Uczenie się pojęć

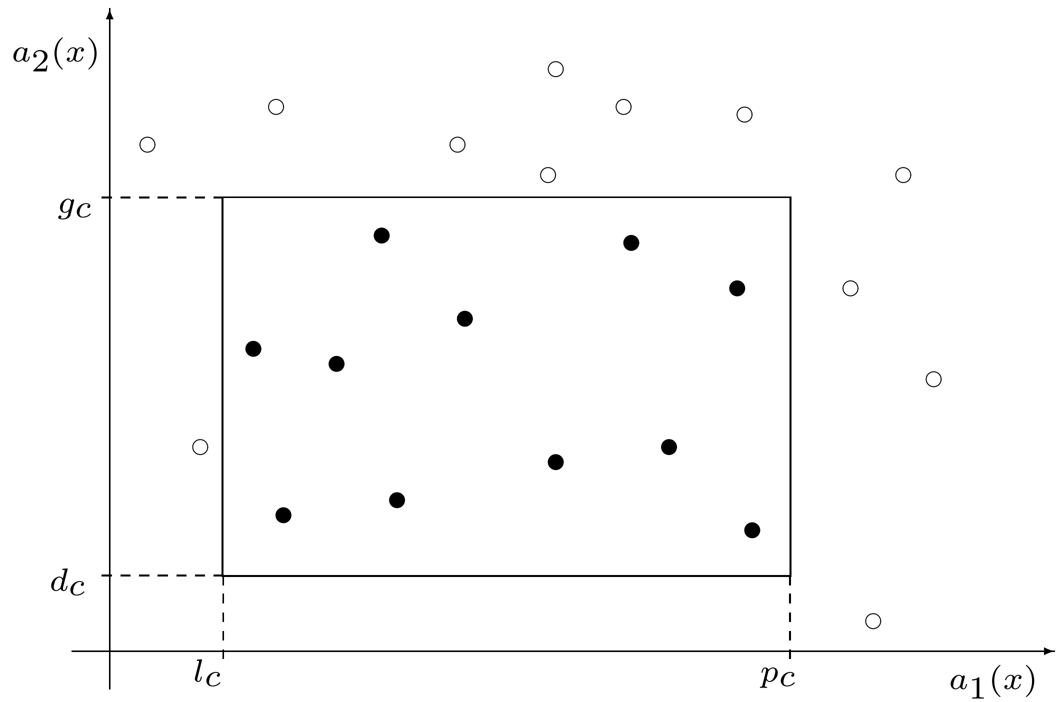
Mamy dziedzinę zbiór X, atrybuty $a \in \mathcal{A}$ jako dowolne funkcje określone na tej dziedzinie $a : X \mapsto A$, i klasę pojęć $\mathcal{C}$. Pojęcia $c \in \mathcal{C}$ są funkcjami $c : X \mapsto C$ gdzie C jest zbiorem kategorii pojęć klasy $\mathcal{C}$.

Pojęcie **pojedyncze** c ma zbiór kategorii $C = \{0, 1\}$. Każde pojęcie pojedyncze wyznacza podzbiór dziedziny zawierający pozytywne próbki tego pojęcia $X^c = \{x \in X \mid c(x) = 1\}$.

W ogólnym przypadku pojęcie **wielokrotne** ma zbiór kategorii o liczności $|C| > 2$. W zbiorze próbek $P \subseteq X$ tego pojęcia możemy wyróżnić zbiór próbek z kategorią $d \in C$ oznaczany P^{cd} lub w skrócie P^d, czyli $P^d = \{x \in P \mid c(x) = d\}$, dla jakiegoś pojęcia c. Jeśli zatem c jest pojęciem pojedynczym to zbiór próbek pozytywnych tego pojęcia $X^c = X^1$ a zbiór $X^0 = X \setminus X^1$ jest zbiorem wszystkich jego próbek negatywnych.

Przykład: zbiór prostokątów na płaszczyźnie[1]

Pojedyncze pojęcie c będzie wyznaczone przez parę punktów (l_c, d_c) i (p_c, g_c), a zbiór próbek pojęcia c mamy jako: $X^c = \{x \in \mathcal{R}^2 \mid l_c \leqslant a_1(x) \leqslant p_c \wedge d_c \leqslant a_2(x) \leqslant g_c\}$



[1] Przykład z prostokątami na płaszczyźnie pochodzi z [Kearns, Vazirani, 1994].

Przykład: zbiór funkcji boolowskich

Dziedziną będzie zbiór n-elementowych łańcuchów binarnych dla pewnego $n \geqslant 1$, czyli $X = \{0, 1\}^n$. Próbkami pojęć będą n-elementowe ciągi zer i jedynek, które można opisać n funkcjami atrybutów $a_1, ..., a_n$ reprezentującymi kolejne bity, gdzie $a_i : X \mapsto \{0, 1\}$.

