Gépi tanulás elmélete

Házi feladat – 2.3. feladat

Kidolgozta: Bódis Attila

1. Bizonyítsd be, hogy az A algoritmus ERM!

   Definíció szerint az A algoritmus jól címkézi a pozitív példákat S-ben, tehát ezekre a hiba 0. A realizálhatósági feltevés miatt a pozitív példák "között" nincs negatív példa, mert különben nem lehetne olyan hipotézis (téglalap), ami jól osztályoz. Mivel az A algoritmus a legkisebb téglalapot választja, amibe belefér az összes pozitív példa, ezért ebben a téglalapban nem lehet negatív, tehát ezekre a hiba szintén 0.

2. Bizonyítsd be, hogy $\frac{4\log\frac{4}{\delta}}{\epsilon}$ példa elég ahhoz, hogy legalább $1-\delta$ valószínűséggel a hiba legfeljebb ϵ legyen!

   Bizonyítás.

   - Mivel $R(S)$ a legkisebb olyan téglalap, amiben az összes pozitív példa benne van, ezért az R^*-nak szükségképpen ezt is le kell fednie, azaz tartalmaznia kell az $R(S)$ téglalapot. Tehát $R(S) \subseteq R^*$.
   - Ha S-ben vannak R_i-beli pozitív példák, akkor az $R(S)$ téglalap tartalmazza ezt a példát, tehát $R(S) \cap R_i \neq \emptyset$. Ebből és az R_i téglalapok definíciójából pedig adódik, hogy az R_i téglalapok uniója lefedi az A algoritmus hibatartományát, azaz az $R^* \setminus R(S)$ halmazt.

     Ezek alapján teljesül a következő:

     $$P(R^* \setminus R(S)) \leq P\left(\bigcup_{i\in\{1,2,3,4\}} R_i\right) \leq \sum_{i\in\{1,2,3,4\}} P(R_i) = 4\cdot\frac{\epsilon}{4} = \epsilon$$

     Tehát, ha az S tartalmaz példákat az R_1, R_2, R_3, R_4 téglalapok mindegyikében, akkor az A algoritmus hibája legfeljebb ϵ.
   - $P(S\text{-ben nincs } R_i\text{-beli példa}) \leq \prod_{j=1}^{m} P(\text{A } j \text{ példa nem } R_i\text{-ben van}) = \left(1-\frac{\epsilon}{4}\right)^m$
   - $P(\text{A hiba legfeljebb } \epsilon) = P(\forall i : S\text{-ben van } R_i\text{-beli példa}) = 1 - P(\exists i : S\text{-ben nincs } R_i\text{-beli példa}) \geq$
     $\geq 1 - \sum_{i\in\{1,2,3,4\}} P(S\text{-ben nincs } R_i\text{-beli példa}) \geq 1 - 4\left(1-\frac{\epsilon}{4}\right)^m$
     Ebből kapjuk a következőket:

     $$\begin{aligned}
     1-\delta &\leq 1-4\left(1-\frac{\epsilon}{4}\right)^m \\
     \delta &\geq 4\left(1-\frac{\epsilon}{4}\right)^m \\
     \delta &\geq 4e^{-\frac{\epsilon m}{4}} \qquad \text{// Itt felhasználjuk, hogy } 1-x \leq e^{-x} \\
     \frac{\delta}{4} &\geq \frac{1}{e^{\frac{\epsilon m}{4}}} \\
     \frac{4}{\delta} &\leq e^{\frac{\epsilon m}{4}} \\
     \log\frac{4}{\delta} &\leq \frac{\epsilon m}{4} \\
     \frac{4\log\frac{4}{\delta}}{\epsilon} &\leq m
     \end{aligned}$$

   □

3. Vizsgáld meg az előző kérdést d dimenzióban!

   Ha d dimenzióban nézzük a feladat, akkor annyi változik, hogy nem 4, hanem $2d$ "levágott" téglalapot kell létrehozni: $R_1, R_2, \ldots, R_{2d}$, melyek mindegyikének a valószíniűségi tömege pontosan $\frac{\epsilon}{2d}$

   Ekkor a korábbi megfontolások alapján kapjuk, hogy:

$$\begin{aligned}
1-\delta &\leq 1-2d\left(1-\frac{\epsilon}{2d}\right)^m \\
\delta &\geq 2d\left(1-\frac{\epsilon}{2d}\right)^m \\
\delta &\geq 2de^{-\frac{\epsilon m}{2d}} \\
\frac{\delta}{2d} &\geq \frac{1}{e^{\frac{\epsilon m}{2d}}} \\
\frac{2d}{\delta} &\leq e^{\frac{\epsilon m}{2d}} \\
\log\frac{2d}{\delta} &\leq \frac{\epsilon m}{2d} \\
\frac{2d\log\frac{2d}{\delta}}{\epsilon} &\leq m
\end{aligned}$$

   Tehát, ekkor $\frac{2d\log\frac{2d}{\delta}}{\epsilon}$ példa kell ahhoz, hogy legalább $1-\delta$ valószínűséggel legfeljebb ϵ legyen az A algoritmus hibája.

4. Mutasd meg, hogy az A algoritmus futásideje polinomiálisan függ a d, az $\frac{1}{\epsilon}$-tól és a $\log\frac{1}{\delta}$-tól!

   Az A algoritmus futásideje a példák számától és a dimenziószámtól függ, mert minden dimenzióban egy minimum és egy maximum keresést kell végrehajtani a pozitív példák között. Ez $\mathcal{O}(dm)$ számú műveletet igényel, tehát valóban polinomiálisan függ a d-től. Ugyanakkor a minták száma, vagyis az m, polinomiálisan függ az $\frac{1}{\epsilon}$-tól és a $\log\frac{1}{\delta}$-tól, ahogy azt korábban megmutattuk.

Megjegyzés:

Előfordulhat olyan eloszlás, amely esetén annak elérése, hogy az R_i téglalapok valószínűsége pontosan $\frac{\epsilon}{4}$ legyen nem lehetséges. Emiatt az a_1, a_2, b_1, b_2 értékeket inkább azon értékek infimumaként kell definiálni, amelyekre a megfelelő közbezárt rész valószínűsége legalább $\epsilon/4$. Ekkor a bizonyítás menete pontosan megegyezik a fentiekkel.