データマイニング工学
第2回レポート

濱崎 直紀
（学籍番号：28G19096）

令和2年1月28日

問題 1

正例包絡では，訓練データが全て正しく判別できるように境界を決めるため，訓練データに対する精度は基本的に 100% となる．よって，正例包絡において訓練データに対する性能が検証データに対する性能より悪くなることはない．

しかし，一般的な機械学習では必ずしも訓練データ全てを正しく判別できるように学習するわけではなく，外れ値を無視するなどの計算が行われる．よって一般的な機械学習では，訓練データに対する性能が検証データに対する性能より悪くなるとは言い切れない．

問題 2

C^* について以下が成り立つ

$$C^* = \hat{C}_{fit} \cup \bigcup_{i=1}^{4} A_i \setminus \{(A_1 \cap A_2) \cup (A_2 \cap A_3) \cup (A_3 \cap A_4) \cup (A_4 \cap A_1)\}$$

よって

$$C^* \subseteq \hat{C}_{\mathrm{fit}} \cup \bigcup_{i=1}^{4} A_i$$

なので

$$C^* \setminus \hat{C}_{\mathrm{fit}} \subseteq \bigcup_{i=1}^{4} A_i$$

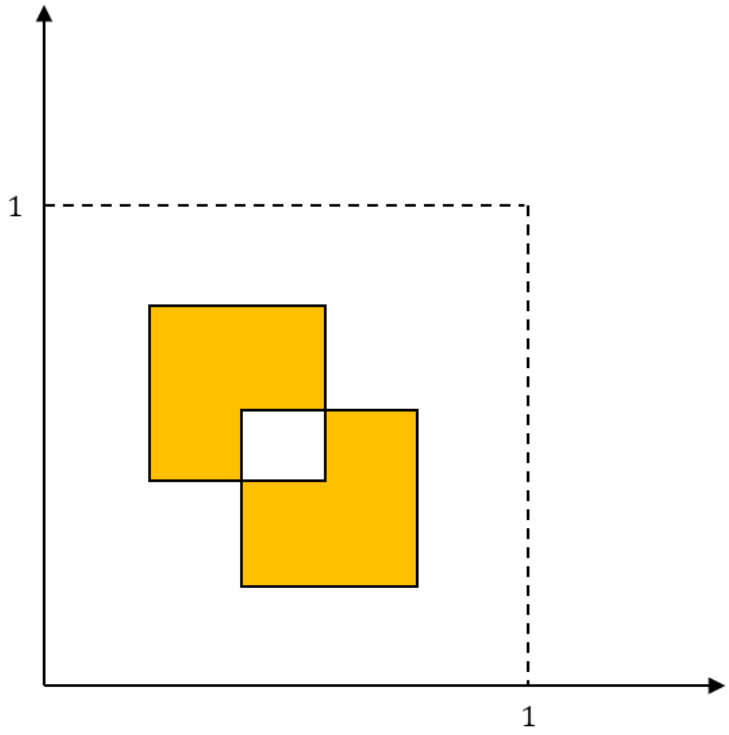
が成り立つ

問題 3

まず，$B_i \subset A_i \Rightarrow \boldsymbol{x}_j \notin B_i,\ \text{for all } j = 1, \ldots, n$ について証明する．
$B_i \subset A_i$ のとき，$\boldsymbol{x}_j \in B_i$ となる $\boldsymbol{x}_j$ が存在すると仮定する．
上式から $\boldsymbol{x}_j \subset A_i$ となる．
また，$\boldsymbol{x}_j \subset \hat{C}_{\mathrm{fit}},\ \text{for all } j = 1, \ldots, n$ であるから A_i と $\hat{C}_{\mathrm{fit}}$ は一部重複することになる．
しかし定義より，A_i と $\hat{C}_{\mathrm{fit}}$ は互いに排反であるのでこれに矛盾する．
よって仮定が間違っていることから，$B_i \subset A_i \Rightarrow \boldsymbol{x}_j \notin B_i,\ \text{for all } j = 1, \ldots, n$ が示された．
次に，$B_i \subset A_i \Leftarrow \boldsymbol{x}_j \notin B_i,\ \text{for all } j = 1, \ldots, n$ について証明する．
$\boldsymbol{x}_j \notin B_i,\ \text{for all } j = 1, \ldots, n$ より，B_i と $\hat{C}_{\mathrm{fit}}$ は排反である．
さらに $B_i \subset C^*$ より，$B_i \subset A_i$ となる．
よって，$B_i \subset A_i \Leftarrow \boldsymbol{x}_j \notin B_i,\ \text{for all } j = 1, \ldots, n$ が示された．
ゆえに，$B_i \subset A_i \Leftrightarrow \boldsymbol{x}_j \notin B_i,\ \text{for all } j = 1, \ldots, n$ が示された．

問題 4

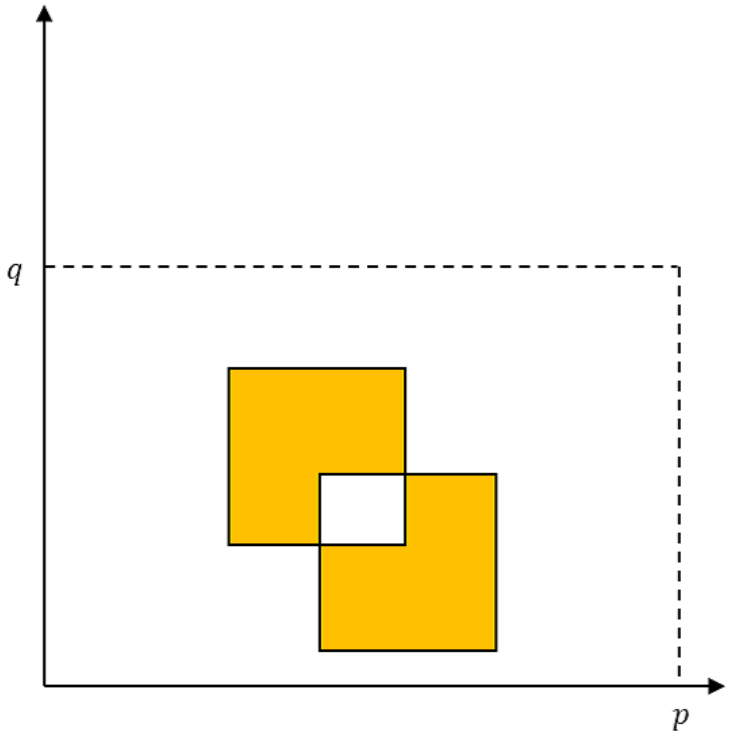
データが単位正方形の上で，一様分布に従うとすると，$R(C)$ は図の着色部分の面積に等しい．



次に，任意の長方形の上で一様に分布している場合の $R(C)$ について考える．
図のように任意の長方形の横と縦の長さをそれぞれ p, q とおき，着色部分の面積を S とおくと

$$R(C) = \frac{S}{pq}$$

となる．


q
p