Pojęcia mogą być reprezentowane przez formuły logiczne, na przykład dla $n = 3$ pojęcie c mogłoby być reprezentowane przez formułę $\neg a_1(x) \wedge (a_2(x) \vee a_3(x))$, co oznacza, że pozytywnymi próbkami są wszystkie ciągi z zerem w pierwszym bicie, i z jedynką przynajmniej w jednym z bitów 2 i 3.

Zbiór wszystkich możliwych różnych pojęć pojedynczych jest równoważny zbiorowi wszystkich różnych funkcji boolowskich dla n-elementowych ciągów binarnych i ma liczność 2^{2^n}.

Hipotezy

Hipotezy będą reprezentowały wynik uczenia się. Zbiór możliwych hipotez oznaczymy $\mathcal{H}$. Każda hipoteza $h \in \mathcal{H}$ jest funkcją $h : X \mapsto C$, podobnie jak pojęcia. Dokładne nauczenie się pojęcia c jest możliwe tylko wtedy gdy $\mathcal{C} \subseteq \mathcal{H}$, bo mamy wtedy pewność, że w zbiorze hipotez jest hipoteza identyczna z pojęciem, którego się uczymy. Na odwrót, dokładne nauczenie się pojęcia może nie być możliwe gdy $\mathcal{H} \subset \mathcal{C}, \mathcal{H} \neq \mathcal{C}$.

Ponieważ próbki uczące opisane są za pomocą atrybutów, więc hipotezy muszą w efekcie mieć postać funkcji: $h : A_1 \times A_2 ... \times A_n \mapsto C$ gdzie $A_1, A_2, ..., A_n$ są przeciwdziedzinami atrybutów określonych na X. Gdyby dla dwóch różnych próbek x_1, x_2 wszystkie atrybuty miały dla nich równe wartości, wtedy dla każdej hipotezy h byłoby $h(x_1) = h(x_2)$.

Informacja trenująca

Dla zbioru próbek $P \subseteq X$ oznaczymy:

P^h zbiór próbek z P **pokrywanych** przez hipotezę h, czyli takich, dla których $h(x) = 1$ (dla pojęć pojedynczych)

P^{hd} zbiór próbek z P, dla których $h(x) = d, d \in C$ (dla dowolnych pojęć)

Próbkę etykietowaną pojęcia c określonego na dziedzinie X zapisujemy $\langle x, c(x) \rangle$, gdzie $x \in X$.

Seria ucząca jest serią próbek etykietowanych, gdzie obiekt x jest wyrażony przez wektor wartości atrybutów, a etykietą jest przypisana mu kategoria.

Praktycznie, seria ucząca jest serią wektorów atrybutów z etykietami.

Błędy w uczeniu się pojęć

Ogólnie: błąd charakteryzuje stopień zgodności klasyfikacji próbek przez pojęcie docelowe i hipotezę.

Błąd próbki $e_P^c(h)$ dla zbioru próbek P

$$e_P^c(h) = \frac{|\{x \in P | h(x) \neq c(x)\}|}{|P|}$$

gdzie

$$r_P^c(h) = |\{x \in P | h(x) \neq c(x)\}|$$

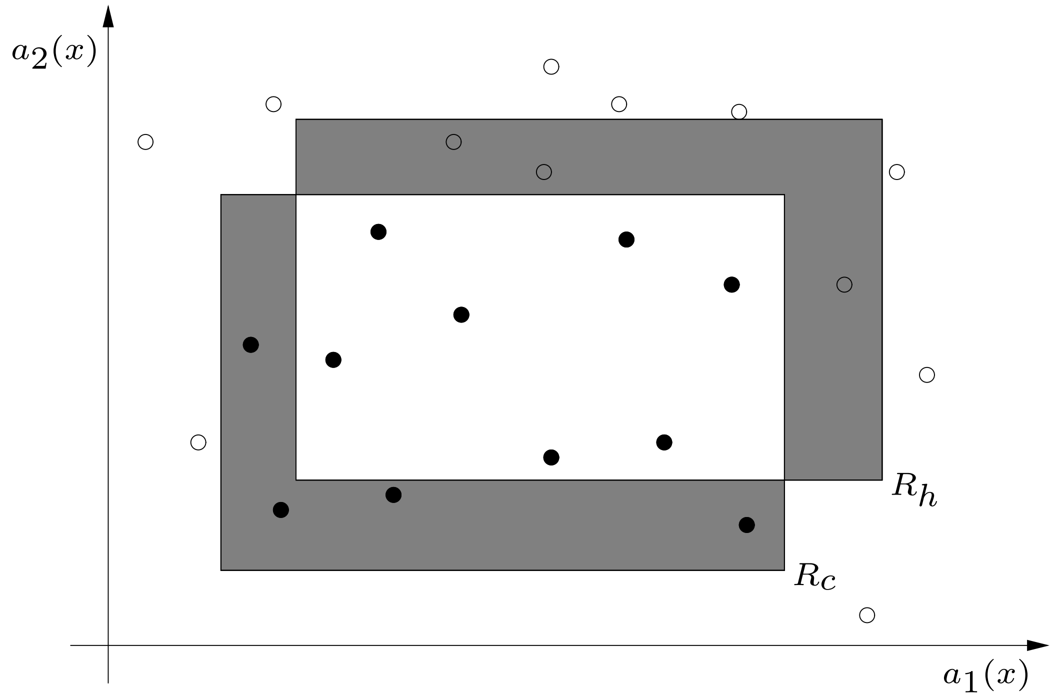
jest liczbą niepoprawnie zaklasyfikowanych próbek ze zbioru P.

