Наивни бајесовски класификатор

Историјски гледано, једна од првих метода која је коришћена за класификацију неког примера на основу вредности његових атрибута била је метода *Naïve Bayes*. Ова метода подразумева да су сви атрибути подједнако важни за класификацију промера и да су међусобно независни. Иако је овакав сценарио доста нереалан, једноставна шема која се користи код методе *Naïve Bayes* ипак даје веома добре резултате. Погледајмо пример са временским подацима на основу којих треба вршити предикцију томе да ли су временске прилике повољне за тенис или не.

Ова интуитивна метода предикције заснива се на Бајесовом правилу за условну вероватноћу, према коме је вероватноћа догађаја A под условом B, једнака количнику вероватноће догађаја A и B, и вероватноће догађаја B:

$$P(A|B) = \frac{P(A,B)}{P(B)} \tag{1}$$

Из (1) добијамо да је:

$$P(A,B) = P(A|B) \cdot P(B)$$

Слично, из $P(B|A) = \frac{P(A,B)}{P(A)}$ имамо:

$$P(A,B) = P(B|A) \cdot P(A) \tag{2}$$

Изједначавањем левих страна из (1) и (2) добијамо $P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$ што се може записати у облику познатијем као Бајесово правило:

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

Претпоставимо желимо да извршимо класификацију пацијената на оне који су позитивни на корону, од оних који то нису на основу једног симптома: $P(bolest|simptom)$. Употребом Бајесовог правила, уместо да се бавимо вероватноћом болести када је присутан симптом, утврђиваћемо вероватноћу појаве симптома када знамо да је болест присутна: $P(simptom|bolest) \cdot P(bolest)$. Именилац у формули нам није од важности, јер је он исти и у случају да процењујемо вероватноћу да је пацијент позитиван и да нас занима вероватноћа да је негативан. Дакле, све што нам је потребно да бисмо одредили вероватноћу да пацијент има болест под условом да има одређени симптом, јесте вероватноћа појаве тог симптома код болести коју разматрамо и *a priori* вероватноћа да се појави дата болест.

Уопштимо претходну причу на одређивање вероватноће да примерак са вредностима атрибута $A_1, A_2, \dots, A_n$ припада класи C_i. Ако претпоставимо да су атрибути међусобно независни, онда се вероватноће класфикација може одередити употребом Бајесовог правила:

$$P(C|A_1A_2 \dots A_n) = \frac{P(C) \cdot P(A_1|C) \cdot P(A_2|C) \cdot \dots \cdot P(A_n|C)}{P(A_1A_2 \dots A_n)}$$

Примена наивног Бајесовског класификатора подразумева да се у фази учења из података израчунају све вероватноће $P(C), P(A_1|C), P(A_2|C), \dots, P(A_n|C)$. У фази предвиђања класе новог примера, користи се следећа формула, за одређивање класе којој нови пример припада са највећом вероватноћом:

$$\text{argmax}_k \, P(C_k) \cdot \prod_{i=1}^{n} P(A_i|C_k)$$

Именилац је за све претпостављене класе исти, тако да не утиче на избор оне са највећом вероватноћом. Зато је изостављен.

Пример.

Нека је дат скуп података као у табели испод.

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

На основу прве табеле може се направити следећа табела учесталости појављивања одређених комбинација вредности атрибута и лабела:

Outlook			Temperature			Humidity			Windy			Play	
	yes	*no*		*yes*	*no*		*yes*	*no*		*yes*	*no*	*yes*	*no*
sunny	2	3	hot	2	2	high	3	4	false	6	2	9	5
overcast	4	0	mild	4	2	normal	6	1	true	3	3		
rainy	3	2	cool	3	1								
sunny	2/9	3/5	hot	2/9	2/5	high	3/9	4/5	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	mild	4/9	2/5	normal	6/9	1/5	true	3/9	3/5		
rainy	3/9	2/5	cool	3/9	1/5								

У њој су дати резултати пребројавања колико пута се сваки пар атрибут-вредност појављује за сваку од лабела (*yes, no*). На пример, из прве табеле можемо видети да је *outlook = sunny* у пет примера, од којих су два са ознаком *yes*, док су три са ознаком *no*. Учесталост пара *outlook – sunny* међу позитивним примерима је 2/9, док је међу негативним примерима она 3/5.

