Лекция 9. Бейсово самообучение

9.1. Теоремата на Бейс

Често в МС ние се интересуваме от определяне на най-добрата хипотеза от някое пространство на хипотези *H* при зададени обучаващите данни *D*. Един от начини да определим, какво разбираме под думата "най-добра", е че това е *най-вероятната* хипотеза при зададени данни *D* и налични в момента знания за априорните вероятности на различни хипотези в *H*. Теоремата на Бейс дава директен метод за изчисляване на тези вероятности. По-точно тази теорема позволява да изчисляваме вероятността на една хипотеза въз основа на нейната априорна вероятност, вероятността да наблюдаваме тези данни при наличието на хипотезата и на вероятността на самите данни.

Означения:

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

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

P(D|h) означава вероятността да наблюдаваме данни *D* при условие, че хипотеза *h* е вярна. В общия случай *условната вероятност P(x|y)* означава вероятността за случване на някое събитие *x* при условие, че вече се случи събитие *y*.

В МС ние се интересуваме от 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 влияние на данни *D*, в отличие от априорната вероятност *P(h)*, която не зависи от 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):*

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)} \quad (9.1)$$

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 колко по-вероятно е, че *D* се наблюдават независимо от *h*, толкова е по-малка поддръжка дава *D* за съществуване на *h*.

В много задачи от МС ние 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-хипотези чрез използване на теоремата на Бейс:

$$h_{MAP} \equiv \arg\max_{h \in H} P(h \mid D) = \arg\max_{h \in H} \frac{P(D \mid h)P(h)}{P(D)} = \arg\max_{h \in H} P(D \mid h)P(h) \quad (9.2)$$

Ние отхвърлихме *P(D)* на последната стъпка, тъй като тя е постоянна и не зависи от *h*.

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

$$h_{ML} = \arg\max_{h \in h} P(D \mid h) \quad (9.3)$$

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

9.2. Пример

За илюстрация да разгледаме пример от медицинската диагностика, където имаме две алтернативни хипотези: 1) пациентът е болен от определена форма на рак и 2) пациентът е здрав. Разполагаме с данни от определен лабораторен тест с два възможни резултата – положителен (+) и отрицателен (-). Имаме априорни знания, че само 0.008 част от цялото население страда от това заболяване. Освен това знаем, че лабораторните тестове са само един приблизителен (неточен) индикатор за заболяването. Тестът връща коректен положителен отговор само в 98% от случаи, когато заболяването е действително налице и коректен отрицателен отговор само в 97% от случаите, кога това заболяване действително не присъства. Във всички други случаи тестът връща неправилен резултат. Тази ситуация може да бъде представена по следния начин:

$$P(\textit{рак}) = 0.008 \qquad P(\neg \textit{рак}) = 0.992$$

$$P(+ \mid \textit{рак}) = 0.98 \qquad P(- \mid \textit{рак}) = 0.02$$

$$P(+ \mid \neg \textit{рак}) = 0.03 \qquad P(- \mid \neg \textit{рак}) = 0.97$$

Да предположим, че при нас е постъпил новият пациент, чийто лабораторен тест има положителен резултат. Трябва ли ние да диагностицираме пациента като заболял от рак или не? Максималната апостериорна хипотеза ще се изчисли на базата на съответната формула (9.2):

$$P(+ \mid рак)P(рак) = 0.98 * 0.008 = 0.0078$$

$$P(+ \mid \neg рак)P(\neg рак) = 0.03 * 0.992 = 0.0298$$

Следователно, $h_{MAP} = \neg рак$. Можем и да изчислим точните вероятности чрез нормализиране на получените стойности по такъв начин, че те да бъдат заедно равни на 1 (например $P(рак|+) = 0.0078/(0.0078 + 0.0298) = 0.21$).

Примерът показва, че резултатът от Бейсовия извод силно зависи от априорните вероятности, които трябва да са известни за директно прилагане на метода. Обърнете внимание, в примера нито една от хипотезите не е напълно приета или отхвърлена – просто тя става по-малко или повече вероятна при условие, че станаха известни повече данни.

