Jensen-Shannon ダイバージェンスを用いた分布に対する回帰樹，クラスタリング，多重比較

慶應義塾大学理工学部　南　美穂子
全米熱帯マグロ類委員会 Cleridy E. Lennert-Cody

本発表では，修正 Jensen-Shannon ダイバージェンスを用いた，分布を目的変数とする回帰樹，分布を分類するクラスタリング，分布間の多重比較法を提案し，その性質について議論する．提案する回帰樹は，情報エントロピー最小化を分割基準とする通常の分類樹の自然な拡張であり，提案するクラスタリング法と統一的な解釈ができる．解析対象の分布は，頻度関数，あるいは，標本から求めたカーネル密度推定関数を想定している．解析例として東部太平洋の $5° \times 5°$ の各小海域で捕獲されたキハダマグロの体長データに基づいた海域のクラスタリングと，前立腺ガン患者と対照群の遺伝子発現量の多重比較における偽陽性率 (FDR) の特徴を示す．

【修正 Jensen-Shannon ダイバージェンス】　信頼度 (例えば標本サイズ）m_1, m_2 を持つ 2 つの分布 f_1, f_2 の修正 Jensen-Shannon ダイバージェンスを以下で定義する．

$$\mathrm{JS}\left((f_1, m_1), (f_2, m_2)\right) = m_1 \mathrm{KL}(f_1; \bar{f}_{1,2}) + m_2 \mathrm{KL}(f_2; \bar{f}_{1,2}) \tag{1}$$

ここで $\bar{f}_{1,2} = \dfrac{m_1}{m_1 + m_2} f_1(x) + \dfrac{m_2}{m_1 + m_2} f_2(x)$ とし，$\mathrm{KL}(\cdot;\cdot)$ は Kullback-Leibler ダイバージェンスを表す．

【情報エントロピー最小化による分類樹】分類樹は，説明変数の値を基準にした 2 分割を繰り返して得られる木構造による分類手法で，2 分割のルールは目的変数の値から計算される不純度を基準にして選択される．頻度分布のデビアンスを不純度としたとき，分割は情報エントロピーを最小化するものを選択することになるが，分割による不純度の減少は，２つの子ノードにおける頻度分布間の修正 Jensen-Shannon ダイバージェンスとしても表現できる．

【分布に対する回帰樹】反応変量が頻度分布，あるいは，母集団からの標本である場合の回帰樹を考える．標本の場合は，頻度分布に変換するかカーネル密度推定関数を求めるものとし，各ユニット $i \in \mathcal{G}$ に対して確率/密度関数 f_i と確信度（標本サイズ）m_i が与えられるとする．このとき，不純度を

$$\mathrm{IMP}_{\mathrm{KL}}(\mathcal{G}) = \sum_{i \in \mathcal{G}} m_i \mathrm{KL}(f_i | \bar{f}(\mathcal{G})) \quad \text{ここで} \quad \bar{f}(\mathcal{G}) = \frac{1}{\sum_{i \in \mathcal{G}} m_i} \sum_{i \in \mathcal{G}} m_i f_i(x) \tag{2}$$

と定めると，分割による不純度の減少は平均分布の情報エントロピーの減少として表され，これは 2 つの子ノードの平均分布間の修正 Jensen-Shannon ダイバージェンスとしても表される (Lennert-Cody et al. 2013).

【分布に対する階層的クラスタリング】各ユニットを確率/密度関数 f_i と確信度（標本サイズ）m_i の組とし，各ユニットが 1 つのクラスターを形成する状態からスタートして距離の最も近いクラスター同士の結合を繰り返す階層的クラスタリングを考える．分布間の距離として修正 Jensen-Shannon ダイバージェンスを用い，クラスターの不純度として (2) を用いると，結合による不純度の増加は 2 つのクラスターの平均分布間の修正 Jensen-Shannon ダイバージェンスで表せる．つまり，クラスター間の距離を平均分布間の修正 Jensen-Shannon ダイバージェンスとしたとき，この階層クラスタリング法は不純度の増加が最小になるような結合を逐次的に選択する方法であることがわかる．

【分布間距離を用いた遺伝子発現量の多重比較】患者群と対照群の遺伝子発現量に相違があるかどうかの多重比較問題を考える．比較方法には，t 統計量を用いた平均値の差の比較，罹患者の一部にでも高発現がある遺伝子を探索する COPA 統計量（Tomlins et al. 2005）や OS 統計量（Tibshirani and Hastie, 2007) による比較，また，Earth Movers Distance で定義した分布間距離による方法 (Nabavi et al. 2016) などがあるが，本研究では修正 Jensen-Shannon ダイバージェンスを用いた方法を提案し，これらの方法との偽陽性率 (FDR) の違いや統計量間の関係性などについて議論する．