Błąd rzeczywisty hipotezy to jest wartość oczekiwana błędu próbki na losowo wybranym zbiorze próbek. Jeśli próbki są wybierane z dziedziny zgodnie z pewnym rozkładem prawdopodobieństwa Ω to błąd rzeczywisty będzie:

$$e_\Omega^c(h) = \mathsf{Pr}_{x \in \Omega}(h(x) \neq c(x))$$

gdzie $\mathsf{Pr}_{x \in \Omega}$ oznacza prawdopodobieństwo przy założeniu wylosowania x ze zbioru X zgodnie z rozkładem Ω.

Błędy w uczeniu się pojęć — przykład



Dla pojęcia i hipotezy, oraz 23 próbek przedstawionych na powyższym rysunku błąd próbki $e^c_P(h) = \frac{7}{23}$, ponieważ 7 próbek zostało błędnie zaklasyfikowanych przez hipotezę (4 pozytywne, i 3 negatywne).

Błąd rzeczywisty $e^c_\Omega(h)$ jest prawdopodobieństwem tego, że próbka wybrana losowo z rozkładem Ω będzie należała do obszaru zacienionego, czyli $\text{Pr}_\Omega(R_c \div R_h)$, gdzie R_c i R_h oznaczają prostokąty odpowiadające pojęciu i hipotezie, a $R_c \div R_h$ oznacza różnicę symetryczną zbiorów.

Zadanie indukcyjnego uczenia się pojęć

Mając dane: dziedzinę X, klasę pojęć $\mathcal{C}$, i przestrzeń hipotez $\mathcal{H}$ ucznia, oraz pewne ustalone, ale nieznane, pojęcie docelowe $c \in \mathcal{C}$, jak również zbiór trenujący $T \subseteq X$ i serię uczącą próbek etykietowanych z tego zbioru, należy znaleźć hipotezę $h \in \mathcal{H}$, która jest najlepszym przybliżeniem pojęcia c według pewnego kryterium.

Jako kryterium, w najprostszym przypadku, można przyjąć minimalizację błędu próbki na zbiorze trenującym $e_T^c(h)$.

W rzeczywistości jednak celem uczenia się jest osiągnięcie najlepszego przybliżenia pojęcia docelowego w przypadku ogólnym. Jeśli próbki wybierane są z dziedziny zgodnie z pewnym rozkładem Ω, i zbiór trenujący również został skonstruowany na podstawie tego rozkładu, to jako kryterium wyboru hipotezy można przyjąć minimalizację błędu rzeczywistego $e_\Omega^c(h)$.

Inne rodzaje indukcyjnego uczenia się

Gdy nieznane są kategorie próbek pojęć, uczeń może być zmuszony sam je sobie stworzyć. Zadanie uczenia się **tworzenia pojęć** jest zatem dwuetapowe: pierwszym etapem jest pogrupowanie próbek serii uczącej na podgrupy odpowiadające wybranym kategoriom, a następnie uczenie się tak określonego pojęcia.

W uczeniu się pojęć zbiorem wartości jest niewielki zwykle zbiór kategorii. Niekiedy może to być cały zbiór liczb rzeczywistych, i wtedy zadanie uczenia się polega na uczeniu się **aproksymacji nieznanej funkcji** na podstawie serii próbek.

Tryby uczenia

Tryb wsadowy — cała seria trenująca od razu, żadnych interakcji ze środowiskiem lub nauczycielem. Najprostszy tryb uczenia się; każdy algorytm może być użyty w tym trybie.

Tryb inkrementacyjny — próbki trenujące są dostarczane pojedynczo i uczeń powinien każdorazowo korygować swoją hipotezę bieżącą. Przydatny gdy nie ma określonego zbioru uczącego a uczenie odbywa się przy bieżącej obserwacji.