Основните формули за изчисляване на вероятности са показани долу:

- *Произведение на вероятностите*: вероятност $P(A \wedge B)$ на конюнкцията на два събития A и B

$$P(A \wedge B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

- *Сума на вероятностите*: вероятност $P(A \vee B)$ на дизюнкцията на два събития A и B

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

- *Теорема на Бейс*: апостериорната вероятност $P(h|D)$ на h при зададени D

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

- *Теорема за пълната вероятност*: ако събитията $A_1, ..., A_n$ са взаимно изключващи се и $\sum_{i=1}^{n} P(A_i) = 1$, то

$$P(B) = \sum_{i=1}^{n} P(B \mid A_i)P(A_i)$$

9.3. Директно прилагане на Бейсов подход към задачата за научаване на понятия

Да разгледаме отново вече въведената по-рано задача за научаване на понятия. Ще смятаме, че алгоритмът за обучение разглежда едно крайно пространство от хипотези H, определено върху пространство от примери X, като задачата му е да научи някое целево понятие $c: X \rightarrow \{0,1\}$. Както винаги предполагаме, че е зададена

някоя последователност от обучаващи примери $\{<x_1,d_1>,\ldots,<x_m,d_m>\}$, където x_i е някой пример от X, а d_i е стойността на целевата функция в x_i (т.е. $d_i = c(x_i)$). Да предположим също, че последователността на примери $\{x_1,\ldots,x_m\}$ е фиксирана така, че обучаващите данни D могат да се запишат само като последователността от стойностите на целевия атрибут: $D = \{d_1,\ldots,d_m\}$.

Можем да конструираме алгоритъм за научаване на понятия, който извежда максималната апостериорна хипотеза, базирайки се на теоремата на Бейс:

Алгоритъм за директно научаване на MAP-хипотеза

1. За всяка хипотеза h в H, изчисли апостериорната ѝ вероятност:

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

2. Върни като изход хипотезата h_{MAP} с най-висока апостериорна вероятност:

$$h_{MAP} = \arg\max_{h \in h} P(h \mid D)$$

Този алгоритъм изисква значителни изчисления, тъй като, за да изчисли $P(h|D)$ той прилага теоремата на Бейс към всяка хипотеза от H. Макар, че това може да бъде практически неприложимо за големи пространства на хипотези, алгоритмът е интересен с това, че предлага един общ стандарт, който може да се ползва за оценяване на поведение на други алгоритми за научаване на понятия.

За да определим задачата за обучение в описания по-горе алгоритъм, трябва да определим стойностите, които ще се ползват за $P(h)$ и $P(D|h)$. Нека да изберем ги в съгласие със следните предположения:

1. Обучаващите данни D не са зашумени (т.е. $d_i = c(x_i)$).
2. Целевото понятие c се съдържа в пространството на хипотези H.
3. Нямаме никакви априорни основания да вярваме, че някаква хипотеза е по-вероятна от друга.

Как при тези предположения да зададем $P(h)$? Съгласно третото предположение, априорните вероятности на всички хипотези в H са еднакви. Освен това, тъй като съгласно второто предположение, истинското описание на понятие се съдържа в H, сумата на априорните вероятности на всички хипотези в H трябва да е 1. Следователно:

$$P(h) = \frac{1}{|H|} \quad \text{за } \forall \text{h} \in \text{H}$$

А как да изберем $P(D|h)$? $P(D|h)$ е вероятността да наблюдаваме стойностите на целевата функция $D = \{d_1,...,d_m\}$ за фиксираното множество от примери $\{x_1,...,x_m\}$, при условие, че хипотезата h е изпълнена (т.е. в един свят, където h е едно коректно описание на целевото понятие c). Тъй като смятаме, че обучаващите данни са коректни (без шум), то вероятността да наблюдаваме класификация d_i при зададена h е точно *1*, ако $d_i = h(x_i)$ и *0*, ако $d_i \neq h(x_i)$. Следователно:

$$P(D \mid h) = \begin{cases} 1, \text{ ако } d_i = h(x_i) \text{ за } \forall d_i \in D \\ 0, \text{ в други случаи} \end{cases} \qquad (9.4)$$

С други думи, вероятността на данни D при зададена хипотеза h е 1, ако D не са противоречат (съвместими с) на h, или 0 в противен случай.

