Лекция 20. Изчислителната теория на самообучение

10.1. Увод

Един от естествените въпроси в машинно самообучение е какви общи закони управляват изкуствени (или естествени) самообучаващи се системи. Дали е възможно да идентифицираме класове задачи за обучение, които са „природно" сложни или леки, не зависимо от това, какъв алгоритъм за самообучение се използва? Можем ли да определим броя на обучаващите примери, необходими или достатъчни, за да гарантираме успешното самообучение? Как се променя този брой, ако на обучаемият ще е позволено да задава въпроси (заявки) към учителя вместо да разполага само с една случайна извадка от обучаващи примери? Може ли някой да определи броя на грешки, които една система за самообучение ще направи преди да научи целевата функция? Да ли е възможно да бъде определена „вградена" изчислителна сложност на различни класове на задачи за самообучение?

Макар че не са още известни общите отговори на всички тези въпроси, вече са изградени фрагменти от една изчислителна теория на самообучение, която позволява да се отговори на тези въпроси за задачи, описани в един определен контекст. В следващите две лекции ние ще се фокусираме върху задачата за индуктивно научаване на една неизвестна целева функция по зададени нейни обучаващи примери и пространство от възможни хипотези. В този контекст ще разгледаме основно въпроси, свързани с броя на примери, достатъчни за успешното научаване на целевата функция, и с броя на грешки, които една система за самообучение прави преди да успее. Както ще видим, е възможно да установим количествени граници на тези величини в зависимост от такива характеристики на задачата за самообучение като:

- Размерът или сложността на пространството от хипотези, използвани от системата за самообучение;
- Точността, с която целевото понятие трябва да бъде апроксимирано (научено);
- Вероятността, с която системата за самообучение трябва да изведе една успешна хипотеза и
- Начинът, по който обучаващите примери се представят на системата.

В повечето случаи ще се фокусираме не на отделни алгоритми за самообучение, а на широки класове от такива алгоритми, характеризиращи се чрез пространства от хипотези, които те разглеждат, чрез представянето на обучаващите примери и т.н. Целта ни е да отговорим на следните въпроси:

- *Сложността на извадката* – колко *примера* са необходими на една система за самообучение, за да осигури (с висока вероятност) сходимостта към успешната хипотеза?
- *Изчислителната сложност* – колко *изчислителни ресурси* са необходими на една система за самообучение, за да осигури (с висока вероятност) сходимостта към успешната хипотеза?
- *Границата на грешки* – колко обучаващи примера ще бъдат неправилно класифицирани от една система за самообучение преди намиране на успешната хипотеза?

Трябва да се отбележи, че съществуват много специфични условия, при които ние можем да търсим отговори на тези въпроси. Например, съществуват множество различни начини да определим, какво означава, че една система за самообучение е „успешна". Ние можем да укажем, че за да успее, тя трябва да изведе една хипотеза, която е идентична на целевото понятие. Алтернативно, ние можем просто да поискаме тя да извежда една хипотеза, която е в съгласие с целевото понятие в повечето от случаи. По същия начин можем да специфицираме как обучаващите примери трябва да постъпват в обучаваща се система – те могат да са представени от един заинтересуван учител, да бъдат получени от самата система при извършване на определени експерименти, или просто да бъдат генерирани по случаен начин от някакъв процес, който се намира извън контрола на системата за самообучение. Както може да се очаква, отговорите на посочените по-горе въпроси ще зависят от тези специфични условия - *модела на самообучение*, който ние сме избрали.

20.2. Вероятностно научаване на една приблизително точна хипотеза

В този раздел ще разгледаме едни конкретни условия за осъществяване на задачата за самообучение, наречени моделът за вероятностно приблизително точно (*probably approximately correct – PAC*) самообучение, а след това ще разгледаме въпроса, колко обучаващи примера и колко изчисления са необходими, за да бъдат научени различни класове на целеви функции в контекста на този PAC модел. За простота ще ограничим дискусията до случая за научаване на двоични понятия от обучаващите данни не съдържащи шум. Обаче, много от резултатите могат да бъдат разширени и към по-общ сценарий за научаване на целевите функции с реални стойности (виж, например Natarajan 1991), а някои – към самообучение от определени типове зашумени данни (виж, например Laird 1988; Kearns and Vazirani 1994).

20.2.1. Постановка на задачата

Както вече беше прието в предишните лекции, чрез X ще означаваме множеството от всички възможни примери, върху които могат да се дефинират целевите функции. Например, X може да представлява множеството от всички хора, описани с атрибути *възраст* (със стойности *млад* или *възрастен*) и *височина* (*нисък* или *висок*). Чрез C ще означаваме някое множество от целеви понятия, които обучаващата се система може да научи върху примерите от X. Всяка целево понятие c от C отговаря на определено подмножество на X или, еквивалентно, на определена двоична функция $c : X \rightarrow \{0, 1\}$. Например едно такова целево понятие c от C може да е „хората, които карат ски". Ако x е положителния пример на c, ще пишем, че $c(x) = 1$; ако x е отрицателния пример, то $c(x) = 0$.

