Bayesovské učení

Bayesovské učení přistupuje k inferenci z pravděpodobnostních pozic. Je založeno na předpokladu, že kvantity, jimiž se zabývá, splňují předpoklady pravděpodobnostních rozložení a že optimální rozhodnutí mohou být dosažena prostřednictvím usuzování nad pozorovanými údaji a jejich příslušnými pravděpodobnostmi.

Strojové učení poskytuje kvantitativní přístup ke zvažování důkazů podporujících alternativní hypotézy. Proto je Bayesovské učení důležité pro strojové učení jako disciplínu.

Bayesovské učení poskytuje základ pro učící algoritmy jenž přímo manipulují s pravděpodobnostmi.

Mezi nejúčinnější metody patří tzv. naivní Bayesovské učení, které pro některé typy problémů poskytuje jednu z nejefektivnějších metod.

$f_1, f_2, \dots f_k$

$p_1 = \frac{f_1}{N}, p_2 = \frac{f_2}{N}, \dots$

Vlastnosti Bayesovského učení

- Každé pozorování (trénovací příklad) může inkrementálně snížit či zvýšit odhadovanou pravděpodobnost, že hypotéza je korektní. Tím je poskytován flexibilnější přístup k učení než v algoritmech, které hypotézu zcela eliminují pokud je zjištěno, že je nekonzistentní s libovolným jediným příkladem.

- Lze kombinovat předchozí znalost s pozorovanými údaji k určení konečné pravděpodobnosti hypotézy. V Bayesovském učení se předchozí znalost používá k prosazení
  1) apriorní pravděpodobnosti pro každou možnou hypotézu,
  2) rozložení pravděpodobnosti pro pozorovaná data pro každou možnou hypotézu.

- Bayesovské metody mohou zpracovávat hypotézy, které se používají pro pravděpodobnostní predikce (např. hypotéza, že „tento pacient s pneumonií má 93% šanci se zcela uzdravit").

- Nové instance lze klasifikovat kombinováním predikcí více hypotéz váhovaných svými pravděpodobnostmi.

- I v případech, kdy jsou Bayesovské metody výpočetně neproveditelné, mohou poskytnout standard pro optimální rozhodování, vůči němuž lze porovnávat ostatní praktické metody.

S Bayesovskými metodami je spjat jeden praktický problém: tyto metody typicky vyžadují nějakou počáteční znalost mnoha pravděpodobností. Pokud tyto pravděpodobnosti nejsou předem známy, tak jsou často odhadovány pomocí nějaké výchozí (základní) znalosti, nebo pomocí dříve získaných dat, či předpokladů o formě pravděpodobnostního rozložení.

Další praktická potíž spočívá ve výrazně náročných výpočetních požadavcích k určení optimální hypotézy (v obecném případě) – složitost je lineární s růstem alternativních hypotéz. V určitých specializovaných situacích však lze výpočetní nároky významně redukovat.

Bayesův teorém

Ve strojovém učení se často zajímáme o stanovení nejlepší hypotézy v nějakém prostoru H za předpokladu, že jsou k dispozici pozorované údaje D pro trénování.

Jednou z cest je požadovat nejpravděpodobnější hypotézu pomocí D a nějaké výchozí znalosti o apriorních pravděpodobnostech různých hypotéz v H. Bayesův teorém poskytuje přímou metodu, jak tyto pravděpodobnosti vypočítat.

Jinými slovy, Bayesův teorém poskytuje způsob, jak určit pravděpodobnost hypotézy pomocí její apriorní pravděpodobnosti, dále pravděpodobností pozorování různých dat za předpokladu, že je dána tato hypotéza, a konečně pomocí pozorovaných dat.

Nechť platí následující označení:

$P(h)$ ... počáteční pravděpodobnost, že hypotéza h platí před tím, než jsme získali pozorování (trénovací data). $P(h)$ se často nazývá apriorní pravděpodobnost hypotézy h a může např. vyjadřovat jakoukoliv předchozí (výchozí) znalost, kterou máme o šanci, že h je správná hypotéza.
Pokud takovou znalost nemáme, pak nezbývá, než jednoduše předpokládat, že každá z možných hypotéz je stejně pravděpodobná (nemusí to být pravda, ale co se dá dělat).

$P(D)$ ... počáteční pravděpodobnost, že budou pozorovány údaje D, tj. pravděpodobnost výskytu D bez znalosti o pravděpodobnosti jakékoliv hypotézy.

$P(D|h)$ ... pravděpodobnost, že zpozorujeme data D je-li dán nějaký „svět" v němž platí hypotéza h.
Obecně $P(x|y)$ označuje pravděpodobnost x za předpokladu y.

Ve strojovém učení nás zajímá problém, jak zjistit pravděpodobnost $P(h|D)$, že platí h máme-li k dispozici pozorovaná trénovací data D.

$P(h|D)$ se nazývá posteriorní pravděpodobnost h, neboť odráží naši důvěru v to, že h platí po tě, co jsme pozorovali trénovací data D.

Za povšimnutí stojí, že posteriorní pravděpodobnost $P(h|D)$ odráží vliv trénovacích dat na platnost hypotézy h, zatímco apriorní pravděpodobnost $P(h)$ je na datech D zcela nezávislá!

Bayesův teorém tvoří základ Bayesovských metod učení, protože poskytuje přímou cestu ke stanovení posteriorní pravděpodobnosti $P(h|D)$ pomocí apriorní pravděpodobnosti $P(h)$ spolu s $P(D)$ a $P(D|h)$:

$$P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)} \qquad (1)$$

(Zde se nebudeme zabývat důkazem; viz příslušnou stať z pravděpodobnostního počtu.)