И така, връщайки се към теоремата на Бейс (9.1) за случая, когато хипотезата h е несъвместима с данни D, от (9.4) следва, че $P(h|D) = 0$.

Сега да разгледаме случая, когато h е съвместима с D. Преди това ще опитаме да оценим $P(D)$, използвайки теоремата за общата вероятност на пълното събитие и факта, че всички хипотези са взаимно изключващи се, т.е. $(\forall i \neq j)(P(h_i \wedge h_j) = 0)$:

$$P(D) = \sum_{h_i \in H} P(D \mid h_i)P(h_i) = \sum_{h_i \in VS_{H,D}} 1 * \frac{1}{|H|} + \sum_{h_i \notin VS_{H,D}} 0 * \frac{1}{|H|} = \sum_{h_i \in VS_{H,D}} 1 * \frac{1}{|H|} = \frac{|V_{H,D}|}{|H|} \qquad (9.5)$$

От тук следва:

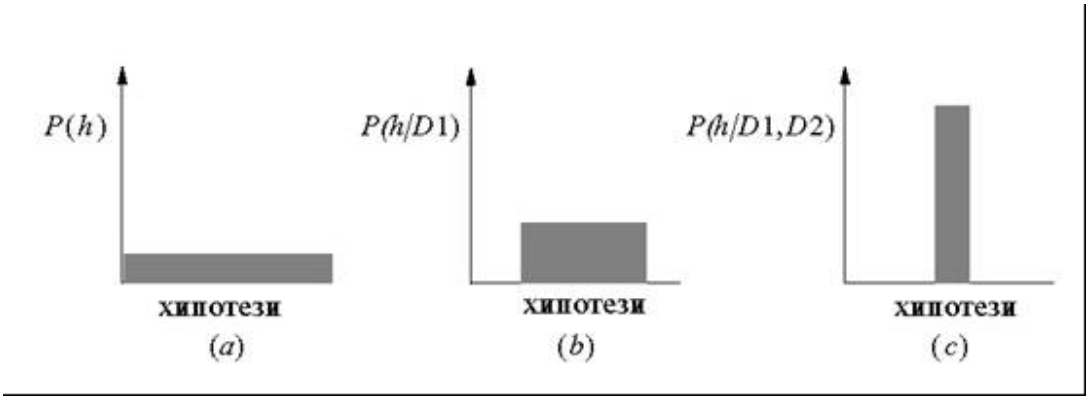
$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)} = \frac{1 * \frac{1}{|H|}}{\frac{|VS_{H,D}|}{|H|}} = \frac{1}{|VS_{H,D}|}$$

И така, теоремата на Бейс извежда, че апостериорната вероятност $P(h|D)$ при предположения приети за $P(h)$ и $P(D|h)$ е:

$$P(h \mid D) = \begin{cases} \frac{1}{|VS_{H,D}|}, \text{ ако } h \text{ е съвместима с } D \\ 0, \quad \text{ако } h \text{ е несъвместима с } D \end{cases} \qquad (9.6)$$

където $|VS_{H,D}|$ е броя на хипотези в H, съвместими с D, т.е. е размер на пространството на версиите. Еволюцията на вероятности, асоциирани с хипотези, е схематично представена на следната рисунка.

Първоначално (a) всички хипотези имат една и съща априорна вероятност. След нарастване на обема на данни до $D1$ (b), а след това и до $D1 \wedge D2$, апостериорните вероятности на несъвместими хипотези стават нула, докато апостериорните вероятности на хипотезите, останали в пространството на версии, нарастват.



Следователно, при направения по-горе избор на *P(h)* и *P(D|h)* всяка съвместима с налични данни хипотеза има апостериорната вероятност $(1/|VS_{H,D}|)$, докато всяка несъвместима – 0.