Ще подразбираме, че всички примери са генерирани по случаен начин от X в съответствие с някое разпределение на вероятности $\mathcal{D}$. Например $\mathcal{D}$ може да бъде разпределението на примери, генерирани при наблюдаване на хора, които посещават някой от големите спортни магазини в София. В общия случай $\mathcal{D}$ може да бъде произволно разпределение, което не е известно на системата за самообучение. Единственото изискване, което налагаме на $\mathcal{D}$, е че то трябва да бъде стационарно, т.е.

да не се променя с течение на време. Обучаващите примери се генерират чрез вземане по случаен начин на пример x от $\mathcal{D}$ и представянето му заедно със стойността на неговата целева функция $c(x)$ на системата за самообучение.

При научаване на целевото понятие обучаващата се система L разглежда определено множество H от възможни хипотези. Например H може да е множеството от всички хипотези, представени чрез конюнкции от атрибутите *възраст* и *височина*. След преглеждането на последователността от обучаващи примери на целевото понятие c, L може да изведе някоя хипотеза h от H, която представлява едно приближение на c. Ще оценяваме успеха на L чрез поведението на h върху нови примери, взети по случаен начин от X в съответствие с $\mathcal{D}$ – същото разпределение на вероятности, което е използвано за генериране на обучаващите примери.

При тази постановка на задачата ще се интересуваме от определяне на поведение на различни системи за самообучение L, използващи различни пространства на хипотези H, когато отделни целеви понятия се взимат от различни класове C. Тъй като изискваме L да бъде достатъчно обща, за да научи всякакво целево понятие от C, не зависимо от разпределението на обучаващите примери, ние често ще се интересуваме от анализа на най-лошия случай от всички възможни целеви понятия от C и всички възможни разпределения на примери $\mathcal{D}$.

20.2.2. Грешка на хипотеза

Тъй като се интересуваме от това, колко близо изходната хипотеза h на система за самообучение се приближава до истинското целево понятие c, ще напомня дефиницията на понятие "истинската грешка" на хипотеза h по отношение на целевото понятие c и разпределението на примери $\mathcal{D}$ (виж. Лекция 18):

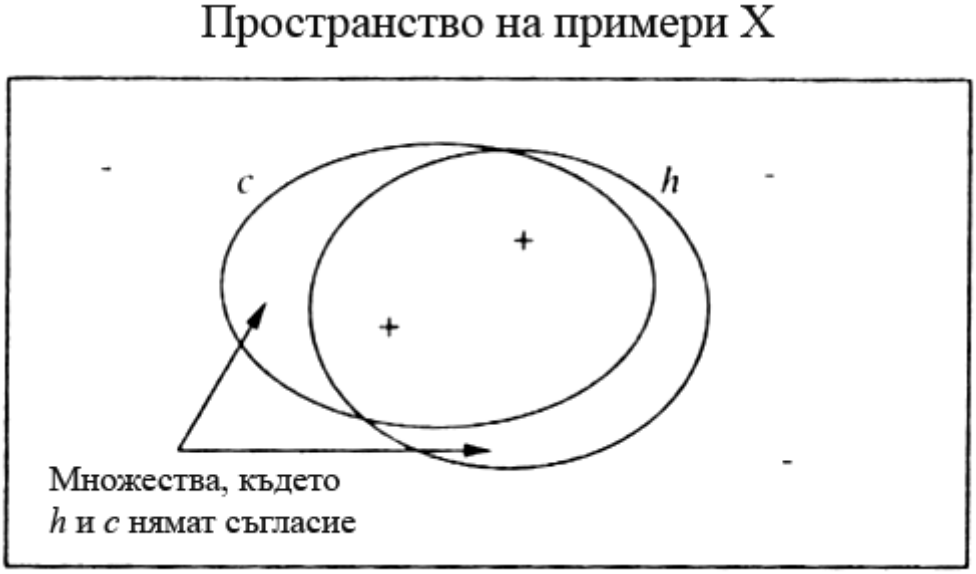
Определение: Истинската грешка $error_{\mathcal{D}}(h)$ на хипотеза h по отношение на целевото понятие c и разпределението $\mathcal{D}$ е вероятността, че h ще класифицира неправилно един произволен пример x взет по случаен начин от $\mathcal{D}$:

$$error_D(h) = \Pr_{x \in D}[c(x) \neq h(x)]$$

На Фиг. 20-1 е показана тази дефиниция в графична форма. Понятията c и h са показани като множества примери от X, които те маркират като положителни. Грешката на h по отношение на c е вероятността, че един случайно взет пример ще попадне в областта, където h и c нямат съгласие (т.е. разликата на техните множества). Обърнете внимание, че ние избрахме да дефинираме грешката върху цялото разпределение на примери, а не просто върху обучаващите примери, тъй като това е истинската грешка, която ние очакваме да срещнем при реалното използване на научената хипотеза h върху последващи примери взети от $\mathcal{D}$.