Tryb epokowy — tryb podobny do inkrementacyjnego, lecz próbki dostarczane są partiami. Mają zastosowanie te same algorytmy.

Tryb korekcyjny — próbki uczące podawane są pojedynczo, lecz bez etykiety, i uczeń ma wyznaczyć wartość swojej hipotezy dla danej próbki, po czym otrzymuje od nauczyciela informację korygującą.

Szacowanie błędów hipotez

Estymacja przedziałowa — szacowanie wartości pewnego nieznanego parametru populacji na podstawie **estymatora** — zmiennej losowej o wartościach wyznaczanych na podstawie losowej próby elementów z tej populacji.

Estymacja przedziałowa polega na wyznaczaniu **przedziałów ufności** dla estymowanego parametru na podstawie estymatora. Robi się to dla ustalonego **poziomu ufności**, czyli prawdopodobieństwa, że rzeczywista wartość parametru znajdzie się w tym przedziale.

Przedziałem ufności dla parametru p o poziomie ufności $1 - \delta$ dla $0 \leqslant \delta < 1$ jest każdy przedział, do którego wartość p należy z prawdopodobieństwem $1 - \delta$.

Przedziały ufności dla błędu rzeczywistego hipotezy można wyliczyć korzystając z rozkładu Bernoulliego (k sukcesów w n próbach). Zakładając, że „sukcesem” będzie pomyłka w klasyfikacji dla kolejnego przypadku wybranego z dziedziny z rozkładem Ω, prawdopodobieństwo tego sukcesu jest błędem rzeczywistym.

Model PAC

W modelu PAC (ang. *probably approximately correct*) przyjmuje się założenia jak w zadaniu indukcyjnego uczenia się, oraz dodatkowo zakłada się, że kolejne próbki zbioru trenującego T generuje **wyrocznia**, którą definiuje się jako zmienną losową EX(c, Ω) zwracającą próbkę etykietowaną $\langle x, c(x) \rangle$ przy czym $x \in X$ jest próbką wylosowaną zgodnie z rozkładem Ω.

Zadaniem ucznia jest wygenerowanie hipotezy $h \in \mathcal{H}$ minimalizującej błąd rzeczywisty $e^c_\Omega(h)$.

Model PAC zmierza do określenia warunków, przy których jest możliwe wyznaczenie hipotezy h o ograniczonym błędzie rzeczywistym z określonym prawdopodobieństwem. Te warunki nazywają się warunkami PAC-nauczalności.

PAC-nauczalność

Klasa pojęć $\mathcal{C}$ jest **PAC-nauczalna** za pomocą $\mathcal{H}$, jeśli istnieje algorytm, który dla dowolnych $0 < \epsilon < 1$ i $0 < \delta < 1$ w wyniku pracy z dostępem do wyroczni EX generującej informacje trenujące, z prawdopodobieństwem co najmniej $(1 - \delta)$ znajduje hipotezę h, dla której $e^c_\Omega(h) \leqslant \epsilon$.

Zauważmy, że hipoteza musi zostać wygenerowana dla dowolnie wybranych parametrów ϵ i δ, i dla dowolnego pojęcia c i rozkładu Ω.

Możnaby podejrzewać, że są to wygórowane wymagania i PAC-nauczalność jest trudna do osiągnięcia i rzadka. Jednak można popatrzeć na to z innej strony. Jeśli rozważamy hipotezę h która często daje istotne błędy ($> \epsilon$), to po przetestowaniu jej na dostatecznej liczbie próbek uczących, wada hipotezy zostanie wykryta, i będzie ona mogła być odrzucona. Na odwrót, jeśli potrafimy wygenerować hipotezę spójną z dowolnie dużą liczbą próbek uczących (tzn. błąd hipotezy będzie $\leqslant \epsilon$), to prawdopodobnie ta hipoteza będzie poprawna, i to prawdopodobieństwo będzie tym większe, im większa jest liczba próbek.

PAC-nauczalność można zatem osiągnąć za pomocą kompetentnego algorytmu generowania hipotez i dostatecznie dużej liczby próbek uczących.