Intuitivně očekáváme, že $P(h|D)$ roste s $P(h)$ a s $P(D|h)$. Je rovněž rozumné očekávat, že $P(h|D)$ klesá s růstem $P(D)$, protože čím je pravděpodobnější, že D pozorujeme nezávisle na h, tím menší podporu (jako důkaz) hypotéze h poskytují data D.

Ve mnoha „učících situacích" zvažuje „žák" nějaký soubor (množinu) možných hypotéz H a zajímá se o nalezení nejpravděpodobnější $h \in H$, přičemž usuzuje na základě pozorovaných dat D.

Každá z maximálně pravděpodobných hypotéz se nazývá maximální aposteriorní hypotéza (MAP).

Můžeme stanovit MAP pomocí Bayesova teorému tak, že spočítáme posteriorní pravděpodobnosti pro všechny uvažované hypotézy.

Říkáme, že h_{MAP} je hypotéza MAP platí-li:

$$h_{MAP} \equiv \underset{h \in H}{\operatorname{argmax}}\ P(h|D) =$$

$$= \underset{h \in H}{\operatorname{argmax}}\ \frac{P(D|h) \cdot P(h)}{P(D)} = \qquad (2)$$

$$= \underset{h \in H}{\operatorname{argmax}}\ P(D|h) \cdot P(h)$$

V posledním kroku byl vynechán člen $P(D)$, protože je to konstanta nezávislá na h.

V některých případech předpokládáme, že každá hypotéza z H má stejnou apriorní pravděpodobnost $P(h_i) = P(h_j)$, $\forall h_i, \forall h_j \in H$. V tom případě lze vztah pro h_{MAP} zjednodušit a lze uvažovat pouze člen $P(D|h)$ k nalezení nejpravděpodobnější hypotézy.

$P(D|h)$ se často nazývá jako „slibnost" („vhodnost") dat D za předpokladu h.

Každá hypotéza, jež maximalizuje $P(D|h)$ se nazývá maximálně slibná (vhodná) (ML) hypotézou h_{ML}:

$$h_{ML} \equiv \underset{h \in H}{\operatorname{argmax}}\, P(D|h)$$

Pozn.: Zde se díváme na Bayesův teorém z praktického hlediska disciplíny strojového učení. Bayesův teorém je ve skutečnosti daleko obecnější nástroj, jenž může být aplikován na libovolnou množinu vzájemně výlučných propozic z H, jejichž pravděpodobnosti dávají v součtu 1 (např. „nebe je modré" a „nebe není modré").

Příklad

Uvažme medicínský diagnostický problém, v němž jsou pro jednoduchost dvě alternativní hypotézy:

1) pacient má konkrétní formu rakoviny
2) pacient ji nemá

Údaje potřebné k posouzení typu nemoci pocházejí z laboratoře, přičemž testy (vyšetření) mají dva možné výsledky: $\oplus$ (pozitivní) a $\ominus$ (negativní).

Nějme dále k dispozici apriorní znalost o celé populaci, že 0.008 (0.8 %) má tuto nemoc.

Dále víme, že laboratorní testy mají jen omezenou spolehlivost < 100 %:

> test vrací korektní výsledek $\oplus$ pro 98 % případů, v nichž pacient skutečně nemoc měl, a korektní výsledek $\ominus$ pro 97 % případů, v nichž pacient skutečně nemoc neměl (tj. v ostatních případech dával test opačný výsledek).

Popsaný problém lze sumarizovat následovně:

$$P(\text{rakovina}) = 0.008 \implies P(\neg\text{rakovina}) = 0.992$$

$$P(\oplus|\text{rakovina}) = 0.98 \implies P(\ominus|\text{rakovina}) = 0.02$$

$$P(\oplus|\neg\text{rakovina}) = 0.03 \impliedby P(\ominus|\neg\text{rakovina}) = 0.97$$

Nyní předpokl., že máme nového pacienta, který obdržel z laboratoře pozitivní výsledek. Jaká má být diagnóza – má či nemá rakovinu? h_{MAP} nalezneme pomocí vztahu (2): $P(D|h) \cdot P(h)$

$$P(\oplus|\text{rakovina})\,P(\text{rakovina}) = 0.98 \cdot 0.008 = 0.0078$$
$$P(\oplus|\neg\text{rakovina})\,P(\neg\text{rakovina}) = 0.03 \cdot 0.992 = 0.0298 \quad \uparrow \text{argmax}$$

Tedy $h_{MAP} = \neg\text{rakovina}$.

Pozn.:

Přesné posteriorní pravděpodobnosti mohou být stanoveny normalizací tak, aby jejich součet = 1:

např. $P(\text{rakovina}|\oplus) = \frac{0.0078}{0.0078+0.0298} = 0.21$

Tento krok je zaručen, protože Bayesův teorém říká, že posteriorní pravděpodobnosti jsou právě ty výše uvedené hodnoty dělené pravděpodobnostmi dat, tj. $P(\oplus)$. Ačkoliv $P(\oplus)$ nebyla poskytnuta přímo jako součást popisu problému, jsme schopni ji spočítat tímto způsobem, neboť víme, že $P(\text{rakovina}|\oplus)$ a $P(\neg\text{rakovina}|\oplus)$ musí dohromady dát 1 (tj. pacient s pozitivním testem buď rakovinu má nebo ji nemá).

Za povšimnutí stojí, že zatímco posteriorní pravděpodobnost rakoviny je výrazně vyšší než apriorní, nejpravděpodobnější hypotéza praví, že pacient rakovinu nemá.