Обърнете също внимание, че грешката зависи строго от неизвестното разпределение на вероятности $\mathcal{D}$. Например, ако $\mathcal{D}$ е едно равномерно разпределение, назначаващо една и съща вероятност на всеки пример от X, то грешката на хипотезата от Фиг. 20-1 ще бъде равна на частта от цялото пространство на примери, която попада в областта, където h и c нямат съгласие. Обаче, същите h и c ще имат значително по-голяма грешка, ако $\mathcal{D}$

назначава много висока вероятност на примери, за които h и c не са съгласни. В един краен случай, когато $\mathcal{D}$ назначава нулева вероятност на примерите, за които $h(x) = c(x)$, грешката на h от Фиг. 20-1, ще е равна на 1, не зависимо от това, че h и c са в съгласие върху едно голямо множество от примери (с нулева вероятност).



Фиг. 20-1. Грешката на хипотеза h по отношение на целевото понятие c. Грешката на h по отношение на c е вероятността, че един случайно взет пример попада в областта, където h и c нямат съгласие за класификация на примера. Точките + и – указват на положителни и отрицателни обучаващи примери. Обърнете внимание, че h има ненулева грешка по отношение на c не зависимо от факта, че h и с имат съгласие за всички пет обучаващи примери, наблюдавани до сега..

И на края, напомням, че грешката на h по отношение на c не е пряко наблюдаема от обучаващата се система. L може само да прецени поведението на h върху *обучаващите примери* и да избере своя изходна хипотеза само върху тези данни. Ще използваме термин "обучаваща грешка", за да означим част от обучаващите примери, неправилно класифицирани от h. Повечето от нашия анализ на сложността на самообучение ще бъде концентриран около въпроса „каква е вероятността, че наблюдаваната обучаваща грешка на h дава едно подвеждащо приближение на истинската грешка $error_{\mathcal{D}}(h)$?".

Обърнете внимание на близката връзка между този въпрос и въпросите, разглеждани в лекция 18. Напомням, че в тази лекция сме дефинирали понятие „извадковата грешка" на h по отношение на множество от примери S като тази част от S, която е класифицирана неправилно от h. Обучаващата грешка, дефинирана по-горе, е просто извадковата грешка в случая, когато S представлява цялото обучаващо множество от примери. В същата лекция сме определили вероятността, че извадковата грешка ще даде подвеждащата оценка за истинската грешка при предположението, че извадковите данни S се взимат независимо от h. Обаче, когато S е цялото обучаващо множество, научената хипотеза h в голяма степен зависи от S! По тази причина в тази лекция ще направим анализ, адресиращ този много важен специален случай.

20.2.3. PAC научаемост

Нашата цел е да опишем класове на целеви понятия, които могат надеждно да бъдат научени от един разумен брой случайно взети обучаващи примери, използвайки за тази цел един разумен обем от изчисления.

Какъв вид предположения за научаемостта трябва да приемем за верни? Можем да опитаме да определим броя на обучаващи примери, необходими да научим хипотеза h, за която $error_{\mathcal{D}}(h) = 0$. За съжаление при описаната по-горе постановка на задачата това предположение се оказва напълно безсмислено по следните две причини: първо, освен ако нямаме обучаващи примери, съответстващи на всеки възможен пример от X (едно напълно нереалистично предположение), е възможно съществуване на няколко хипотези съвместими с предоставените за обучение примери и обучаващата се система не може със сигурност да избере само една от хипотезите като.отговаряща на целевото понятие. Второ, приемайки, че обучаващите примери се взимат по случаен начин, винаги съществува ненулева вероятност, че обучаващите примери, използвани от обучаващата се система, са подвеждащи. (Например, макар че ние може често да видим скиори с различна височина, в един определен ден съществува един малък шанс да видим, че всички наблюдавани скиори са над 2 м.)

За да се приспособим към тези два затруднения, ще ослабим нашите изисквания към обучаваща се система по следните два начина. Първо, няма да изискваме от системата да извежда хипотеза с нулевата грешка – само ще искаме тази грешка да бъде ограничена от някоя константа - ε, която може да бъде произволно малка. Второ, няма да изискваме обучаващата се система да успява при *всяка* последователност от случайно взети обучаващи примери – вместо това ще искаме само нейната вероятност за неуспех да бъде ограничена от някоя малка константа - δ, която може да бъде произволно малка. С други думи ще искаме само обучаващата се система *вероятностно* да научи някоя хипотеза, която е *приблизително точна* – от тук и следва термин „вероятностно приблизително точно (*PAC*) самообучение".

Да разгледаме някой клас C от възможни целеви понятия и една система за самообучение L, използваща пространство от хипотези H. Ще говорим, че понятийният клас C е PAC научаем от L, използвайки H, ако за всяко целево понятие c от C, L ще извежда с вероятност $(1 - \delta)$ хипотеза h с $error_{\mathcal{D}}(h) < \varepsilon$ след използване на едно разумно количество обучаващи примери и изпълнение на един разумен обем изчисления.

Определение: Нека един клас от понятия C е дефиниран върху множество от примери X с размер n и има една система за самообучение L, използваща пространство от хипотези H. Ще наричаме класът C ***PAC-научаем*** от L, използващ H, ако за всички $c \in C$, разпределения $\mathcal{D}$ върху X, ε - такова, че $0 < \varepsilon < ½$ и δ, такова че $0 < \delta < ½$, системата за самообучение L ще извежда с вероятност $(1 - \delta)$ хипотеза $h \in H$, такава че $error_{\mathcal{D}}(h) \leq \varepsilon$ за време, което е полиномиално по отношение на $1/\varepsilon$, $1/\delta$, n и $size(c)$.