Нека даље треба направити предикцију да ли су следећи временски услови погодни за играње тениса:

Outlook	Temperature	Humidity	Windy
sunny	cool	high	true

Исход *yes* има следећу вероватноћу:

$$P(yes) = P(outlook = sunny|yes) \cdot P(temperature = cool|yes) \cdot P(humidity = high|yes) \cdot P(windy = true|yes) = \frac{2}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{9}{14} = 0.0053$$

Исход *no* има следећу вероватноћу:

$$P(no) = P(outlook = sunny|no) \cdot P(temperature = cool|no) \cdot P(humidity = high|no) \cdot P(windy = true|no) = \frac{3}{5} \cdot \frac{1}{5} \cdot \frac{4}{5} \cdot \frac{3}{5} \cdot \frac{5}{14} = 0.0206$$

Дакле, *no* је више вероватно него *yes*. Добијене вредности се могу изразити као вероватноће које у збиру дају 1 ако извршимо њихову нормализацију:

$$P(yes) = \frac{0.0053}{0.0053 + 0.0206} = 0.2046 = 20.5\%$$

$$P(no) = \frac{0.0206}{0.0053 + 0.0206} = 0.7953 = 79{,}5\%$$

Лапласовско глачање

Примена наивног бајесовског класификотора може бити проблематична уколико се у скупу података одређена вредност неког атрибута не појављује у комбинацији са сваком ознаком класе. Претпоставимо на пример, да је у подацима за „Тенис“ пример вредност *sunny* атрибута *outlook* увек „везана“ за лабелу *no*. Тада ће вероватноћа да је *outlook=sunny* када је предикција *yes*, P(*outlook=sunny*|*yes*), бити једнака 0. Будући да се ова вероватноћа множи вероватноћама појединих вредности осталих атрибута, коначна вероватноћа лабеле *yes* ће бити 0. Најједноставнији начин за решавање проблема непостојеће комбинације вредност атрибута - лабела у скупу за учење јесте „додавање“ по једног вештачког примера за сваку могућу вредност атрибута. Ови примери неће бити буквално додавани у тренинг скуп, већ ћемо приликом рачунања учесталости појављивања парова атрибут-вредност у комбинацији са неком од ознака класе додати 1 на вредност бројиоца и m на вредност имениоца, где је m_j број различитих вредности неког атрибута $a_j, j = 1, \dots l$. На пример, за ознаку *play = no, outlook* је *sunny* за три примера, *rainy* за два и *overcast* ни за један пример, те су за укупно пет негативних примера вероватноће сваке вредност атрибута *outlook* једнаке редом: 3/5, 2/5, и 0/5. Да бисмо се решили нуле, додаћемо по један фиктивни пример за сваки пар атрибут-вредност, односно укупно три фиктивна примера са негативним исходом. Нове вероватноће ће сада бити: 4/8, 2/8 и 1/8. Овиме смо се осигурали да *outlook= overcast* има малу вероватноћу различиту од нуле.

Вероватноћа да атрибут $a_j, j = 1, \dots, l$ има вредност $v_i, i = 1, \dots m_j$ за лабелу c_k се сада рачуна тако што на број n_{ij} тренинг примера који имају вредност v_i атрибута a_j и лабелу класе c_k додајемо један тренинг пример, а на укупан број n_k примера са лабелом c_k додајемо још онолико примера m_j колико је могући вредности атрибута a_j.

$$P(a_j = v_i|c_k) = \frac{n_{ij} + 1}{n_k + m_j}$$

Овај поступак представља тзв. Лапласовско глачање. Уместо Лапласовског процељивања верватноће, могуће је користити и m-процену:

$$P(a_j = v_i|c_k) = \frac{n_{ij} + mp_i}{n_k + m}$$

Где је m произвољна константа, а p_i је a priori претпостављена вероватноћа сваке вредности v_i атрибута a_j.

Недостајуће вредности атрибута

У случају да је за нови пример који треба класификовати вредност неког од атрибута непозната, у израчунавању ће она једноставно бити прескочена, што неће утицати на резултат класификације. Ако нам је у примеру чију смо класификацију раније извршили непознат *outlook*, онда ће израчунавања *yes* и *no* исхода бити следећа:

$P(yes) = P(temperature = cool|yes) \cdot P(humidity = high|yes) \cdot P(windy = true|yes) = \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{9}{14} = 0.0238$

$P(no) = P(temperature = cool|no) \cdot P(humidity = high|no) \cdot P(windy = true|no) = \frac{1}{5} \cdot \frac{4}{5} \cdot \frac{3}{5} \cdot \frac{5}{14} = 0.0343$

Добијене вредности су знатно веће од оних које смо добили раније, али је резултат класификације остао исти, са вероватноћом 41% да треба, односно 59% да не треба играти тенис.

Ако вредност атрибута недостаје у неком од тренинг примера, онда она једноставно неће бити уврштена у пребројавање, тако да неће имати утицаја на односе вероватноћа који се пребројавањем добијају.