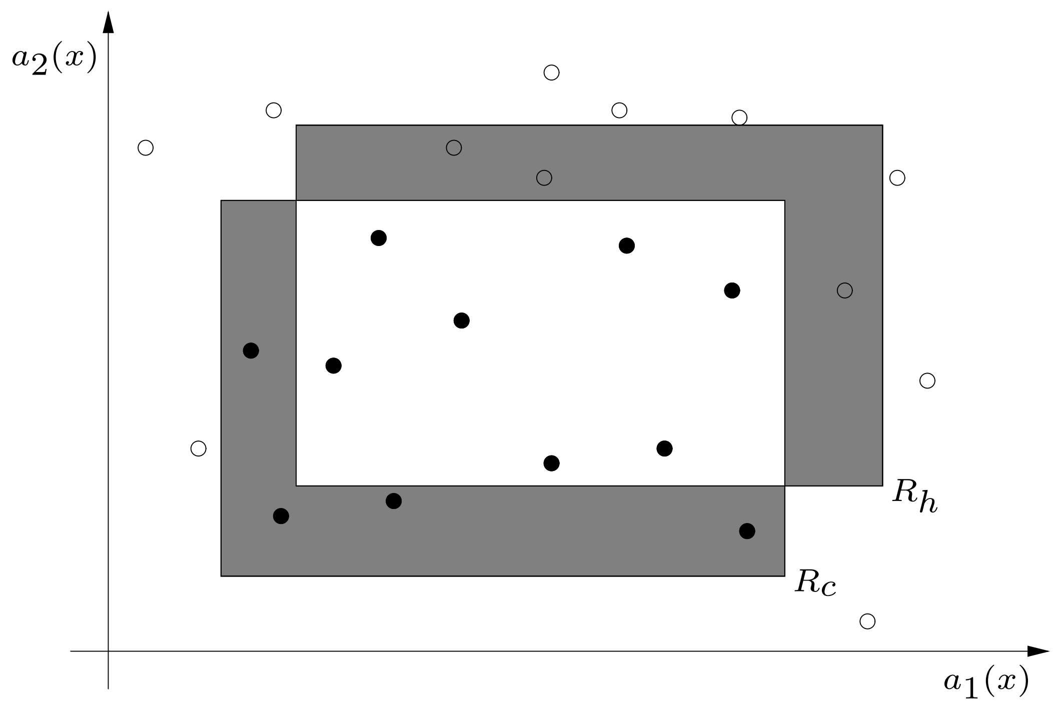
Efektywna PAC-nauczalność

Klasa pojęć $\mathcal{C}$ jest **efektywnie PAC-nauczalna** jeśli jest PAC-nauczalna i istnieje algorytm PAC-uczenia się, który działa w czasie wielomianowym ze względu na $1/\delta$, $1/\epsilon$, rozmiaru próbki z X i rozmiaru pojęcia z $\mathcal{C}$.

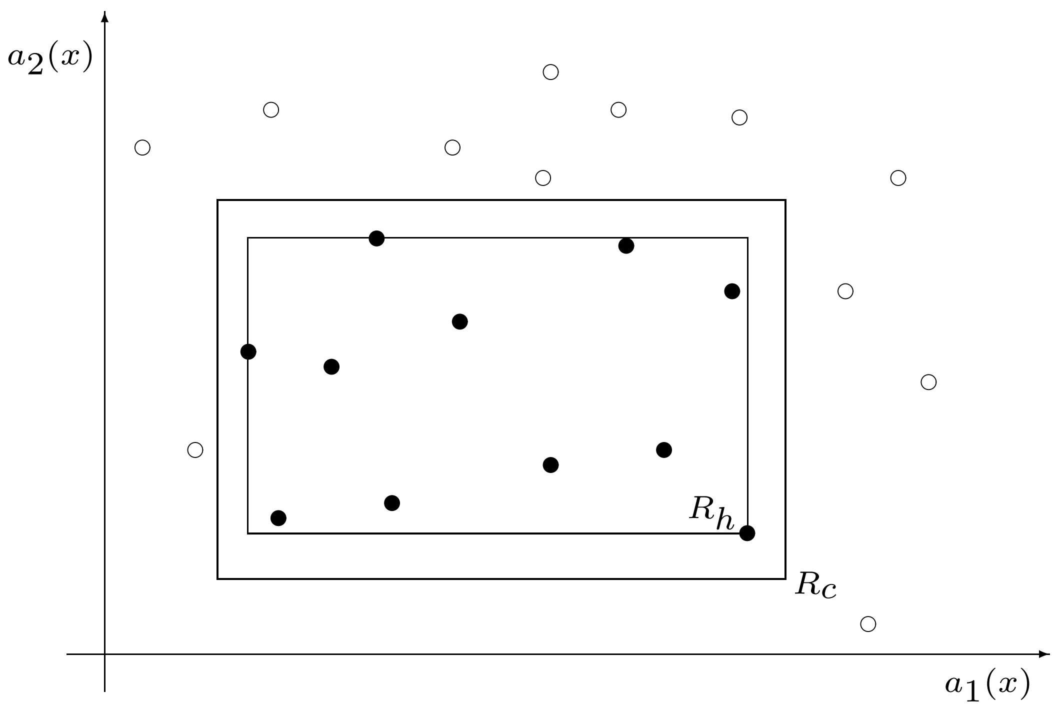
Użyte tu rozmiary próbek i pojęć mają stanowić miarę ich złożoności przekładającą się na nakład obliczeń niezbędny do ich przetwarzania. Dla próbek rozmiarem może być liczba atrybutów. Dla pojęć wyznaczenie rozmiaru jest trudniejsze. Na przykład, dla dziedziny prostokątów można przyjąć stały rozmiar równy 4, ponieważ każde pojęcie można opisać czterema liczbami. Dla dziedziny funkcji boolowskich jako rozmiar pojęcia można przyjąć liczbę literałów niezbędnych do zapisania funkcji prawdy pojęcia. Będą zatem istniały pojęcia o mniejszych i większych rozmiarach.