Нашето определение изисква две неща от L. Първо, L трябва с произволно висока вероятност $(1 - \delta)$ да извежда хипотеза, имаща произволно малка истински грешка (ε). Тя трябва да го прави ефективно – за време, което расте най-много полиномиално по отношение на $1/\varepsilon$ и $1/\delta$, което определя строгостта на нашите изисквания към изходната

хипотеза, и по отношение на n и *size(c)*, които определят присъщата сложност на основното пространство от примери X и класа на понятия C (n е размерът на примери в X). Например, ако примерите в X са конюнкции от k двоични признака, то $n = k$. Вторият параметър на пространството – размерът на c (*size(c)*) представлява дължина на кодиране на c в C при наличието на някое представяне на C. Например, ако понятията в C са конюнкции, включващи до k двоични признака, като всяко понятие е описано със списъка на индексите на признаците в конюнкцията, то *size(c)* ще е броят на двоичните признаци, *действително използвани* за описание на c.

На пръв поглед нашата дефиниция на PAC самообучение се отнася само за изчислителните ресурси, необходими за самообучение, докато на практика ние обикновено повече се интересуваме от броя на обучаващи примери, необходими за тази цел. Обаче и двете неща са тясно свързани: ако L се нуждае от определено минимално време за обработка на всеки обучаващ пример, то за да бъда клас C PAC научаем от L, *L трябва да учи от едно полиномиално количество от обучаващи примери*. Фактически един типичен подход да покажеш, че някой клас от целеви понятия C е PAC научаем, е да покажеш отначало, че всяко целево понятие в C може да бъде научено от едно полиномиално количество обучаващи примери, а след това, да покажеш, че времето за обработка на един пример е също така ограничено полиномиално.

Преди да продължим, трябва да посочим наличието на едно ограничително предположение, неявно заложено в нашата дефиниция на PAC научаемост. Тази дефиниция неявно предполага, че пространството на хипотези, използвано от системата за самообучение L, съдържа за всяко целево понятие от C по една хипотеза с произволно малка грешка. Това следва от изискването, че системата успява, когато границата на грешката ε е произволно близка до нула. Естествено е, че това трудно да бъде осигурено, ако C не е известно предварително (какво е, например, C за една програма, която трябва да се научи да разпознава лица по снимки?), освен ако H е направено така, че представлява множество-степен от X. Както вече беше отбелязано в една от първите лекции, подобното „безпристрастно" пространство от хипотези няма да поддържа точно обобщение от едно разумно количество обучаващи примери. Въпреки това резултатите, базирани върху модела за PAC самообучение, дават полезни представи за относителната сложност на различни задачи за самообучение и за скоростта, с която класификационната точност се подобрява с добавяне на нови обучаващи примери. Повече от това, по нататък в раздел 20.3.1 ние ще отхвърлим това рестриктивно предположение и ще разгледаме случай, в който обучаващата се система не прави никакви предварителни предположения за вида на целевото понятие.

20.3. Извадковата сложност за крайни пространства от хипотези

Както беше отбелязано, PAC научаемост е в голяма степен се определя от броя на обучаващи примери, необходими за обучаващата се система. Нарастване на този брой от необходими обучаващи примери в зависимост от размера на задачата, наречено „*извадкова сложност*" (sample complexity), е характеристика, предизвикваща обикновено най-голям интерес. Причината за това е, че в най-често срещани практически условия факторът, който в най-голяма степен ограничава успех на една система за самообучение, е ограниченото наличие на обучаващи данни.

В настоящия раздел ще намерим една обща граница на извадковата сложност за един много широк клас от системи за самообучение, наречени „логични системи за самообучение" (consistent learners). Напомням, че една система са самообучение е *логична*, ако извежда хипотези, които абсолютно точно класифицират (съответстват на) обучаващите данни. Изискването една система да е логична е доста разумно, тъй като ние обикновено предпочитаме хипотези, които са в съгласие с обучаващите данни, пред тези, които не го правят. Напомням, че много от разгледаните по-рано алгоритми (включително тези от първите две лекции) представляват логични системи за самообучение.

Можем ли ние да изведем граница (оценка) за броя на обучаващи примери, необходими на произволна логична система за самообучение, не зависимо от конкретния алгоритъм, който се използва за извеждане на съвместими („логични") хипотези? Отговорът е „да". За да направим това, трябва да припомним определението за пространството на версии. Пространството на версии $VS_{H,D}$ е множество от всички хипотези $h \in H$, които правилно класифицират обучаващите примери D:

$$VS_{H,D} \equiv \{h \in H \mid (\forall < x, c(x) > \in D) \quad h(x) = c(x)\}$$