9.4. MAP-хипотези и "логични" алгоритми за обучение

Проведеният по-горе анализ показа, че при приетите вероятностни предположения всяка съвместима с данни хипотеза е MAP-хипотеза. Този извод може директно да се приложи към един общ клас алгоритми за обучение, наречени "логични". Ще казваме, че един алгоритъм за обучение е *логичен*, ако той извежда като изход хипотеза, която класифицира без никакви грешки всички обучаващи примери. Нашият анализ показа, че *всеки логичен алгоритъм за обучение извежда MAP-хипотеза при предположение за еднакво (равномерно) разпределение на вероятностите на хипотезите в H (т.е. $P(h_i) = P(h_j)$ за всички i, j), и че данните са детерминирани и са без грешки (т.е. P(D|h) = 1, ако D и h са съвместими, и 0 – в противен случай)*.

Да разгледаме, като пример, алгоритъм за научаване на понятия FIND-S, разгледан в лекция 1. FIND-S претърсва пространството на хипотези от частното към общото, извеждайки максимално специфична хипотеза, съвместима с данни (т.е. максимално специфичен член на пространството на версиите). Тъй като алгоритмът извежда една съвместима с данни хипотеза, тя ще бъде MAP-хипотеза при описани по-горе разпределения на вероятности на *P(h)* и *P(D|h)*. Ясно е, че FIND-S не работи в явен вид с вероятностите, обаче изясняване на характера на разпределение на вероятности на *P(h)* и *P(D|h)* дава удобен начин за описание на поведение на алгоритъма.

С други думи Бейсовият подход позволява да анализира поведение на различни алгоритми за обучение, дори когато те не работят в явен вид с вероятностите. Чрез идентификация на вероятностните разпределения на *P(h)* и *P(D|h)*, при които алгоритмът извежда оптималните (т.е. MAP) хипотези, можем да разберем неявни предположения, при наличие на които алгоритмът ще работи оптимално.

От тази гледна точка, използването на Бейсовия подход е близко по духа с изясняването на индуктивното пристрастие на алгоритъма. Да напомня, че ние дефинирахме индуктивното пристрастие на един алгоритъм за обучение като множеството от предположения B, достатъчни за дедуктивно обяснение на направения от алгоритъма индуктивен извод. Например, описахме индуктивното пристрастие на алгоритъма за елиминиране на кандидати като предположение, че целевото понятие се съдържа в пространството на хипотези H. Показахме, че изходът на този индуктивен алгоритъм дедуктивно следва от неговите входове плюс това неявно предположение. Бейсовата интерпретация предоставя един алтернативен начин за характеризиране на използваните в индуктивни алгоритми неявни предположения. Сега вместо да моделираме индуктивния извод чрез една еквивалентна дедуктивна система, ние ще го моделираме чрез една еквивалентна система за *вероятностни разсъждения*, базираща се на теоремата на Бейс. При този подход неявните предположения, върху които се базира работата на индуктивния алгоритъм, приемат вид "априорните вероятности на H са зададени от разпределение $P(h)$, а влиянието на данни за отхвърляне или приемане на една хипотеза се задава чрез $P(D|h)$".

Една система за вероятностен извод, базирана на теоремата на Бейс, ще има същото входно-изходно поведение както и индуктивните алгоритми, ако на нея са зададени тези предполагаеми разпределения на вероятностите.

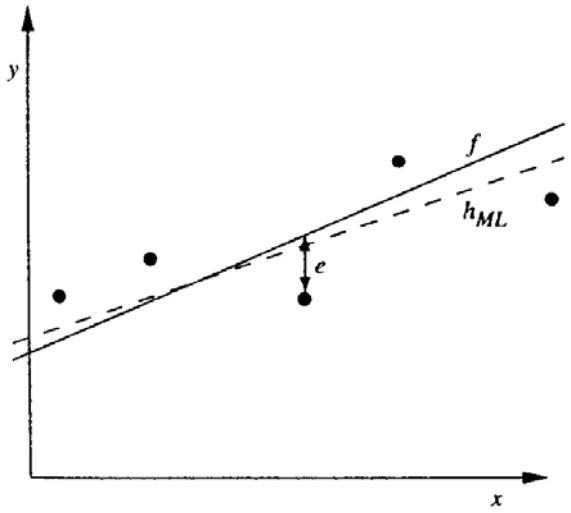
9.5. Максимално възможна хипотеза и хипотеза, минимизираща квадрата на грешката