Uvedený příklad svědčí, že výsledek Bayesovské inference závisí velmi silně na apriorních pravděpodobnostech, které musí být k dispozici, aby bylo možno metodu aplikovat přímo.

Důležité rovněž je, že v příkladu nejsou hypotézy ani zcela přijaty, ani zcela zamítnuty. Spíše se stávají více či méně pravděpodobné podle toho, jak jsou získávána data.

Pro zopakování: základní pravděpodobnostní vztahy:

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

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

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

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i), \quad A_1, \ldots, A_n \text{ jsou navzájem výlučné jevy}, \sum_i P(A_i) = 1$$

Bayesův teorém a učení se konceptům

Bayesův teorém poskytuje principiální způsob, jak počítat posteriorní pravděpodobnosti každé z hypotéz za předpokladu trénovacích dat.

Můžeme jej tedy použít jako základ pro učící algoritmus, který pomocí vypočtených pravděpodobností ohodnotí jednotlivé hypotézy a jako výsledek poskytne tu nejpravděpodobnější. Tato metoda bývá někdy nazývána jako „brutální učení pomocí B.T.", tj. metoda hrubé síly.

Je zajímavé, že i ostatní metody (jsou-li aplikovatelné) poskytují jako svůj výstup týchž hypotézy, i když explicitně s pravděpodobnostmi nemanipulují a jsou často výrazně efektivnější.

Brutální Bayesovské učení se konceptům

Je dán konečný prostor H, obsahující hypotézy, definovaný nad instančním prostorem X.

Úkolem je naučit se nějaký cílový koncept:

$$c: X \to \{0, 1\}$$

K dispozici je nějaká posloupnost trénovacích příkladů $\langle\langle x_1, d_1\rangle, \ldots, \langle x_m, d_m\rangle\rangle$, kde x_i je nějaká instance z X a d_i je cílová hodnota pro x_i (tj., $d_i = c(x_i)$). Pro jednoduchost předpokládejme, že sekvence instancí $\langle x_1, \ldots, x_m\rangle$ je konstantní, takže trénovací data D mohou být jednoduše popsána jako posloupnost cílových hodnot $D = \langle d_1, \ldots, d_m\rangle$.

Přímý algoritmus učení se konceptu, jenž vrací maximální a-posteriori hypotézu a je založen na Bayesově teorému, je následující:

1. $\forall h \in H$ vypočítej posteriotní pravděpodobnost:

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

2. Jako výsledek vrať hypotézu h_{MAP} s nejvyšší posteriotní pravděpodobností:

$$h_{MAP} = \underset{h \in H}{\operatorname{argmax}}\ P(h|D)$$

Uvedený algoritmus však může být výpočetně velmi náročný, protože aplikuje Bayesův teorém na každou hypotézu v H k výpočtu $P(h|D)$. Proto může být pro rozsáhlé prostory hypotéz nepraktický, přesto poskytuje standard pro srovnávání ostatních algoritmů na učení konceptů.

Abychom mohli definovat učící problém pro algoritmus hrubé síly, je nutno specifikovat hodnoty pro $P(h)$ a $P(D|h)$; $P(D)$ bude stanoveno, jakmile určíme obě předchozí veličiny – viz dále.

Rozložení pravděpodobností $P(h)$ a $P(D|h)$ zvolíme v souladu s výchozí znalostí o problému. Měly by být konzistentní s následujícími předpoklady:

1. Trénovací data jsou bez šumu (tj. $d_i = c(x_i)$).
2. Cílový koncept c je obsažen v prostoru hypotéz H.
3. Nemáme žádný a-priorní důvod předpokládat, že některá hypotéza je pravděpodobnější než jiná.

Za předpokladu splnění uvedených tří bodů je nutno stanovit hodnoty $P(h)$. Nemáme-li důvod dávat přednost některým hypotézám (bod 3.), je rozumné každé $h \in H$ přiřadit stejnou pravděpodobnost. Protože předpokládáme, že cílový koncept $\in H$, pak se musí součet $P(h)$ rovnat 1:

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

Jak nyní zvolit $P(\mathcal{D}|h)$?

$P(\mathcal{D}|h)$ je pravděpodobnost pozorování cílových hodnot $\mathcal{D} = \langle d_1, \dots, d_m \rangle$ pro pevný soubor instancí $\langle x_1, \dots, x_m \rangle$, za předpokladu „světa" v němž hypotéza h platí (tj. existuje „svět", v němž h je korektním popisem cílového konceptu c).

Protože předpokládáme nezašuměná tréninková data, tak pravděpodobnost pozorování klasifikace d_i za předpokladu h je 1 pro $d_i = h(x_i)$ a 0 pro $d_i \neq h(x_i)$. Tudíž

$$P(\mathcal{D}|h) = \begin{cases} 1 & \text{pro } d_i = h(x_i), \ \forall d_i \in \mathcal{D} \\ 0 & \text{jinak} \end{cases}$$

Jinými slovy, pravděpodobnost dat $\mathcal{D}$ za předpokladu hypotézy h je 1 je-li $\mathcal{D}$ konsistentní s h, jinak 0.

Disponujeme-li nyní volbami pro $P(h)$ a $P(\mathcal{D}|h)$, máme nyní zcela definovaný problém pro algoritmus brutálního Bayesovského učení.

Uvažme první krok algoritmu, v němž se používá Bayesův teorém k výpočtu posteriorních pravděpodobností $P(h|\mathcal{D})$ každé hypotézy h za předpokladu pozorování tréninkových dat $\mathcal{D}$.