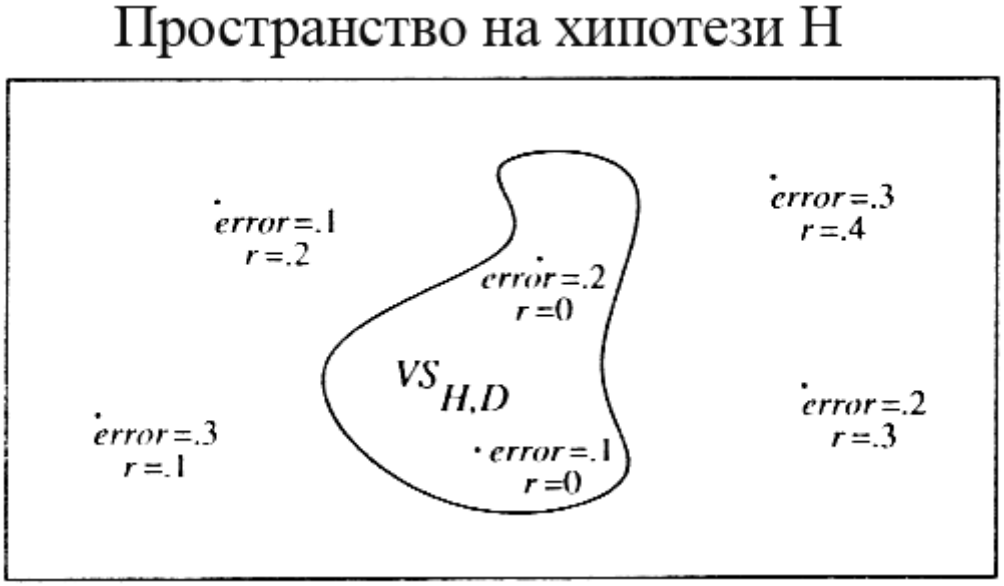
Важността на пространството на версии се състои в това, че всяка „логична" система за самообучение извежда хипотези, принадлежащи към пространството на версии, не зависимо от пространството на примери X, пространството на хипотези H или обучаващите данни D. Следователно, за да ограничим броя на примери, необходими на една система за самообучение, трябва само да ограничим броя на примери, необходими, за да осигурим, че пространството на версии не съдържа никакви неприемливи хипотези. Това утвърждение се задава по-точно със следното определение:

Определение: Нека H е пространство от хипотези, c е целево понятие, $\mathcal{D}$ е разпределение на примери и D е множество от обучаващи примери на c. Пространството на версии $VS_{H,D}$ се нарича ***ε-изчерпващо*** по отношение на c и $\mathcal{D}$, ако всяка хипотеза h от $VS_{H,D}$ има истинската грешка по-малка от ε по отношение на c и $\mathcal{D}$:

$$(\forall h \in VS_{H,D}) \qquad error_{\mathcal{D}}(h) < \varepsilon$$

Тази дефиниция е илюстрирана на Фиг. 20-2.

Пространството на версии е ε-изчерпващо само в случая, когато всички хипотези, съвместими с наблюдавани обучаващи примери (т.е. тези с нулева *обучаваща* грешка), имат истинската грешка по-малка от ε. Естествено е, че от гледната точка на системата за самообучение всичко, което тя може да знае, е че тези хипотези пасват на обучаващи данни еднакво добре – те имат нулева обучаваща грешка. Само един външен наблюдател, който знае идентичността на целевото понятие, може да определи със сигурност, дали пространството на версии е ε-изчерпващо. Учудващо е, че вероятностният аргумент ни позволява да ограничим вероятността, че пространството на версии ще бъде ε-изчерпващо след определен брой обучаващи примери, дори без да знаем идентичността на целевото понятие или разпределението, от което се взимат обучаващи примери. Тази граница представена във вид на следната теорема (Hausseler 1988):



Фиг. 20-2. *Изчерпване на пространството на версии.* Пространството на версии $VS_{H,D}$ е подмножеството на хипотези $h \in H$, които имат нулевата обучаваща грешка (означени на фигурата с $r = 0$). Ясно е, че истинската грешка $error_{\mathcal{D}}(h)$ (означена на фигурата с *error*) може и да не е нулева, дори за хипотези с нулевата грешка върху обучаващите данни. Пространството на версии се нарича ε-изчерпващо, когато всички хипотези h, останали в $VS_{H,D}$, имат $error_{\mathcal{D}}(h) < \varepsilon$.

Теорема: ε-изчерпване на пространството на версии. Ако пространството на хипотези H е крайно и D е последователността от $m \geq 1$ случайно взети примери на някое целево понятие c, то за всяко $0 \leq \varepsilon \leq 1$ вероятността, че пространството на версии $VS_{H,D}$ не е ε-изчерпващо (по отношение на c), е по-малка или равна на

$$| H | e^{-\varepsilon m}$$

Теоремата ограничава вероятността, че m обучаващи примера няма да са достатъчни да елиминират всички „лоши" хипотези (т.е. хипотези с истинската грешка по-голяма от ε) за всяка „логична" система за самообучение, използваща пространство от хипотези H.