В предишния раздел показахме, че Бейсовият анализ може да се използва, за да покажем, че един алгоритъм за обучение може да извежда MAP-хипотези дори, ако той не използва явно Бейсовото правило или да не изчислява изобщо вероятностите.

В този раздел ще разгледаме задачата за научаване на непрекъсната целева функция – задача, решавана от множество различни подходи за обучение, между които е обучението чрез невронни мрежи, линейната регресия и т.н. С помощта на Бейсовия анализ ние ще покажем, че *при определени предположения всеки алгоритъм за обучение, който минимизира средноквадратичната грешка между предсказването на изходната хипотеза и обучаващите данни, дава като изход максимално възможна хипотеза*. Важността на този резултат е в това, че той предоставя Бейсовото обоснование на методи, използвани от невронните мрежи и др., опитващи се да намерят оптималното решение чрез минимизиране на сумата от квадрати на грешката върху обучаващи данни.

Да разгледаме следната постановка на задачата. Системата за самообучение L разглежда някое пространство от примери X и пространство от хипотези H, състоящо от определен клас непрекъснати функции, определени на X (т.е. всяка h от H е функцията във вида $h : X \to \Re$, където $\Re$ представлява множеството от реални числа). Задачата пред L е да научи неизвестна целева функция $f : X \to \Re$, избрана от H. На системата е предоставено множество от m обучаващи примера, в които стойността на целевия атрибут е изкривена от случаен шум, който има нормално разпределение на вероятността. По-точно, всеки обучаващ пример е двойката от вида $< x_i, d_i >$, където $d_i = f(x_i) + e_i$. Тука $f(x_i)$ е свободна от шума стойност на целевата функция, а e_i е случайна величина, представяща шум. Предполага се, че стойностите на e_i се постъпват независимо една от друга и са разпределени по нормален закон със средната стойност нула. Задачата на системата е да изведе максимално възможна хипотеза или, еквивалентно, максимална апостериорна (MAP) хипотеза, предполагайки, че всички хипотези са равновероятни априори.

Един прост пример на такава задача е научаване на една линейна функция. На рисунката по-долу линейната целева функция f е показана с плътна линия, а множеството от зашумени обучаващи примери на тази функция – с точки. Пунктираната линия съответства на хипотеза, минимизираща сума от квадрати на грешките в тези данни, т.е. максимално възможна хипотеза h_{ML}. Обърнете внимание, че максимално възможната хипотеза не винаги съвпада с правилната хипотеза f, тъй като тя е изведена само от една ограничена извадка от зашумени обучаващи данни.

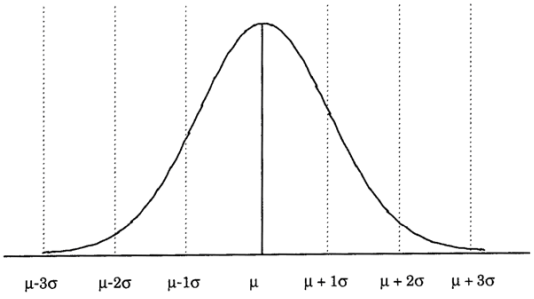


Преди да докажем, че при тази постановка на задачата хипотезата, минимизираща сумата от квадрати на грешките върху обучаващи данни, е също така и максимално възможната хипотеза, нека съвсем накратко да припомним две основни понятия от теорията на вероятностите – плътността на вероятност и нормалното разпределение. Тъй като ще разглеждаме вероятността на непрекъснати случайни величини от типа на e_i, трябва да дефинираме понятие за плътността на вероятност. Основната причина за това е, че искаме сумата на вероятности на всички възможни стойности на случайната величина да бъде равна на 1. В случая на непрекъсната величина това не може да бъде постигнато чрез назначаване на някоя крайна стойност на вероятността на всяка една от безкрайното множество от възможни значения на случайната величина. По тази причина ще говорим за *плътността на вероятност* на непрекъснати случайни величини и ще изискваме интеграл от тази плътност на вероятност по всички възможни стойности да е равен на 1. Плътността на вероятност $p(x_0)$ се определя като:

$$p(x_0) \equiv \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} P(x_0 \le x < x_0 + \varepsilon)$$