Použijeme tedy známý vztah:

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

Uvažme napřed případ, kdy h je nekonsistentní s tréninkovými daty $\mathcal{D}$. Zde je $P(\mathcal{D}|h)$ definováno jako 0, takže dostaneme:

$$P(h|\mathcal{D}) = \frac{0 \cdot P(h)}{P(\mathcal{D})} = 0 \qquad (h \text{ je nekonsistentní s } \mathcal{D})$$

Nyní uvažme případ, kdy h je s $\mathcal{D}$ konsistentní. Zde je $P(\mathcal{D}|h)$ definováno jako 1, takže dostaneme:

$$P(h|\mathcal{D}) = \frac{1 \cdot \frac{1}{|H|}}{P(\mathcal{D})} = \frac{1 \cdot \frac{1}{|H|}}{\frac{|VS_{H,\mathcal{D}}|}{|H|}} = \frac{1}{|VS_{H,\mathcal{D}}|} \qquad (h \text{ konsist. s } \mathcal{D})$$

kde $VS_{H,\mathcal{D}}$ je podmnožina hypotéz z H, které jsou konsistentní s $\mathcal{D}$ (tj. $VS_{H,\mathcal{D}}$ tvoří prostor verzí).

Lze si ověřit, že $P(\mathcal{D}) = \frac{|VS_{H,\mathcal{D}}|}{|H|}$, protože součet přes všechny hypotézy $P(h|\mathcal{D})$ se musí rovnat 1 a dále proto, že počet hypotéz z H konsistentních s $\mathcal{D}$ je dle definice $|VS_{H,\mathcal{D}}|$. Případně lze $P(\mathcal{D})$ odvodit z teorému o celkové pravděpodobnosti $P(B) = \sum_i P(B|A_i)P(A_i)$ a ze skutečnosti, že hypotézy jsou navzájem výlučné ($\forall i \neq j$, $P(h_i \wedge h_j) = 0$):

$$P(\mathcal{D}) = \sum_{h_i \in H} P(\mathcal{D}|h_i)P(h_i) = \sum_{h_i \in VS_{H,\mathcal{D}}} 1 \cdot \frac{1}{|H|} + \sum_{h_i \notin VS_{H,\mathcal{D}}} 0 \cdot \frac{1}{|H|} =$$

$$= \sum_{h_i \in VS_{H,\mathcal{D}}} \frac{1}{|H|} = \frac{|VS_{H,\mathcal{D}}|}{|H|}$$

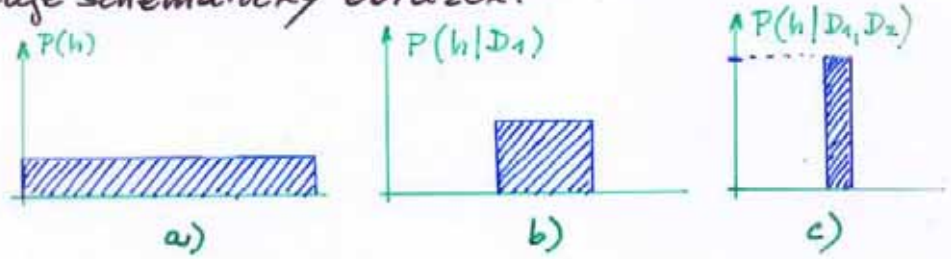
Shrnutí:

Bayesův teorém implikuje, že posteriorní pravděpodobnost $P(h|D)$ za předpokladů $P(h)$ a $P(D|h)$ je:

$$P(h|D) = \begin{cases} \frac{1}{|VS_{H,D}|} & \text{když } h \text{ je konsistentní s } D \\ 0 & \text{jinak} \end{cases}$$

přičemž $|VS_{H,D}|$ je počet hypotéz z H konsistentních s D.

Vývoj pravděpodobností spjatých s hypotézami ilustruje schématicky obrázek:



a) Původně mají všechny hypotézy stejné pravděpodobnosti.

b) Jak přibývají trénovací data, posteriorní pravděpodobnosti nekonsistentních hypotéz se stávají nulové, zatímco celková pravděpodobnost = 1 je sdílena rovnoměrně zbylými konsistentními pravděpodobnostmi.

Závěr:

Na základě volby $P(h)$ a $P(D|h)$ má každá konsistentní hypotéza posteriorní pravděpodobnost $= 1/|VS_{H,D}|$ a každá nekonsistentní má 0. Tudíž každá konsistentní hypotéza je MAP hypotéza.

Optimální Bayesovský klasifikátor

Dosud jsme uvažovali otázku „Která hypotéza je nejpravděpodobnější, jsou-li dána trénovací data?".

Často však bývá otázka položena jinak: „Jaká je nejpravděpodobnější klasifikace nové instance, jsou-li dána určitá trénovací data?".

Může se zdát, že na druhou otázku lze najít jednoduše odpověď pomocí aplikace principu MAP-hypotézy na novou instanci. Ve skutečnosti lze klasifikovat lepším způsobem.

Uvažme prostor hypotéz obsahující 3 hypotézy: h_1, h_2 a h_3. Nechť jsou jejich posteriorní pravděpodobnosti za předpokladu určitých trénovacích dat 0.4, 0.3 a 0.3.

h_1 tedy je MAP-hypotéza. Předpokládejme, že se objeví nová instance x pozitivně klasifikovaná h_1 a negativně pomocí h_2 a h_3. Uvážíme-li všechny hypotézy, pak pravděpodobnost, že x je pozitivní bude 0.4 (podle h_1) a že je negativní 0.6. Nejpravděpodobnější klasifikace (negativní) je v tomto případě různá od klasifikace generované hypotézou MAP.

Obecně se nejpravděpodobnější klasifikace nové instance získá kombinováním predikcí všech hypotéz, přičemž se k váhování použijí posteriorní pravděpodobnosti.