Нека да използваме този резултат, за да определим броя на обучаващи примери, необходими за намаляване на тази вероятност от неуспех до някакво желано ниво δ:

$$| H | e^{-\varepsilon m} < \delta \quad (20.1)$$

Решавайки това неравенство относително m получаваме:

$$m \geq \frac{1}{\varepsilon}(\ln | H | + \ln(1/\delta)) \quad (20.2)$$

Полученото неравенство определя общата граница на броя на обучаващи примери, достатъчни на всяка „логична" система за самообучение, за да научи успешно всяко целево понятие от H за всяка желана стойност на δ и ε. Този брой обучаващи примери m е достатъчен, за да осигури, че всяка съвместима хипотеза ще бъде вероятно (с вероятността $(1 - \delta)$) приблизително (с грешка ε) точна. Обърнете внимание, че m расте линейно по отношение на $1/\varepsilon$ и логаритмически по отношение на $1/\delta$. То също така расте логаритмически по отношение на размера на пространство от хипотези H.

Трябва да се отбележи, че получената граница може да е съществено надценена. Например, макар, че вероятността от пропадане да бъде изчерпано пространството от хипотези трябва да лежи в интервала [0, 1], границата, задавана от теоремата, расте линейно по отношение на $|H|$. За достатъчно голямо пространство от хипотези тази граница може лесно да стане по-голяма от единица. Като резултат границата, задавана от неравенството (20.2), може значително да надхвърля броя на необходими обучаващи примери. На практика в доста случаи може да бъде намерена една много по-тясна граница, както и граница, която покрива безкрайно големи пространства от хипотези. Тези граници ще бъдат разгледани в следващата лекция.

20.3.1. Агностично самообучение и несъвместими хипотези

Неравенство (20.2) е важно, тъй като ни показва, колко обучаващи примера са достатъчни, за да осигурим (с вероятност $(1 - \delta)$), че всяка хипотеза от H, имаща нулева обучаваща грешка, ще има истинската грешка не по-голяма от ε. За съжаление, ако H не съдържа някое определено целево понятие c, то хипотезата с нулева обучаваща грешка не може винаги да бъде намерена. В този случай това, което най-много можем да поискаме от нашата система за самообучение, е тя да извежда такава хипотеза от H, която има *минимална грешка* върху обучаващите примери. Системата за самообучение, която не прави никакви предположения за това, дали целевото понятие е представимо в H, а просто намира хипотезата с минимална обучаваща грешка, често се нарича *агностична,* тъй като не прави никакви предварителни предположения за това, дали $C \subseteq H$.

Макар, че неравенството (20.2) се базира върху предположение, че системата за самообучение извежда хипотеза с нулева обучаваща грешка, подобната граница може да бъде намерена и за този по-общ случай, в който системата приема хипотези и с ненулева обучаваща грешка. За да дефинираме по-точно този случай, нека да приемем, че D е определено множество от обучаващи примери, достъпни за системата, докато $\mathcal{D}$ – означава вероятностното разпределение върху цялото множество от примери. Чрез $error_D(h)$ ще означаваме обучаващата грешка на хипотеза h. В частност $error_D(h)$ се определя като част от обучаващите примери в D, които класифицирани неправилно от h. Обърнете внимание, че $error_D(h)$, дефинирано върху конкретната извадка от обучаващите данни D, може да се различава от истинската грешка $error_{\mathcal{D}}(h)$, дефинирана върху цялото вероятностното разпределение на примери $\mathcal{D}$. Нека с h_{best} да означим хипотезата от H, имаща най-малка обучаваща грешка върху обучаващите примери. Колко обучаващи примера са достатъчниq за да осигурим (с висока степен на вероятност), че истинската грешка на тази хипотеза $error_{\mathcal{D}}(h_{best})$ ще бъде не по-голяма от $\varepsilon + error_D(h_{best})$? Очевидно е, че въпросът, разгледан в предишния раздел, е само един специален случай на този въпрос, когато $error_D(h_{best}) = 0$.

За да се отговори на този въпрос е полезно да привлечем *общите граници на Хоефдинг* (Hoeffding bounds), които понякога се наричат *адитивни граници на Чернов* (additive Chernoff bounds). Границите на Хоефдинг характеризират отклонението между истинската вероятност на някое събитие и неговата наблюдаема честота на срещане при m независими опита. По-точно тези граници се прилагат при експерименти, включващи m различни опити на Бернули (т.е. m независими хвърления на една монета, която има една определена вероятност да падна по лице). Това е точна аналогия с

предположенията, които сме правили при изчисляване на грешка на една хипотеза в лекция 18: вероятността че монетата ще падна по лице съответства на вероятността, че хипотезата ще класифицира неправилно един случайно избран пример. *m* независими хвърляния на монетата съответстват на честотата на неправилното класифициране на *m* примера.