Нормалното разпределение е една гладка камбановидна функция, която напълно се описва чрез своята средна стойност μ и стандартното отклонение σ:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Сега, използвайки тези понятия, да покажем, че хипотезата, минимизираща сумата на квадрати на грешките върху обучаващите данни, е максимално възможна хипотеза. Да започнем с вече въведеното по-рано определение за максимално възможна хипотеза, обаче заменяйки вероятността с плътност на вероятност:

$$h_{ML} = \arg\max_{h \in h} p(D \mid h)$$

Както винаги, предполагайки, че множеството от обучаващи примери $<x_1,...,x_m>$ е фиксирано, можем да разглеждаме D като $D = <d_1,...,d_m>$. Предполагайки, че обучаващите примери са взаимно независими при зададена h, можем да препишем $p(D|h)$ като произведение от различните $p(d_i|h)$:

$$h_{ML} = \arg\max_{h \in H} \prod_{i=1}^{m} p(d_i \mid h)$$

Тъй като по предположение шумът e_i има нормалното разпределение със средното нула и неизвестната дисперсия σ^2, то и всяка d_i също ще се подчинява на нормалното разпределение със същата дисперсия, но със средното $\mu = f(x_i)$. Следователно, същото важи и за $p(d_i|h)$, а тъй като смятаме, че h е коректно описание на целевата функция f, можем да заменим $\mu = f(x_i) = h(x_i)$. Получаваме:

$$h_{ML} = \arg\max_{h \in H} \prod_{i=1}^{m} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(d_i-\mu)^2}{2\sigma^2}} = \arg\max_{h \in H} \prod_{i=1}^{m} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(d_i-h(x_i))^2}{2\sigma^2}}$$

Сега приложим една трансформация (взимане на логаритъма), която е общо приета при работата с нормалното разпределение. Тя е приложима, тъй като $\ln p$ е монотонна функция от p, така че максимизиране на $\ln p$ също така ще максимизира и p:

$$h_{ML} = \arg\max_{h \in H} \left(\sum_{i=1}^{m} \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}(d_i - h(x_i))^2\right)$$

Първия член на този израз е константа и не зависи от h, следователно може да бъде пренебрегнат. Получаваме:

$$h_{ML} = \arg\max_{h \in H} \sum_{i=1}^{m} -\frac{1}{2\sigma^2}(d_i - h(x_i))^2 = \arg\min_{h \in H} \sum_{i=1}^{m} \frac{1}{2\sigma^2}(d_i - h(x_i))^2$$

И на края, отново премахвайки константа, независеща от h, получаваме:

$$h_{ML} = \arg\min_{h \in H} \sum_{i=1}^{m} (d_i - h(x_i))^2$$

И така, ние доказахме, че максимално възможната хипотеза h_{ML} е тази, която минимизира сума от квадрати на грешки между наблюдаваните обучаващи стойности d_i и предсказанията на хипотезата $h(x_i)$. Това твърдение е вярно при предположение, че наблюдаваните обучаващи стойности d_i са генерирани с добавяне на шума към истинските стойности на целевата функция, като този случаен шум се взима независимо за всеки пример от нормалното разпределение с нулева средна стойност. Дали е разумно да приемаме това допускане? Да, защото освен това, че позволява по-лесен математически анализ, нормалното разпределение е едно добро приближение към различни типове шум, наблюдавани във физически системи.

