероятно е, че данните *D* се наблюдават независимо от *h*, толкова по-малка поддръжка дава *D* за съществуването (валидността) на *h*.

В много задачи от областта на МС разглеждаме определено множество от възможни хипотези *H* и се интересуваме от намирането на най-вероятната от тях при наблюдавани данни *D* (или от максимално вероятна – ако има няколко). Такава максимално вероятна хипотеза се нарича *максимална апостериорна* (MAP) хипотеза. Можем да определим MAP хипотезата с използване на теоремата на Бейс:

$$h_{MAP} = argmax_{h \in H} P(h|D) = argmax_{h \in H} P(D|h)P(h)$$

В горното равенство вероятността *P(D)* е игнорирана, тъй като тя е постоянна и не зависи от *h.*

В някои случаи може да се приеме, че всички хипотези в *H* имат еднаква априорна вероятност. В такива случаи в горната формула има смисъл да разглеждаме само множителя *P(D|h)*, който често се нарича *възможност* (likelihood) на данните *D* при зададена *h*. Всяка хипотеза, максимизираща *P(D|h)*, се нарича *максимално възможна* (ML) хипотеза:

$$h_{ML} = argmax_{h \in H} P(D|h)$$

Често в машинното самообучение *H* се свързва с пространство от възможни целеви функции, подлежащи на научаване, а *D* – с обучаващи примери за някаква целева функция. Теоремата на Бейс е приложима и в по-общ контекст, където *H* се разглежда като множество от взаимно изключващи се предположения, чиято обща вероятност е равна на 1.

Досега разисквахме въпроса „каква е най-вероятната хипотеза при зададени обучаващи данни?“. Много често в практиката е по-важен въпросът „каква е най-вероятната *класификация на нов пример* при зададени обучаващи данни?“ Макар да изглежда, че отговорът на този въпрос може да бъде получен просто чрез прилагане на MAP хипотеза към новия пример, на практика е възможен по-добър подход.

Оптимален Бейсов класификатор

Нека например да разгледаме пространство от хипотези, съдържащо три хипотези *h1, h2* и *h3*. Нека техните апостериорни вероятности при зададени обучаващи данни са съответно 0.4, 0.3 и 0.3. Следователно *h1* е MAP хипотеза. Да предположим, че постъпва нов пример, който се класифицира като положителен от *h1* и като отрицателен от *h2* и *h3*. Отчитайки всички хипотези, вероятността, че примерът е положителен, е 0.4, а че примерът е отрицателен – 0.6. Най-вероятната класификация в този случай е различна от класификацията, генерирана от MAP хипотезата.

В общия случай най-вероятната класификация на даден нов пример се получава от комбиниране на предсказанията на всички хипотези, претеглени от техните апостериорни вероятности. Ако за възможна класификация на новия пример може да приеме една от стойностите v_i от множеството V, то вероятността $P(v_i|D)$ за коректна класификация на новия пример като v_i се получава с използване на формулата

$$P(v_i|D) = \sum_{h_j \in H} P(v_i|h_j) P(h_j|D)$$

Оптималната Бейсова класификация на новия пример е стойността v_{i*}, за която $P(v_{i*}|D)$ е максимална:

$$v_{i*} \equiv argmax_{v_i \in V} \sum_{h_j \in H} P(v_i|h_j) P(h_j|D)$$

Всяка система, която класифицира в съгласие с горната формула, се нарича *оптимален Бейсов класификатор*. Никой друг метод за класификация, използващ същото пространство на хипотези и същите априорни знания, *не може да подобри* (като средно) този метод. Той максимизира вероятността, че новият пример е класифициран правилно при зададените налични данни, пространството от хипотези и априорните вероятности на хипотезите.

Макар че оптималният Бейсов класификатор постига най-доброто поведение, което може да бъде постигнато при наличните обучаващи данни, неговото прилагане е много скъпо от изчислителна гледна точка. Това се дължи на необходимостта да бъдат изчислени апостериорните вероятности на всички възможни хипотези в H, след което трябва да бъдат комбинирани предсказанията на всички хипотези, за да бъде класифициран разглежданият нов пример.

Наивен Бейсов класификатор

Да разгледаме случая, когато примерът *x* за научаваното понятие се описва с конюнкция от атрибутни стойности, а целевата функция *f(x)* приема дискретни стойности от дадено крайно множество *V*.

Дадени са множество от обучаващи примери *D* и нов пример, подлежащ на класификация: $x = \langle a_1, \ldots, a_n \rangle$. Алгоритмът за обучение трябва да предскаже стойността на целевата функция за този пример, т.е. да класифицира този пример.