Může-li možná klasifikace nového příkladu nabýt libovolné hodnoty $v_j \in V$, pak pravděpodobnost $P(v_j|D)$, že korektní klasifikace nové instance je v_j udává vztah:

$$P(v_j|D) = \sum_{h_i \in H} P(v_j|h_i) P(h_i|D)$$

Optimální klasifikace nové instance je pak hodnota v_j, pro niž je $P(v_j|D)$ maximální.

Bayesova optimální klasifikace:

$$v_j = \underset{v_j \in V}{\operatorname{argmax}} \sum_{h_i \in H} P(v_j|h_i) P(h_i|D)$$

Ilustrace: $V = \{\oplus, \ominus\}$, $P(h_1|D) = 0.4$, $P(\ominus|h_1) = 0$, $P(\oplus|h_1) = 1$
$P(h_2|D) = 0.3$, $P(\ominus|h_2) = 1$, $P(\oplus|h_2) = 0$
$P(h_3|D) = 0.3$, $P(\ominus|h_3) = 1$, $P(\oplus|h_3) = 0$

takže

$$\sum_{h_i \in H} P(\oplus|h_i) P(h_i|D) = 0.4$$

$$\sum_{h_i \in H} P(\ominus|h_i) P(h_i|D) = 0.6$$

$$\underset{v_j \in \{\oplus, \ominus\}}{\operatorname{argmax}} \sum_{h_i \in H} P(v_j|h_i) P(h_i|D) = \ominus$$

Libovolný systém, který používá pro klasifikaci nových příkladů vztah pro Bayesův optimální klasifikátor, se nazývá optimální Bayesovský učící se systém.

Neexistuje žádná jiná klasifikační metoda, která za předpokladu použití téhož prostoru hypotéz a téže apriorní znalosti bude v průměru dávat lepší výsledky.

Tato metoda maximalizuje pravděpodobnost, že nové instance budou klasifikovány korektně, za předpokladu, že jsou k dispozici určitá data pro naučení, prostor hypotéz a apriorní pravděpodobnosti nad těmito hypotézami.

Gibbsův algoritmus

I když Bayesův optimální klasifikátor poskytuje nejlepší výsledky dosažitelné pomocí trénovacích dat, může (a často bývá) být velmi náročný z výpočetního hlediska.

Tato náročnost plyne ze skutečnosti, že se počítají posteriorní pravděpodobnosti pro každou hypotézu z H a pak se kombinují predikce všech hypotéz pro klasifikaci každé nové instance.

Gibbsův algoritmus poskytuje méně náročnou (a nikoliv optimální) alternativu:

1. Zvol náhodně hypotézu $h \in H$ v souladu s distribucí (rozložením) posteriorních pravděpodobností nad H.
2. Použij h k predikci klasifikace (nové) instance x.

Gibbsův algoritmus tedy jednoduše aplikuje na novou instanci náhodně zvolenou hypotézu s tím, že výběr je ovlivněn okamžitým rozložením posteriorních pravděpodobností.

Bylo ukázáno, že tento postup překvapivě vede jen ke dvojnásobné (maximálně) očekávané klasifikační chybě (Haussler, 1994).

Zajímavá je jedna implikace: Předpokládá-li učící se systém rovnoměrné rozložení apriorních pravděpodobností nad H a jsou-li cílové koncepty ve skutečnosti voleny pomocí takového rozložení ve fázi učení, pak klasifikace nové instance pomocí náhodně zvolené ~~instance~~ hypotézy $\in H$ bude – vzhledem k rovnoměrnému rozložení – mít maximálně dvojnásobnou očekávanou chybu oproti Bayesovu optimálnímu klasifikátoru.

Naivní Bayesovský klasifikátor

Jednou z vysoce praktických metod strojového učení je tzv. naivní Bayesovský klasifikátor (NBK).

V mnoha oblastech je jeho výkonnost z klasifikačního hlediska zcela srovnatelná s umělými neuronovými sítěmi a s rozhodovacími stromy. Ukážeme si princip NBK a příklad aplikace.

NBK je použitelný v případech, kdy lze každou instanci x popsat jako konjunkci hodnot atributů a kde cílová funkce $f(x)$ může nabýt libovolné hodnoty z nějaké konečné množiny V.

Musí být k dispozici soubor trénovacích příkladů pro funkci $f(x)$. Tyto příklady, stejně jako nové instance, jsou popsány n-ticí hodnot atributů $\langle a_1, a_2, \dots, a_n \rangle$.

Naučený systém má za úkol predikovat cílovou hodnotu (klasifikaci) této nové instance.

Bayesovský přístup ke klasifikaci nové instance spočívá v přiřazení nejpravděpodobnější cílové hodnoty v_{MAP} za předpokl., že je dána n-tice $\langle a_1, \dots, a_n \rangle$ popisující tuto instanci:

$$v_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j \mid a_1, a_2, \dots, a_n)$$

S použitím Bayesova teorému lze poslední vztah přepsat na

$$v_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} \frac{P(a_1, a_2, \dots, a_m \mid v_j) \cdot P(v_j)}{P(a_1, a_2, \dots, a_m)} =$$

$$= \underset{v_j \in V}{\operatorname{argmax}} \; P(a_1, a_2, \dots, a_m \mid v_j) \cdot P(v_j)$$

Nyní je zapotřebí stanovit (odhadnout) oba členy pomocí trénovacích dat.

Je snadné odhadnout $P(v_j)$ prostě pomocí výpočtu četnosti, s níž se každá cílová hodnota v_j vyskytuje v trénovacích datech.