Едно много важно ограничение на използваната от нас постановка на задачата е, че ние допускаме шум само в стойностите на *целевия атрибут* на обучаващите

примери и *не допускаме наличие на шума в атрибутните стойности,* описващи самите обучаващи примери.

9.6. Принцип за минималната дължина на описание

Да припомним, че в лекция 2 ние обсъждахме едно популярно индуктивно пристрастие, наречено "принцип на Окам", което гласеше: "Избирай най-късото обяснение за наблюдаваните данни". Сега ще добавим няколко аргумента към все още продължаващ се дебат за валидността на този принцип. Ще разгледаме него от гледна точна на Бейсовия извод, както и един друг - много свързан с него принцип, наречен принцип за минималната дължина на описание (MDL).

MDL-принципът е мотивиран от интерпретацията на определението за MAP-хипотеза от гледната точка на теорията на информация. Да се върнем към вече известната дефиниция:

$$h_{MAP} = \arg\max_{h \in H} P(D \mid h)P(h)$$

която може еквивалентно да се препише в термини на максимизирането на log_2

$$h_{MAP} = \arg\max_{h \in H} (\log_2 P(D \mid h) + \log_2 P(h))$$

или алтернативно, в термините на минимизацията:

$$h_{MAP} = \arg\min_{h \in H} (-\log_2 P(D \mid h) - \log_2 P(h)) \quad (9.7)$$

Полученото утвърждение може да се интерпретира като утвърждение, че късите хипотези са за предпочитане, ако се подразбира определена схема за кодиране на хипотези и данни. За да разберем това, трябва да се върнем за малко към теорията на информация.

Да разгледаме задачата за създаване на кода за предаване на съобщения, взети по случаен начин, където вероятността за срещане на съобщение i е p_i. Искаме да създадем най-компактния код, т.е. такъв код, който минимизира очаквания брой битове, които трябва да изпратим за кодиране на някое случайно избрано съобщение. Ясно е, че за да минимизираме тази очаквана дължина на кода, трябва да назначаваме по-къси кодове на тези съобщения, които са по-вероятни. Shannon и Weaver (1949) показаха, че оптималният код (т.е. код, минимизиращ очакваната дължина на съобщение) използва $-log_2 p_i$ битове за кодиране на съобщение i. Ще използваме означение $L_C(i)$ за запис на *дължината на описание на съобщение i с помощта на кода C*.

В термините на теорията на информация, ф-ла (9.7) може да се интерпретира по следния начин:

- $-\log_2 P(h)$ е дължината на описанието на h при използване на оптималното кодиране на пространството на хипотези H. С други думи, това е размер на описанието на хипотезата h при използване на оптималното представяне, т.е. $L_{C_H}(h) = -\log_2 P(h)$, където C_H е оптимален код за описание на пространството на хипотези H.
- $-\log_2 P(D|h)$ е дължината на описанието на обучаващите данни D при зададената хипотеза h и използване на нейното оптимално кодиране, т.е. $L_{C_{D|h}}(D|h) = -\log_2 P(D|h)$, където $C_{D|h}$ е оптималният код за описание на данни D при предположението, че и двамата – и изпращач, и получател, знаят хипотезата h.

Следователно, можем да препишем ф-ла (9.7) за MAP-хипотеза във вид:

$$h_{MAP} = \arg\min_{h \in H}(L_{C_H}(h) + L_{C_{D|h}}(D \mid h))$$

MDL-принципът предлага избиране на такава хипотеза, която минимизира сумата на тези две дължини на описания. За да го използваме на практика, трябва да изберем конкретни кодове C_1 и C_2 за представяне на хипотези и представяне на данни при наличието на тази хипотеза. Ако сме го направили, то MDL-принципът гласи:

Избери h_{MDL} *така, че* $h_{MDL} = \arg\min_{h}(L_{C_1}(h) + L_{C_2}(D \mid h))$ (9.8)

Нашият анализ показа, че ако избрахме C_1 да бъде оптималното кодиране на хипотези, а C_2 – оптималното кодиране на данни чрез хипотези, то $h_{MDL} = h_{MAP}$.