PAC-nauczalność dla prostokątów

Odniesiemy obecnie pojęcie PAC-nauczalności do dziedziny prostokątów na płaszczyźnie. Błąd rzeczywisty dowolnej hipotezy h względem dowolnego pojęcia c i rozkładu prawdopodobieństwa Ω wynosi: $e^c_\Omega(h) = \mathrm{Pr}_\Omega(R_c \div R_h)$, gdzie R_c i R_h oznaczają prostokąty odpowiadające pojęciu i hipotezie,
a $R_c \div R_h = (R_c \setminus R_h) \cup (R_h \setminus R_c)$ oznacza różnicę symetryczną zbiorów.



Rozważmy algorytm najciaśniejszego dopasowania, generujący dla danego zbioru trenującego T najmniejszy prostokąt zawierający wszystkie próbki pozytywne. Wtedy $R_h \subseteq R_c$, a zatem $R_c \div R_h = R_c \setminus R_h$.



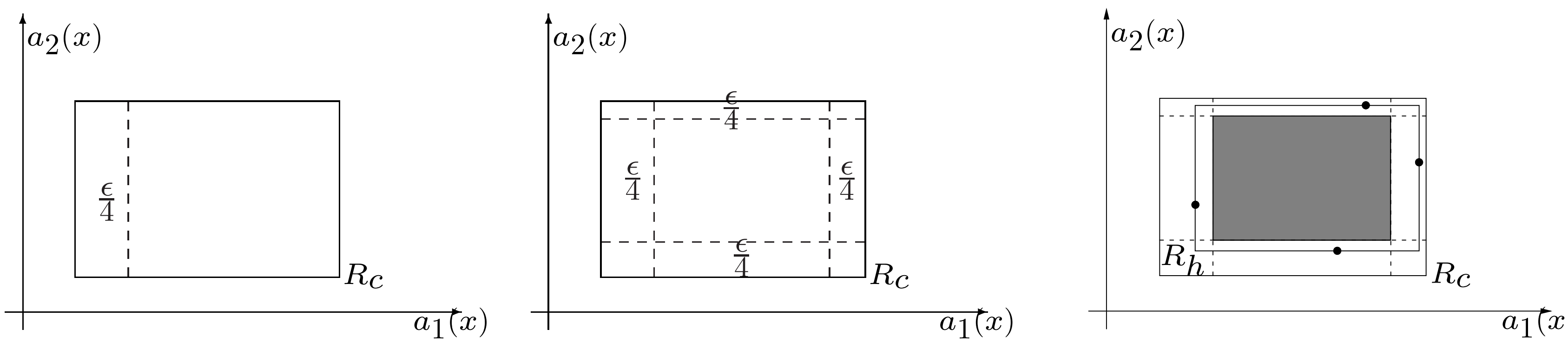
Chcemy związać błąd rzeczywisty hipotezy z prawdopodobieństwem jego otrzymania. Jakie jest prawdopodobieństwo, że nowo wygenerowany punkt należący do R_c nie będzie należał do R_h?

Możemy przyjąć, że:

$$\text{Pr}_\Omega(R_c) \geqslant \epsilon$$

(Ponieważ gdyby, na odwrót, było $\text{Pr}_\Omega(R_c) < \epsilon$ to również byłoby $e^c_\Omega(h) = \text{Pr}_\Omega(R_c \setminus R_h) < \epsilon$ i błąd rzeczywisty byłby zawsze poniżej dopuszczalnej granicy.)

Rozważmy przesuwanie lewego boku prostokąta pojęcia R_c tak długo, aż prawdopodobieństwo wylosowania punktu w prostokącie zmniejszy się o $\frac{\epsilon}{4}$. Na pewno jest to możliwe (rysunek po lewej).



Możemy powtórzyć tę operację dla wszystkich boków prostokąta R_c (rysunek środkowy). Aby błąd hipotezy h nie przekraczał ϵ, wystarczy zapewnić, że we wszystkich pasach bocznych prostokąta R_c znajdzie się przynajmniej po jednej próbce trenującej (rysunek po prawej). Wtedy dla najciaśniejszego dopasowania będzie:

$$\text{Pr}_\Omega(R_c \setminus R_h) \leqslant \epsilon$$

Jak możemy zagwarantować, że co najmniej po jednym punkcie znajdzie się w każdym z bocznych pasów prostokąta pojęcia?