Oproti tomu stanovení $P(a_1, a_2, \dots, a_m \mid v_j)$ stejným způsobem je neproveditelné (ledaže bychom disponovali neobyčejně rozsáhlou množinou trénovacích dat). Potíž spočívá ve faktu, že počet takových různých výrazů $P(a_1, a_2, \dots, a_m \mid v_j)$ se rovná počtu možných instancí násobeno počtem možných cílových hodnot. Takže bychom potřebovali „vidět" každou instanci v instančním prostoru mnohokrát, pokud bychom chtěli získat statisticky spolehlivé odhady.

Naivní Bayesovský klasifikátor je založen na zjednodušujícím předpokladu, že hodnoty atributů jsou podmíněně nezávislé za předpokladu dané cílové ~~[illegible]~~hodnoty.

Jinými slovy, předpokládáme, že jsou-li dány cílové hodnoty instance, pak pravděpodobnost zpozorování (výskytu) konjunkce $a_1, a_2, \dots, a_m$ je dána pouze součinem pravděpodobností individuálních atributů:

$$P(a_1, a_2, \dots, a_m \mid v_j) = \prod_i P(a_i \mid v_j)$$

(Obecně není tento předpoklad korektní.)

Použijeme-li uvedené zjednodušení, dostaneme upravený vztah:

NBK:
$$v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} \; P(v_j) \prod_i P(a_i \mid v_j)$$

kde v_{NB} označuje cílovou hodnotu poskytnutou naivním Bayesovským klasifikátorem.

V NBK je počet různých termů $P(a_i \mid v_j)$, které je nutno odhadnout pomocí trénovacích dat, rovno právě počtu různých hodnot atributů násobeno počtem různých cílových hodnot. To je mnohem menší počet než kdybychom měli odhadovat $P(a_1, \dots, a_n \mid v_j)$ bez uvažovaného zjednodušení založeného na předpokladu nezávislosti.

Shrnutí:

Naivní Bayesovské učení zahrnuje učící krok, v němž jsou odhadnuty různé $P(v_j)$ a $P(a_i|v_j)$ pomocí frekvencí výskytů v trénovacích datech.

Soubor těchto odhadů koresponduje s naučenými hypotézami. Tyto hypotézy jsou pak používány ke klasifikaci každé nové instance pomocí rovnice NBK.

Kdykoliv je splněn předpoklad BNK o podmíněné nezávislosti, je naivní Bayesovská klasifikace v_{NB} identická s klasifikací MAP.

Bayesovské učící metody (naivní) nevyžadují explicitní prohledávání prostoru možných hypotéz. V tomto případě je prostorem možných hypotéz prostor možných hodnot, které lze přiřadit různým členům $P(v_j)$ a $P(a_i|v_j)$. Hypotézy jsou jednoduše formovány pomocí výpočtu frekvencí různých kombinací dat mezi trénovacími příklady.

Ilustrativní příklad

Uvažme případ, kdy je nutno zjistit (naučit se) na základě pozorovaných dat, zda je vhodné jít hrát tenis nebo ne:

den	předpověď	teplota	vlhkost	vítr	hrát tenis?
1	slunečno	teplo	vysoká	slabý	NE
2	slunečno	teplo	vysoká	silný	NE
3	zataženo	teplo	vysoká	slabý	ANO
4	déšť	středně	vysoka	slabý	ANO
5	déšť	chladno	normální	slabý	ANO
6	déšť	chladno	normální	silný	NE
7	zataženo	chladno	normální	silný	ANO
8	slunečno	středně	vysoká	slabý	NE
9	slunečno	chladno	normální	slabý	ANO
10	déšť	středně	normální	slabý	ANO
11	slunečno	středně	normální	silný	ANO
12	zataženo	středně	vysoká	silný	ANO
13	zataženo	teplo	normální	slabý	ANO
14	déšť	středně	vysoká	silný	NE

Data použijeme k naučení systému rozeznat cílovou hodnotu ANO/NE pro novou instanci:

⟨slunečno, chladno, vysoká, silný⟩ ... ?

$$v_{NB} = \underset{v_j \in \{ANO, NE\}}{\operatorname{argmax}} P(v_j) \prod_i P(a_i|v_j)$$

⇓

$$v_{NB} = \underset{v_j}{\operatorname{argmax}}\, P(v_j) \cdot P(\text{slunečno} \mid v_j) \cdot P(\text{chladno} \mid v_j) \cdot P(\text{vysoká} \mid v_j) \cdot P(\text{silný} \mid v_j)$$

Poslední vztah byl konkretizován pomocí nové instance.

Pro výpočet v_{NB} nyní potřebujeme zjistit 10 pravděpodobností pomocí tabulky trénovacích dat.

Pravděpodobnosti obou různých cílových hodnot ANO/NE určíme snadno z existujících četností mezi 14 řádky:

$P(\text{hrát tenis} = \text{ANO}) = 9/14 = 0.64$

$P(\text{hrát tenis} = \text{NE}) = 5/14 = 0.36$

Obdobně stanovíme podmíněné pravděpodobnosti:

$P(\text{vítr} = \text{silný} \mid \text{hrát tenis} = \text{ANO}) = 3/9 = 0.33$

$P(\text{vítr} = \text{silný} \mid \text{hrát tenis} = \text{NE}) = 3/5 = 0.60$

atd. Použitím uvedených a stejným způsobem vypočítaných hodnot pro zbývající atributy určíme v_{NB}:

$P(\text{ANO}) \cdot P(\text{slunečno} \mid \text{ANO}) \cdot P(\text{chladno} \mid \text{ANO}) \cdot P(\text{vysoká} \mid \text{ANO}) \cdot P(\text{silný} \mid \text{ANO}) = 0.0053$

$P(\text{NE}) \cdot P(\text{slunečno} \mid \text{NE}) \cdot P(\text{chladno} \mid \text{NE}) \cdot P(\text{vysoká} \mid \text{NE}) \cdot P(\text{silný} \mid \text{NE}) = 0.0206$

$$v_j = \underset{v_j \in \{\text{ANO}, \text{NE}\}}{\operatorname{argmax}} (0.0053, 0.0206) \Rightarrow v_j = \text{NE}$$

Uvedená konstelace hodnot atributů nové instance tedy tadi na tenis nejít.