Границите на Хоефдинг постулират, че ако обучаващата грешка $error_D(h)$ се измерва върху множество *D*, съдържащо *m* случайно избрани примера, то:

$$Pr[error_{\mathcal{D}}(h) > error_D(h) + \varepsilon] \leq e^{-2m\varepsilon^2}$$

Това дава ни границата върху вероятността, че една произволно избрана хипотеза има една доста подвеждаща обучаваща грешка. За да осигурим, че *най-добрата* хипотеза, намерена от *L*, има грешка, ограничена по този начин, ние трябва да разгледаме вероятността, че някоя от |*H*| хипотезите може да има по-голяма грешка:

$$\Pr[(\exists h \in H)(error_{\mathcal{D}}(h) > error_D(h) + \varepsilon)] \leq |H| e^{-2m\varepsilon^2}$$

Ако наречем тази вероятност δ и запитаме, колко примера *m* са достатъчни, за да бъде равна δ на някаква желана стойност, ще получим следното:

$$m \geq \frac{1}{2\varepsilon^2}(\ln|H| + \ln(1/\delta)) \qquad (20.3)$$

Това представлява едно обобщение на неравенство (20.2) за случая, в който системата за самообучение избира най-добрата хипотеза $h \in H$, която може да има и ненулевата обучаваща грешка. Обърнете внимание, че *m* зависи логаритмически от |*H*| и от $1/\delta$, както това става и при по-ограничен случай на неравенство (20.2). Обаче в този по-малко ограничителен случай *m* е вече расте като квадрат от $1/\varepsilon$, а не линейно от $1/\varepsilon$, както в случай (20.2).

20.3.2. Конюнкции от Булеви литерали са PAC научаеми

И така, вече имаме границата, указваща броя на обучаващи примери достатъчни за вероятностно приближено научаване на целевото понятие. Можем да използваме нея за определяне на извадковата сложност и PAC научаемост на някои специфични класове от понятия.

Да разгледаме клас *C* от целеви понятия, които се описват чрез конюнкции от булеви (двоични) литерали. Един булев *литерал* е всякаква булева променлива (например *възрастен*) или нейно отрицание (например $\neg$ *възрастен*). По този начин конюнкциите от булеви литерали включват целевите понятия от типа на „*възрастен* $\wedge \neg$ *висок*”). Дали *C* е PAC научаем? Можем да покажем, че отговорът е „да” чрез показване че, първо, всяка „логична” система за самообучение ще изисква само един полиномиален брой обучаващи примери, за да научи произволно понятие *c* от *C*, а след това, чрез предлагане на един специфичен алгоритъм, който използва полиномно време за обработка на всеки обучаващ пример.

Да разгледаме една „логична" система за самообучение L, използваща пространство от хипотези H, идентично на C. Можем да използваме неравенство (20.2), за да изчислим броя m на случайно взети обучаващи примери, достатъчни, за да осигурим, че L ще изведе с вероятност $(1 - \delta)$ хипотеза с грешка, чийто максимум е равен на ε. За да направим това ние се нуждаем само от определяне на размера $|H|$ на пространството от хипотези.

Нека да разгледаме пространството на хипотези H, определено от конюнкции на литерали, базирани върху n булеви променливи. Размерът $|H|$ на такова пространство от хипотези е 3^n. За да се уверим в това трябва да разгледаме факта, че във всяка конкретна хипотеза има само три възможности за всяка променлива: да бъде включена променливата като литерал в хипотезата, да бъде включено нейното отрицание като литерал или променливата да бъде игнорирана. При условие, че имаме n такива променливи, има 3^n различни хипотези.

Заменяйки $|H| = 3^n$ в неравенство (20.2) получаваме следната граница за извадковата сложност на задачата за научаване на конюнкции от до n булеви функции:

$$m \geq \frac{1}{\varepsilon}(n \ln 3 + \ln(1/\delta)) \qquad (20.4)$$

Например, ако една „логична" система опитва да научи някое целево понятие, описано от конюнкции от до 10 булеви литерала, и искаме 95% вероятност, че тя ще научи хипотезата с грешка по-малка от 0.1, то на системата ще е достатъчно да бъдат предоставени m случайно избрани обучаващи примери, където

$$m = \frac{1}{0.1}(10 \ln 3 + \ln(1/0.05)) = 140.$$

Обърнете внимание, че m расте линейно по отношение на броя на литерали n и на $1/\varepsilon$, и логаритмически по $1/\delta$. А какъв е обемът от необходимите изчислителни ресурси? Естествено, че той ще зависи от конкретния алгоритъм за самообучение. Обаче докато такъв алгоритъм ще изисква не повече от полиномиален брой изчисления за един обучаващ пример и ще ползва не повече от полиномиален брой обучаващи примери, общият обем на необходимите изчисления ще бъде също полиномиален.

В случая на научаване на конюнкции от булеви литерали, един такъв алгоритъм, отговарящ на описаните по-горе изисквания, е вече ни известен от лекция 1. Това е алгоритъм FIND-S, който инкрементално изчислява най-специфичната хипотеза, съвместима с обучаващите примери. За всеки нов положителен пример този алгоритъм изчислява пресичането на литерали, общи за текущата хипотеза и нов обучаващ пример, използвайки време, което е линейно по отношение на n. По тази причина FIND-S PAC-научава клас от понятия, представени от конюнкции от n булеви литерали.

Теорема. PAC научаемост на булеви конюнкции. Клас C от конюнкции от булеви литерали е PAC научаем чрез алгоритъм FIND-S, използващ $H = C$.

Доказателство. Неравенство (20.4) показва, че извадковата сложност на този клас от понятия е полиномиален по отношение на n, $1/\delta$ и $1/\varepsilon$ и не зависи от размера на целевото понятие $size(c)$. За инкрементална обработка на всеки обучаващ пример, алгоритъм FIND-S се нуждае от изчисления, чийто брой е линеен по отношение на n и

независи от $1/\delta$ и $1/\varepsilon$ и *size(c)*. Следователно, този класс от понятия е PAC научаем от алгоритъма FIND-S.