Нумеричке вредности атрибута

Наивни бајесовски класификатор се може користити и у случају атрибута са нумеричким вредностима, и то коришћењем претпоставке да ове вредности имају нормалну тј. Гаусовску расподелу. Уместо рачунања учесталости појединих вредности атрибута за одређену класу, сада се одређују средња вредност и стандардна девијација. Нека су нам „тенис“ подаци сада дати тако да су температура и влажност ваздуха изражене бројчано, онда ћемо за остале атрибуте пребројавање вршити као и раније, док ћемо за температуру и влажност одредити средњу вредност и стандардну девијацију. Стандардну девијацију је квадратни корен варијансе узорка и рачунамо је по формули:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(v_i - \mu)^2}{n - 1}}$$

Где је n број различитих вредности атрибута, v_i је i-та вредност атрибута, а μ је средња вредност.

Густина вероватноће за нормалну расподелу са средњом вредношћу μ и стандардном девијацијом σ рачуна се по формули:

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

Outlook			Temperature			Humidity			Windy			Play	
	yes	*no*		*yes*	*no*		*yes*	*no*		*yes*	*no*	*yes*	*no*
sunny	2	3		83	85		86	85	false	6	2	9	5
overcast	4	0		70	80		96	90	true	3	3		
rainy	3	2		68	65		80	70					
				64	72		65	95					
				69	71		70	91					
				75			80						
				75			70						
				72			90						
				81			75						
sunny	2/9	3/5	*mean*	73	74.6	*mean*	79.1	86.2	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	*std. dev.*	6.2	7.9	*std. dev.*	10.2	9.7	true	3/9	3/5		
rainy	3/9	2/5											

Дакле, ако је потребно класификовати нови пример код кога је *temperature* = 66*, humidity = 90* онда је:

$$f(temperature = 66|yes) = \frac{1}{\sqrt{2\pi}\cdot 6.2} e^{\frac{(66-73)^2}{2\cdot 6.2^2}} = 0.034$$

$$f(temperature = 66|no) = \frac{1}{\sqrt{2\pi}\cdot 7.9} e^{\frac{(66-74.6)^2}{2\cdot 7.9^2}} = 0.0221$$

$$f(humidity = 90|yes) = \frac{1}{\sqrt{2\pi}\cdot 10.2} e^{\frac{(90-79.1)^2}{2\cdot 10.2^2}} = 0.0221$$

$$f(humidity = 90|no) = \frac{1}{\sqrt{2\pi}\cdot 9.7} e^{\frac{(90-74.6)^2}{2\cdot 9.7^2}} = 0.0381$$

Густина вероватноће $f(x)$ представља вероватноћу да x припада ε околини вредности x, тј. да се налази на произвољно малој удаљености ε од x. Ипак, у израчунавању *yes* и *no* исхода, густину вероватноће ћемо користити као вероватноћу:

$P(yes) = P(outlook = sunny|yes)\cdot f(temperature = 66\,|yes)\cdot f(humidity = 90|yes)\cdot P(windy = true|yes) = \frac{2}{9}\cdot 0.034\cdot 0.0221\cdot \frac{3}{9}\cdot \frac{9}{14} = 0.000036$

$P(no) = P(outlook = sunny|no)\cdot f(temperature = 66|no)\cdot f(humidity = 90|no)\cdot P(windy = true|no) = \frac{3}{5}\cdot 0.0221\cdot 0.0381\cdot \frac{3}{5}\cdot \frac{5}{14} = 0.000108$

$$P(yes) = \frac{0.000036}{0.000036 + 0.000108} = 25\%$$

$$P(no) = \frac{0.000108}{0.000036 + 0.000108} = 75\%$$

Бајесовски модели за класификацију докумената

Значајан домен машинског учења је класификација докумената, у којој је свака инстанца један документ, а класа инстанце је тема документа. Документи могу бити вести, а класе могу бити домаће вести, иностране вести, финансијске вести и спорт. Документе карактеришу речи које се појављују у њима, а један од начина примене машинског учења на класификацију докумената је третирање присуства или одсуства сваке речи као логички атрибут (TRUE, FALSE). Наивни Бајес је популарна техника за овај домен машинског учења, јер је врло брза и прилично тачна.

Међутим, ово не узима у обзир број појављивања сваке речи, што је потенцијално корисна информација при одређивању категорије документа. Документи се не посматрају као скупови, већ као „торбе речи“ у којима се различите речи могу појављивати више пута.