Normalizací získaných hodnot tak, aby součet = 1, můžeme spočítat podmíněnou pravděpodobnost, že cílová hodnota = NE za předpokladu pozorování uvedených hodnot atributů. Pro náš příklad:

$$P(\text{NE} \mid \text{slunečno}, \text{chladno}, \text{vysoká}, \text{silný}) = \frac{0.0206}{0.0206 + 0.0053} = 0.795$$

Odhad pravděpodobností

Dosud jsme odhadovali pravděpodobnosti pomocí poměrné části výskytu jednotlivých jevů. Např. $P(\text{vítr} = \text{silný} \mid \text{hrát tenis} = \text{NE})$ pomocí podílu n_c/n kde $n = 5$ je celkový počet trénovacích příkladů pro něž hrát tenis = NE, a $n_c = 3$ je počet těch případů pro něž vítr = silný.

Tato pozorování poskytují dobrý odhad pravděpodobnosti v mnoha případech, avšak pro velmi malá n_c bývá tento odhad špatný či nespolehlivý.

Uvažme pro ilustraci případ, kdy

$P(\text{vítr} = \text{silný} \mid \text{hrát tenis} = \text{NE}) = 0.08$

a máme k dispozici vzorek obsahující pouze 5 příkladů pro něž hrát tenis = NE. Potom nejpravděpodobnější hodnota pro $n_c = 0$. (protože na síle větru zde nezáleží)

Tím vzniknou 2 potíže:

1) n_c/n poskytuje podhodnocený odhad
2) nulová hodnota ovládne v budoucnu všechny dotazy obsahující hodnotu vítr = silný (bude vynásobeno nulou).

Aby bylo možno se tomu vyhnout, zavádí Bayesovský klasifikátor odhad pravděpodobností pomocí tzv. m-odhadu:

$$\frac{n_c + mp}{n + m}$$

kde n_c a n mají dříve uvedený význam, p je apriorní odhad pravděpodobnosti, který chceme stanovit, m je konstanta zvaná ekvivalentní velikost vzorku, která určuje jak mnoho je nutno váhovat p relativně k pozorovaným datům.

Typickou metodou k určení p, chybí-li další informace, je předpokládat rovnoměrné rozložení apriorních pravděpodobností.
Tj. pro k možných hodnot atributu $p = \frac{1}{k}$.

Např. při odhadu

$P(\text{vítr} = \text{silný} \mid \text{hrát tenis} = \text{NE})$

víme, že atribut vítr nabývá dvou možných hodnot, takže rovnoměrné apriorní pravděpodobnosti budou $p = 0.5$.

Je-li $n = 0$ pak m-odhad $\equiv n_c/n$.

Pro $m \neq 0$ a $n \neq 0$, pak n_c/n a p budou kombinovány vzhledem k váze m.

m se nazývá ekvivalentní velikostí vzorku proto, že vztah $(n_c + mp)/(n + m)$ můžeme interpretovat jako zvětšení n skutečných pozorování pomocí přídavných m virtuálních vzorků s distribucí odpovídající p.

Příklad aplikace: Naučit se klasifikovat text

Uvažme případ, kdy instancemi jsou textové dokumenty. Cílovým konceptem mohou např. být „elektronické články, které jsou pro mne zajímavé", nebo „www stránky zabývající se strojovým učením".

Pokud by se počítač byl schopen naučit automaticky klasifikovat např. texty získané elektronicky z Internetu, ušetřil by nám – jako inteligentní filtr – mnoho nezáživné práce tím, že by vybral pouze relevantní dokumenty.

K uvedenému účelu se velmi dobře hodí NBK.

Uvažme instanční prostor X obsahující všechny možné textové dokumenty (řetězce slov a interpunkčních znaků všech možných délek).
Dále mějme trénovací příklady (textové dokumenty) pro neznámou cílovou funkci $f(x)$, která může nabýt libovolné hodnoty z konečné množiny V.

Úkolem je naučit se pomocí těchto trénovacích příkladů predikovat cílovou hodnotu pro budoucí textové dokumenty.

Pro ilustraci uvažme cílovou funkci klasifikující dokumenty do dvou tříd: zajímavé / nezajímavé.

Libovolný textový dokument lze reprezentovat velmi jednoduše: atributem je každá pozice slova v dokumentu; hodnotou je slovo na této pozici.

Např. určitý odstavec může mít 100 slov (100 slovních pozic). První hodnotou je první slovo atd.