Tak by się nie stało, gdyby co najmniej jeden z pasów bocznych został ominięty przez wszystkie punkty zbioru uczącego. Prawdopodobieństwo ominięcia jednego konkretnego pasa bocznego przez wszystkie próbki serii T wynosi najwyżej: $(1 - \frac{\epsilon}{4})^{|T|}$, zatem prawdopodobieństwo ominięcia któregokolwiek z nich jest cztery razy większe: $4(1 - \frac{\epsilon}{4})^{|T|}$.

Chcielibyśmy, żeby to prawdopodobieństwo nie przekroczyło δ:

$$4(1 - \frac{\epsilon}{4})^{|T|} \leqslant \delta$$

co korzystając z nierówności $1 + \alpha \leqslant e^{\alpha}$ przy $\alpha = -\frac{\epsilon}{4}$ można przekształcić do:

$$|T| \geqslant \frac{4}{\epsilon}(\ln 4 + \ln \frac{1}{\delta})$$

Przedstawione rozumowanie dowodzi PAC-nauczalności zagadnienia prostokątów (algorytmem najciaśniejszego dopasowania próbek pozytywnych), i jednocześnie daje minimalną liczbę próbek, gwarantującą spełnienie warunków PAC-nauczalności.

Warunki ogólne PAC-nauczalności

Można wykazać, że w przypadku ogólnym, jeśli istnieje algorytm generujący hipotezy spójne z dostatecznie długą serią uczącą, to klasa pojęć jest PAC-nauczalna.

Konkretnie, dla liczby próbek, zwanej złożonością próbkową (*sample complexity*):

$$N \geqslant \frac{1}{\epsilon}(\ln\frac{1}{\delta} + ln|\mathcal{H}|)$$

jeśli algorytm generuje hipotezy spójne z tymi próbkami, dla dowolnych ϵ i δ, to z prawdopodobieństwem co najmniej $(1 - \delta)$ hipotezy te posiadają błąd mniejszy niż ϵ.

Zauważmy, że powyższe twierdzenie mówi, że dowolna hipoteza spójna z danej długości zbiorem próbek będzie spełniała warunki PAC-nauczalności. Zauważmy również, że liczba próbek określona w twierdzeniu jest warunkiem wystarczającym PAC-nauczalności.

Rozmiar przestrzeni hipotez

Wartość $|\mathcal{H}|$ we wzorze na złożoność próbkową jest rozmiarem przestrzeni hipotez, która np. dla zbioru funkcji boolowskich wynosi 2^{2^n}. Zatem złożoność próbkowa tej przestrzeni rośnie tak jak 2^n. Ponieważ liczba możliwych próbek pozytywnych również wynosi 2^n, zatem osiągnięcie PAC-nauczalności dla klasy funkcji boolowskich może wymagać zbadania prawie wszystkich próbek.

Przyczyna tego jest oczywista — przestrzeń hipotez zawiera hipotezy (funkcje boolowskie) klasyfikujące dowolną próbkę w dowolny sposób. Dla dowolnej liczby N próbek, hipotezy zgodne z nimi dzielą się w równej liczbie na takie, które sklasyfikują próbkę x_{n+1} jako pozytywną jak i negatywną. Jeśli rozpatrujemy rzeczywiście dowolne funkcje boolowskie, to bez zbadania (prawie) wszystkich tych próbek trudno byłoby skutecznie nauczyć się takiej funkcji.

Jak z tego wynika, skuteczne uczenie się można uzyskać tylko w takim przypadku, gdy klasa rozważanych funkcji boolowskich (przestrzeń hipotez $\mathcal{H}$) będzie istotnie ograniczona. Ograniczenie tej klasy, z jednoczesnym zapewnieniem ogólnego warunku $\mathcal{C} \subseteq \mathcal{H}$, jest możliwe jedynie przy wykorzystaniu pewnych dodatkowych informacji o klasie $\mathcal{C}$.

Można wyróżnić kilka ogólnych schematów wykorzystania dodatkowej informacji o klasie $\mathcal{C}$ zaby umożliwić skuteczne uczenie się.