Съгласно Бейсовия подход класифицирането на примера се свежда до избор на най-вероятната стойност на целевата функция v_{MAP} при зададените атрибутни стойности $<a_1, \ldots, a_n>$, описващи примера:

$$v_{MAP} = argmax_{v_j \in V} P(v_j | a_1, \ldots, a_n)$$

Използвайки теоремата на Бейс, получаваме:

$$v_{MAP} = argmax_{v_j \in V} \frac{P(a_1, \ldots, a_n | v_j) P(v_j)}{P(a_1, \ldots, a_n)} =$$

$$argmax_{v_j \in V} P(a_1, \ldots, a_n | v_j) P(v_j) \qquad (1)$$

Сега трябва да опитаме да оценим двата терма в равенство (1) на базата на обучаващите данни. $P(v_j)$ е лесно да се оцени чрез изчисляване на честотата на срещане на всяка стойност v_j на целевата функция в множеството от обучаващи данни. Но количествена оценка на другия терм по този начин е практически невъзможна, освен ако не разполагаме с много голямо множество от обучаващи данни.

Проблемът е в това, че броят на термовете от този вид е равен на броя на всички възможни примери (комбинации от атрибутните стойности), умножен по броя на възможните стойности на целевата функция. Следователно, за да получим едно надеждно приближение за подобна оценка, трябва да прегледаме многократно всеки възможен пример в пространството от примери.

Един възможен начин за избягване тези усложнения е да предположим, че атрибутните стойности са *условно независими при зададената стойност на целевия атрибут*, т.е. че за дадената стойност на целевата функция вероятността за съществуване на конюнкцията на атрибутните стойностите $a_1, \ldots, a_n$ е равна на произведението на вероятностите на отделните атрибутни стойности:

$$P(a_1, \ldots, a_n | v_j) = \prod_{i=1}^{n} P(a_i | v_j)$$

Заместваме този терм в (1) и получаваме подхода, известен като *наивен Бейсов класификатор:*

$$v_{NB} = argmax_{v_j \in V} P(v_j) \prod_{i=1}^{n} P(a_i | v_j) \qquad (2)$$

Ще обърнем внимание, че в случая на наивния Бейсов класификатор броят на различните термове $P(a_i|v_j)$, чиито стойности трябва да бъдат оценени чрез обучаващите данни, е равен на броя на възможните атрибутни стойности, умножен по броя на възможните стойности на целевия атрибут – това число е значително по-малко от броя на термовете, които са необходими за оценяването на $P(a_1, \ldots, a_n \mid v_j)$.

Пример: Научаване на класификация на текстове

В качеството на илюстративен пример ще разгледаме един общ алгоритъм за научаване на класификация на текстове, базиран на наивния Бейсов класификатор.

Нека разгледаме пространството *X*, съдържащо всички възможни *текстови документи* (т.е. всички възможни низове от думи и пунктуационни знаци с произволна дължина). Дадени са обучаващи примери за някаква неизвестна целева функция *f(x),* която приема стойности от дадено крайно множество *V*. Задачата е да се научим от тези обучаващи примери да предсказваме стойността на целевия атрибут за нови текстови документи. За илюстрация ще разгледаме класификацията на документите само на две групи – *интересни* и *неинтересни*.

Най-напред е необходимо да се намерят отговори на два основни въпроса:

- как ще представяме произволен текстов документ в термините на атрибутните стойности;
- как да бъдат оценени вероятностите, необходими за работата на наивния Бейсов класификатор.

Предложеният подход е много прост: при зададен текстов документ, съдържащ например текст на английски език, се дефинира по един атрибут за всяка отделна позиция на дума в текста (за всеки конкретен пореден номер на дума в текста), а като стойност на този атрибут – съответната английска дума, намираща се на конкретната позиция.

Например цитираният по-долу параграф може да се опише със 111 позиции на думите. Стойността на първия атрибут е думата ”our”, на втория – “approach”, и т.н.

Our approach to representing arbitrary text documents is disturbingly simple: Given a text document, such as this paragraph, we define an attribute for each word position in the document and define the value of that attribute to be the English word found in that position. Thus, the current paragraph would be described by 111 attribute values, corresponding to the 111 word positions. The value of the first attribute is the word "our," the value of the second attribute is the word "approach," and so on. Notice that long text documents will require a larger number of attributes than short documents. As we shall see, this will not cause us any trouble.

Следователно, броят на необходимите атрибути е равен на броя на думите в текста, т.е. по-дългите текстови документи изискват по-голям брой атрибути от по-късите документи.