20.3.3. PAC научаемост на други класове от понятия

Неравенство (20.2) дава една обща основа за намиране на граници на извадковата сложност за научавани целеви понятия за всеки определен клас *C*. По-горе ние приложихме го към класа от конюнкции на булеви литерали. Неравенството може да бъде използвано, за да се покаже, че и много други класове от понятия имат `полиномиална извадковата сложност.

„Безпристрастни" системи за самообучение

Не всички класове от понятия имат полиномиално ограничена извадковата сложност, определяна съгласно границата, задавана от формула (20.2). Например, да разгледаме „безпристрастния" клас от целеви понятия *C*, който съдържа *всички* научаеми понятия по отношение на пространство на примери *X* (т.е. класът от целеви понятия, за описание на които не се прилага индуктивното пристрастие – виж лекция 2). Този клас *C* от всички целеви понятия, които могат да бъдат дефинирани върху *X*, съответства на множество-степен на *X*, което съдържа $|C| = 2^{|X|}$ понятия. Да предположим, че примерите в *X* се дефинират чрез *n* булеви признака. В този случай, възможни са $|X| = 2^n$ различни примера и, следователно, $|C| = 2^{|X|} = 2^{2^n}$ различни понятия. Естествено е, че за да научи такъв „безпристрастен" клас от понятия една система за самообучение трябва също да използва „безпристрастното" пространство от хипотези $H = C$. Заменяйки $|H| = 2^{2^n}$ в формула (20.2) получаваме извадковата сложност за „безпристрастния" клас от понятия относително *X*:

$$m \geq \frac{1}{\varepsilon}(2^n \ln 2 + \ln(1/\delta)) \quad (20.5)$$

Следователно „безпристрастният" клас на целеви понятия има експоненциална извадкова сложност при използване на PAC модела, съгласно формула (20.5). Макар че формули (20.2) и (20.5) не задават тесни горни граници, на практика може да се докаже, че извадковата сложност за „безпристрастния" клас от целеви понятия е експоненциална по отношение на *n*.

k-терм ДНФ и k-КНФ понятия

Съществуват класове от понятия, които имат полиномиална извадкова сложност, но никога не могат да бъдат научени за полиномиално време. Един интересен пример на такъв клас е класът от изрази в дизюнктивна нормална форма, съставени от *k* терма (*k*-ДНФ). *k*-ДНФ изрази са изрази от вида на $T_1 \vee T_2 \vee \ldots \vee T_k$, където всеки терм T_i представлява конюнкция от *n* булеви атрибути или техни отрицания. Предполагайки $H = C$, лесно е да се покаже, че |H| е най-много $k3^n$ (тъй като има само *k* терма, умножени по 3^n възможности за всеки терм). Трябва да се отбележи, че $k3^n$ е една прекалено висока оценка за размера на пространството от хипотези, тъй като в нея по два пъти влизат случаи, когато $T_i = T_j$ и когато T_i *е_по-общо_от* T_j. Това не ни пречи да използваме тази горна граница на $|H|$, за да получим една горна граница на извадковата сложност, подставяйки този израз във формула (20.2):

$$m \geq \frac{1}{\varepsilon}(n \ln 3 + \ln k + \ln(1/\delta)) \quad (20.6)$$

Получената формула показва, че извадковата сложност на *k*-терм ДНФ е полиномиална по отношение на $1/\delta$ и $1/\varepsilon$, n и k. Макар че извадковата сложност е полиномиална, изчислителната сложност *не е* полиномиална, тъй като може да се покаже, че тази задача за обучение е еквивалентна на други задачи, за които е известно, че не могат да бъдат решени за полиномиално време. Така че, макар че *k*-терм ДНФ има полиномиална извадкова сложност, тя няма полиномиална изчислителна сложност за една система за самообучение, използваща $H = C$.

Един учудващ факт за *k*-терм ДНФ е, че макар този клас от целеви понятия не е PAC научаем, съществува един друг строго по-голям клас от понятия, който *е* PAC научаем! Това е възможно, тъй като по-големият клас от понятия има полиномиална изчислителна сложност за един пример, както и полиномиална извадкова сложност. Този по-голям клас е класът на изрази в конюктивно нормална форма, съставени от k терма, т.е. $T_1 \wedge T_2 \wedge \ldots \wedge T_k$, където всеки терм T_i представлява дизюнкция от до n булеви атрибути. Лесно е да се покаже, че *k*-КНФ изрази включват в себе си *k*-ДНФ, тъй като всеки *k*-КНФ израз може лесно да бъде преписан като *k*-ДНФ израз (но не е и обратно). Макар че *k*-КНФ изрази са по изразителни от *k*-ДНФ, те имат както полиномиална извадкова сложност, така и полиномиална времева сложност. Следователно клас понятия, описани като *k*-КНФ изрази, е PAC научаем от един ефективен алгоритъм, използващ $H = k$-КНФ (виж Kearns and Vazirani (1994) за повече подробности).