Интуитивно можем да мислим за MDL-принцип като за рекомендацията за избор на най-кратък начин за прекодиране на обучаващите данни, имайки предвид както размер на хипотезата, така и цената за допълнителното кодиране на данни при тази хипотеза.

Да разгледаме като илюстрация, как MDL-принципът може да се приложи към задачата за научаване на класификационни дървета от обучаващите данни. По-точно, към задачата за допълнителното подрязване на дървета, която може да се разглежда като търсене на оптималното (в смисъла на MDL-принципа) класификационно дърво в множеството от класификационни дървета H, покриващи обучаващите данни D. Първо, трябва да изберем начина на кодиране на хипотези – класификационни дървета (C_1). Очевидно е, че дължината на кодирането на едно дърво зависи от броя на възли и дъги в него. Например, като такава мярка може да се използва дължината на всички пътищата от корена до всяко листо на дървото (т.е. броят на всички тестове в системата от правила, в които може да бъде транслирано дървото):

$L_{C_1}(h) = \sum_i d_i$, където d_i е дълбочина на листото i в дървото h.

По-сложен е въпросът, как до оценим сложността на кодиране на данни при зададено класификационно дърво (C_2). Един от възможните подходи е използване на ентропията. Съществуват различни конкретни формули за тази цел, като най-

често се използва формулата, отчитаща разпределението на примери, покрити от конкретна хипотеза:

$$L_{C_2}(D \mid h) = \sum_i n_i Entropy_i(h), \text{ където}$$

$$Entropy_i(h) = -\frac{n_i(+)}{n_i}\log_2\frac{n_i(+)}{n_i} - \frac{n_i - n_i(+)}{n_i}\log_2\frac{n_i - n_i(+)}{n_i}$$

n_i е броят на примери, покрити от листото i, а $n_i(+)$ – броят на положителни примери, покрити от листото i.

По този начин MDL-принципът може да доведе до предпочитане на едно по-къса хипотеза, която прави няколко грешки, пред една по-дълга (по-сложна) хипотеза, която перфектно покрива обучаващи данни. От тази гледна точка принципът за минималната дължина на описание третира дискутирания от нас по-рано въпрос за прекалено нагаждане към данни.

Дали MDL-принципът доказва един път за винаги, че по-късите хипотези са по-добри? Не, единствено, което ние показахме, е, че *ако* представянето на хипотезите е избрано по такъв начин, че размерът на хипотезата h е $-\log_2 P(h)$, и че *ако* избрахме такова представяне на изключенията, че дължината на кодирането на D при избраното h е $-\log_2 P(D|h)$, *то тогава* MDL-принципът избира хипотезите с най-голяма апостериорна вероятност. Обаче, за да покажем, че ние избрахме точно такова представяне, трябва да знаем както априорните вероятности $P(h)$, така и на $P(D|h)$. Няма никакви причини да вярваме, че избраната от MDL-принципа хипотеза е най-добрата при *произволни* начини за кодиране C_1 и C_2. По тази причина в литература за практическото прилагане на този принцип често присъстват аргументи, опитващи се да доказват правилността на избора на кодиранията C_1 и C_2.

Резюме

- Бейсовите методи осигуряват основата за методите на вероятностно обучение, които използват (и се нуждаят от) знания за априорните вероятности на алтернативни хипотези и за вероятности от наблюдаване на различни данни при зададени хипотези. Бейсовите методи позволяват назначаване на определени апостериорни вероятности на всяка хипотеза, базирайки се на предполагаемите априорните вероятности и наблюдаваните данни.
- Бейсовите методи могат да се ползват за определяне на най-вероятната хипотеза при зададените данни - максимална апостериорна (MAP) хипотеза. Тя е оптималната хипотеза в смисъл, че не съществува никоя друга хипотеза, която е по-вероятна от нея.
- Принципът за минимална дължина на описание предлага избор на хипотезата, която минимизира дължината на описание на хипотезата плюс дължината на описание на данни при зададена хипотеза. Теоремата на Бейс и основните резултати от теорията на информация могат да се ползват като обосновка за този избор.