Předpokl. 700 dokumentů označených jako „zajímavé" a 300 dokumentů označených jako „nezajímavé" (tuto práci, tj. přípravu trénovacích dat, je obvykle nutno udělat „ručně").

$$v_{NB} = \underset{v_j \in \{\text{zajímavé}, \text{nezajímavé}\}}{\operatorname{argmax}} P(v_j) \prod_{i=1}^{100} P(a_i \mid v_j) =$$

$$= \underset{v_j}{\operatorname{argmax}} P(v_j)\, P(a_1 = 1.\text{slovo} \mid v_j)\, P(a_2 = 2.\text{slovo} \mid v_j) \cdots$$

$$\cdots P(a_{100} = 100.\text{slovo} \mid v_j)$$

Naivní Bayesovská klasifikace v_{NB} je tedy klasifikace, která maximalizuje pravděpodobnost pozorování slov, která byla skutečně nalezena v dokumentu, vzhledem k onomu „naivnímu" předpokladu podmíněné nezávislosti – tato nezávislost, že $P(a_1, \ldots, a_{100} \mid v_j) = \prod_{i=1}^{100} P(a_i \mid v_j)$ v uvedeném kontextu znamená, že pravděpodobnost výskytu slova na určité pozici nezávisí na výskytu ostatních slov na ostatních pozicích za předpokladu klasifikace dokumetu $= v_j$. Tento předpoklad není správný. Např. pravděpodobnost výskytu slova „učení" na určité pozici může být větší, pokud slovu předchází „strojové".

Abychom se však vyhnuli nutnosti počítat velké množství pravděpodobnostních výrazů, uvedené zjednodušení přijmeme. Naštěstí, jak ukázaly praktické experimenty, NBK pracuje i tak velice dobře, pouze je o něco snížena přesnost klasifikace.

K výpočtu v_{NB} potřebujeme stanovit $P(v_j)$ a $P(a_i = w_k | v_j)$, kde w_k označuje k-té slovo v nějakém slovníku slov.

$P(v_j)$ určíme snadno jako poměrce:

$$P(\text{zajímavý}) = 700/(700+300) = 0.7$$
$$P(\text{nezajímavý}) = 300/(700+300) = 0.3$$

Stanovení podmíněných pravděpodobností je problematičtější, protože musíme odhadnout jednu pravděpodobnost pro každou kombinaci ~~každého~~ textové pozice, slova a cílové hodnoty.

Např. v angličtině je 50 000 různých slov, takže pro 100 slov v dokumentu a 2 cílové hodnoty musíme určit cca $2 \cdot 100 \cdot 50\,000 \approx 10$ milionů termů z trénovacích dat.

Lze ovšem učinit rozumný předpoklad, redukující takový počet. Budeme předpokládat, že pravděpodobnost výskytu konkrétního slova (např. „čokoládě") nezávisí na konkrétní slovní pozici, kterou uvažujeme (např. a_{23} versus a_{95}). Tj. předpokládáme vzájemnou nezávislost atributů a identické rozložení za předpokladu cílové klasifikace:

$$\forall i,j,k,m: P(a_i = w_k | v_j) = P(a_m = w_k | v_j)$$

Jinak řečeno, odhadujeme celý soubor pravděpodobností $P(a_1 = w_k | v_j)$, $P(a_2 = w_k | v_j)$ ... pomocí jediné pozičně nezávislé pravděpodobnosti $P(w_k | v_j)$, kterou použijeme bez ohledu na pozici slova.

Výsledný efekt tedy bude, že odhadneme pouze $2 \cdot 50\,000 = 100\,000$ různých termů pro $P(w_k | v_j)$. I to je hodně, ale již je to schůdnější.

Ke kompletnímu návrhu učícího algoritmu zbývá ještě určit metodu odhadu pravděpodobností. Můžeme použít m-odhad:

$$P(w_k | v_j) \approx \frac{n_k + 1}{n + |\text{slovník}|}$$

kde n je celkový počet slovních pozic ve všech trénovacích příkladech, jejichž cílová hodnota je v_j, n_k je počet výskytů slova w_k mezi n slovními pozicemi, a konečně $|\text{slovník}|$ je celkový počet různých slov (a dalších lexikálních elementů) nalezených mezi trénovacími údaji.

Sumarizace

Během učení jsou zprvu extrahována slova z trénovacích dokumentů (všech kategorií, $\Theta_1, \Theta_2, \ldots$) a je vytvořen slovník (obsahuje od každého slova jen 1 výskyt). Spočítá se frekvence výskytu těchto slov ve třídě zajímavý a nezajímavý, čímž se získají nezbytné odhady pravděpodobností. Dále se pro nový dokument použijí tyto pravděpodobnosti pro zařazení do příslušné v_j.

Pozn.: Pokud se v nových instancích, které jsou automaticky klasifikovány, vyskytují slova neobsažená v původně vytvořeném slovníku, ignorují se.

Algoritmus

1. Vytvoř kolekci všech slov, interpunkcí, ... z příkladů:

slovník ← soubor jednotlivých slov resp. dalších lexikálních elementů

2. Vypočti požadované $P(v_j)$ a $P(w_k | v_j)$ {

Pro každou cílovou hodnotu $v_j \in V$:

$docs_j$ ← podmnožinu dokumentů z příkladů, pro něž je cílová hodnota v_j

$$P(v_j) \leftarrow \frac{|docs_j|}{|\text{příklady}|}$$

$Text_j$ ← jeden dokument vzniklý konkatenací všech členů $docs_j$

n ← celkový počet různých slovních pozic v $Text_j$

Pro každé slovo $w_k \in$ slovník {

n_k ← počet výskytů slova w_k v $Text_j$

$$P(w_k | v_j) \leftarrow \frac{n_k + 1}{n + |\text{slovník}|}$$

}

}

Fáze klasifikace:

pozice ← všechny slovní pozice v klasifikovaném dokumentu Doc obsahující lexikální elementy a slova ∈ slovník

$$v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \prod_{i \in \text{pozice}} P(a_i | v_j)$$

Uvedený algoritmus v testech pomocí dat z Internetu klasifikoval s přesností 85–90% (podle aplikace).