実データで学ぶ人工知能講座
演習課題：rectangles

Matthew J. Holland*
大阪大学 産業科学研究所

問題 1. データに関する複数の問題.

A. このデータを生成する関数名は何か.

B. その関数はどこで定義されているか.

C. 入力データの確率分布は一様分布である．それがわかる箇所はどこか．また，一様分布の範囲はどこからどこまでか.

D. そのファイルを開き，学習用の標本数を元の半分・二倍にして，それぞれのケースで再実行してみること.

問題 2. 正例包絡のアルゴリズムに関する複数の問題.

A. アルゴリズムのクラス名は何か.

B. そのクラスはどこで定義されているか.

C. 具体的な手順は単純なので，それをコード内の変数名を引用しながら，自分の言葉で説明すること.

問題 3. 学習用のデータでの成績と，検証用のデータでの成績の差（識別率の差）を計算すること．正例包絡アルゴリズムを使った場合，前者が後者より悪くなることはあり得るか．別のアルゴリズムを使った場合はどうか．その理由も説明すること.

問題 4. 入力データが一様分布に従うなら，事象 $\{X \in A\}$ の確率は，A の面積を分布の範囲全体の面積で割ったものになる．この演習はシミュレーションなので C^* は自ら設定している．これを踏まえて，スライド中に扱った誤差 $R(C)$ をどのように計算すれば良いか.

問題 5. 前の問題の解答を踏まえて，正例学習アルゴリズムの期待損失，つまり $R(\widehat{C}_{\text{fit}})$ を計算すること.

問題 6. スライド中に，学習用のデータが $n = 100$ のとき，正例包絡の期待損失について

$$\mathbf{P}\left\{R(\widehat{C}_{\text{fit}}) > 0.1\right\} \leq 4\left(1 - \frac{0.1}{4}\right)^n$$

*作者の連絡先：matthew-h@ar.sanken.osaka-u.ac.jp.

という上界を導き出している．このバウンドの正しさと緩さを調べること．たとえば，$n = 100$ 個の標本データを $T = 100000$ 回ほど独立に生成し，それぞれの正例包絡アルゴリズムの成績 $R(\widehat{C}_{\mathrm{fit}})$ を記録する．その結果として得られる T 個の成績値の分布 (分位値など) を見れば，知りたい確率の良い近似が得られる．理論値と比較して，上界がきちんと成り立っているか．また，その上界の緩さについて考察すること．