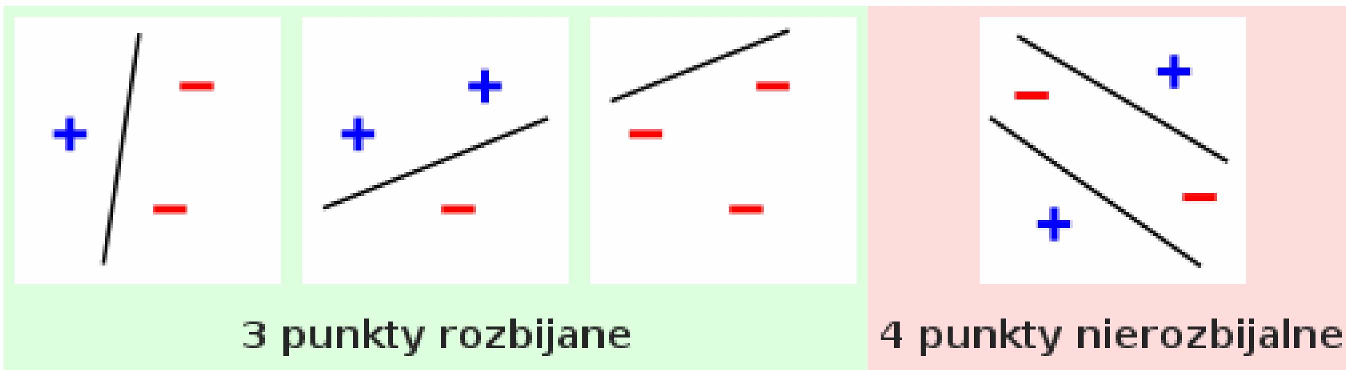
Jednym z tych schematów jest wymaganie aby hipoteza była prosta. Na przykład, w algorytmie uczenia drzew decyzyjnych, przyjęto heurystyczne założenie, że generowane są hipotezy minimalizujące entropię.

Inną metodą jest użycie zredukowanej. łatwiejszej do przeszukiwania przestrzeni hipotez, przy założeniu, że znaleziona tą drogą hipoteza będzie dostatecznie zbliżona do rzeczywistej funkcji klasy.

Jeszcze innym podejściem jest wykorzystanie posiadanej wiedzy od problemie.

Rozbijanie zbioru próbek

Jeśli funkcje ze zbioru hipotez $\mathcal{H}$ są zdolne podzielić zbiór P zawierający m próbek na wszystkie 2^m sposobów, niezależnie od sposobu etykietowania tych próbek, to mówimy, że $\mathcal{H}$ **rozbija** (ang. *shatters*) zbiór P.



Jak widać na powyższych obrazkach, każdy zbiór trzech punktów na płaszczyźnie może być rozbity prostą, ale nie dla każdego zbioru czterech punktów traktowanych jako próbki etykietowane istnieje prosta, która je poprawnie rozdziela. Można powiedzieć, że zbiór hipotez $\mathcal{H}$ składający się z prostych na płaszczyźnie rozbija każdy zbiór trzech próbek niewspółliniowych, ale nie rozbija zbioru czterech próbek (bo istnieją dla nich takie sposoby etykietowania, że nie da się ich rozdzielić).

Wymiar Wapnika-Czerwonenkisa

Chcemy oszacować zdolność danego algorytmu klasyfikacji do nauczenia się klasyfikacji różnych pojęć. Pewne algorytmy klasyfikacji mogą być zdolne do nauczenia się jednych pojęć, a niezdolne do nauczenia się innych.

Klasa klasyfikatorów f rozbija zbiór punktów $(x_1, x_2, ..., x_n)$ jeśli dla dowolnego przypisania etykiet tym punktom, istnieje klasyfikator z tej klasy, który prawidłowo klasyfikuje te punkty.

Wymiar Wapnika-Czerwonenkisa (VCdim, *Vapnik-Chervonenkis dimension*) klasy f jest największą liczbą punktów, którą klasa f rozbija.

Na przykład, klasa klasyfikatorów realizowanych za pomocą prostych w przestrzeni R^2 ma VCdim równy 3.

Wymiar Wapnika-Czerwonenkisa jest użyteczną miarą ekspresywności zbioru hipotez. Aby hipoteza spójna ze zbiorem treningowym miała szansę na dobrą generalizację, liczność zbioru treningowego musi znacznie przekraczać VCdim dla danej przestrzeni hipotez i określonego zagadnienia.

Związek wymiaru Wapnika-Czerwonenkisa z PAC-nauczalnością

Twierdzenie: Przestrzeń hipotez $\mathcal{H}$ jest PAC-nauczalna iff posiada skończony wymiar VCdim.

Twierdzenie: Zbiór hipotez $\mathcal{H}$ jest właściwie PAC-nauczalny jeśli:

1. $m \geq (\frac{1}{\epsilon}) \max[4 \lg(\frac{2}{\delta}), 8 \text{ VCdim } \lg(\frac{13}{\epsilon})]$,
2. istnieje algorytm generujący hipotezę $h \in \mathcal{H}$ spójną ze zbiorem treningowym w czasie wielomianowym (ze względu na m i n).