Нека је n_i број појављивања речи i у документу, а P_i је вероватноћа да се реч i нађе у узорку извученом из свих докумената категорије H. Претпоставка је да ова вероватноћа не зависи од контекста и позиције речи у документу. Са оваквом расподелом, вероватноћа да је документ класе H управо документ E рачуна се по формули:

$$P(E|H) \approx N! \cdot \prod_{i=1}^{k} \frac{P_i^{n_i}}{n_i!}$$

Где је $N = n_1 + n_2 + \cdots + n_k$ број речи у документу. Факторијели су ту да би се узели у обзир различити редоследи појављивања речи у документу. P_i се процењује израчунавањем релативне учесталости речи i у текстовима свих докумената класе H коришћених за тренирање.

На пример, нека документи класе H имају речник од само две речи: *plavo* и *crveno* и нека су вероватноће појављивања ових речи у документима класе H: $P(crveno|H) = 75\%$ и $P(plavo|H) = 25\%$. Нека је E документ $plavo\ crveno\ plavo$ чија је дужина $N = 3$ речи. Постоје четири могуће торбе од три речи направљене коришћењем речника који садржи две речи: $\{crveno\ crveno\ crveno\}$, $\{plavo\ plavo\ plavo\}$, $\{crveno\ crveno\ plavo\}$,$\{crveno\ plavo\ plavo\}$. Њихове вероватноће су:

$$P(\{crveno\ crveno\ crveno\}|H) = 3! \cdot \frac{0.75^3}{3!} \cdot \frac{0.25^0}{0!} = \frac{27}{64}$$

$$P(\{plavo\ plavo\ plavo\}|H) = 3! \cdot \frac{0.75^0}{0!} \cdot \frac{0.25^3}{3!} = \frac{1}{64}$$

$$P(\{crveno\ crveno\ plavo\}|H) = 3! \cdot \frac{0.75^2}{2!} \cdot \frac{0.25^1}{1!} = \frac{27}{64}$$

$$P(\{crveno\ plavo\ plavo\}|H) = 3! \cdot \frac{0.75^1}{1!} \cdot \frac{0.25^2}{2!} = \frac{9}{64}$$

Документу E одговара последња торба речи, па је вероватноћа да је документ који припада класи H уједно документ E једнака $\frac{9}{64} = 14\%$. Ако бисмо имали неку класу H' за коју знамо да је $P(crveno|H) = 10\%$ и $P(plavo|H) = 90\%$ онда би вероватноћа да документ који припада класи H' буде документ E била једнака $P(\{crveno\ plavo\ plavo\}\,|\,H') = 3! \cdot \frac{0.1^1}{1!} \cdot \frac{0.9^2}{2!} = 24.3\%$.

Ипак, претходне бројке не значе обавезно да је вероватноћа да документ E припада класи H' већа него да он буде класе H. Да бисмо до краја применили Бајесово правило, потребно је да

знамо и *a priori* вероватноћу да документи буду класе H или H'. На пример, ако бисмо знали да су документи класе H троструко чешћи од докумената класе H', онда би коришћењем Бајесовог правила добили да је документ документ E припада класи H:

$$P(H|\{crveno\ plavo\ plavo\}) = P(H) \cdot P(\{crveno\ plavo\ plavo\}|H) = 0.75 \cdot 0.14 = 0.105$$

$$P(H'|\{crveno\ plavo\ plavo\}) = P(H') \cdot P(\{crveno\ plavo\ plavo\}|H') = 0.25 \cdot 0.24 = 0.06$$

Наивни Бајес даје једноставан приступ, са јасном семантиком за представљање, коришћење и учење уз пробабилистички приступ. Помоћу њега се могу постићи импресивни резултати. Често се показује да Наивни Бајес надмашује софистицираније класификаторе над многим скуповима података.

Међутим, постоји много скупова података над којима Наиве Баиес нема тако добре резултате и лако је увидети зашто. Будући да се атрибути третирају као да су потпуно независни, додавање сувишних атрибута доводи до ремећења процеса учења.

Као крајњи пример, ако у податке о времену уврстите нови атрибут са истим вредностима као и код атрибута температура, ефекат атрибута температура би се удвостручио: све вероватноће би би се квадрирале, што би му дало много већи утицај у одлуци. Зависности између атрибута неизбежно смањују моћ Наивног Бајес-а. Резултати се могу побољшати коришћењем подскупова атрибута у поступку доношења одлука, уз пажљиви одабир оних којих ће бити употребљени.

Претпоставка нормалне расподеле за нумеричке атрибуте је још једно ограничење за Наивног Бајеса. Међутим, ништа нас не спречава да користимо друге расподеле за нумеричке атрибуте. Ако се зна да одређени атрибут вероватно има неку другу расподелу, могу се користити стандардни поступци процене те друге расподеле. Друга могућност је да се подаци преведу у дискератн облик, па онда примени Наивни Бајес.