След като сме избрали подходящо кодиране на текстовите документи, вече можем да приложим наивния Бейсов класификатор. Да предположим, че разполагаме с набор от 700 документа, които са класифицирани като „*неинтересни*“, и други 300 документа, които са класифицирани като „*интересни*“. Предоставен ни е нов документ, който трябва да класифицираме като „интересен“ или „неинтересен“.

Нека предположим, че този документ е съставен от цитирания по-горе параграф.

В такъв случай от равенство (2) получаваме

$$v_{NB} = argmax_{v_j \in \{\text{интересни,неинтересни}\}} P(v_j) \prod_{i=1}^{111} P(a_i | v_j) =$$

$$argmax_{v_j \in \{\text{интересни,неинтересни}\}} P(v_j) P(a_1 = \text{"our"} | v_j)$$

$$P(a_2 = \text{"approach"} | v_j) \ldots (a_{111} = \text{"trouble"} | v_j)$$

Използваното в наивния Бейсов класификатор предположение за условната независимост на атрибутите в нашия контекст означава, че вероятностите на думите, намиращи се на определени позиции, не зависят от думите, намиращи се на други позиции. Очевидно обаче това предположение *е невярно*.

Например, вероятността да открием на някоя позиция думата “learning” е по-голяма, ако преди нея се намира думата “machine”. Въпреки че посоченото базово предположение е невярно, на практика нямаме голям избор, тъй като без него броят на вероятностите, които трябва да бъдат изчислени, е изключително голям. За радост наивният Бейсов класификатор работи добре и в повечето случаи на очевидно нарушаване на предположението за условна независимост, както е показано в редица литературни източници.

За да изчислим получения израз, трябва да имаме приближения за стойностите на $P(v_j)$ и $P(a_i = w_k|v_j)$, като с w_k тук ще означаваме k-тата дума в речника на английския език. Първият терм може да бъде оценен лесно като пропорцията (относителния дял) на съответния клас в обучаващите данни: *P(интересни) = 0.3*; *P(неинтересни) = 0.7*. Оценката на условните вероятности (например $P(a_1 =$ *“our”*$|$*неинтересни)*) е по-сложна, тъй като изисква да се оцени по един такъв терм за всяка комбинация от текстова позиция, английска дума и съответната целева стойност.

В речника на английския език има около 50000 различни думи; работим с 2 стойности на целевия атрибут и 111 позиции в текста, подлежащ на класифициране. Следователно трябва да оценим *2*111*50000 ≈ 10 000 000* комбинации за обучаващите данни.

Тук е възможно да се направи още едно разумно предположение, значително намаляващо броя на вероятностите, които трябва да оценим. Ще предположим, че вероятността да срещнем някоя конкретна дума w_k (например “chocolate”) не зависи от конкретната позиция на думата в текста. Формално това означава, че предполагаме, че атрибутите са независими и равномерно разпределени при зададена целева класификация, т.е. че $P(a_i = w_k|v_j) = P(a_m = w_k|v_j)$ за всички $i, j, k, m.$

По този начин можем да оценим цялото множество от вероятности $P(a_1 = w_k|v_j)$, $P(a_2 = w_k|v_j)$, … чрез една единствена, независима от позицията вероятност $P(w_k|v_j)$. Това означава, че са необходими оценките на само $2*50000$ различни стойности за различните $P(w_k|v_j)$.

За да завършим проектирането на нашия алгоритъм за самообучение, трябва да изберем подходящ метод за оценяването (пресмятането) на необходимите вероятности. Ще използваме известната от литературата формула

$$P(w_k|v_j) = \frac{n_k+1}{n+|\text{Речник}|},$$

където n е общият брой позиции на думите във всички обучаващи примери, които имат стойност на целевия атрибут v_j; n_k е броят на срещанията на думата w_k сред тези n позиции; $|Речник|$ е общият брой на различните думи, намерени в обучаващите данни.

Резюме

За да изчисли най-вероятната класификация на всеки нов пример, оптималният Бейсов класификатор комбинира предсказанията на всички алтернативни хипотези, претеглени от техните апостериорни вероятности.

Наивният Бейсов класификатор е метод за Бейсово обучение, който е доказал своята приложимост за множество практически задачи. Той се нарича „наивен“, защото използва улесняващото предположение, че атрибутните стойности са условно независими при зададена класификация на примера. Когато това предположение е изпълнено, наивният Бейсов класификатор извежда MAP класификацията. Дори когато това предположение не е изпълнено, използваният класификационен метод остава много ефективен.