データサイエンス研究室

情報数理分野

データサイエンス研究室では，さまざまなデータから有用な情報を抽出する手法を，統計科学，機械学習，情報理論，情報幾何学の視点から開発しています．またその手法を用いて実社会のさまざまな予測問題に取り組んでいます．

生息分布の予測モデル構築

推定されたクマの生息分布

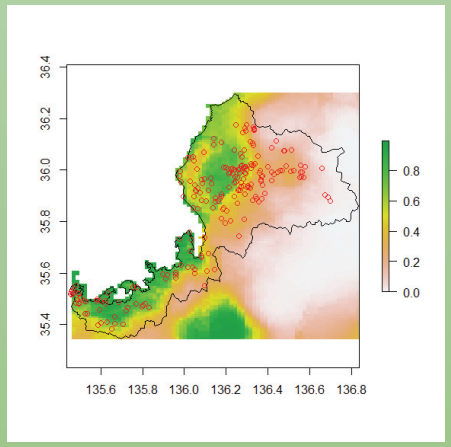


表 6-3.ツキノワグマの生息分布に相関のある環境変数(反復 30 回)

	ツキノワグマに使用する環境変数(赤字は相関あり、太字強い相関)
f_1	年平均気温
f_2	月平均気温の標準偏差
f_3	最寒月平均最低気温
f_4	気温のレンジ(最暖月最高気温 - 最寒月平均最低気温)
f_5	年平均降水量
f_6	月平均降水量の変動係数
f_7	最も暖かい 3 か月間の合計降水量
f_8	年最大積雪深
f_9	陸地面積
f_{10}	メッシュ内の道路総延長
f_{11}	メッシュ内の道路密度
f_{12}	人口密度

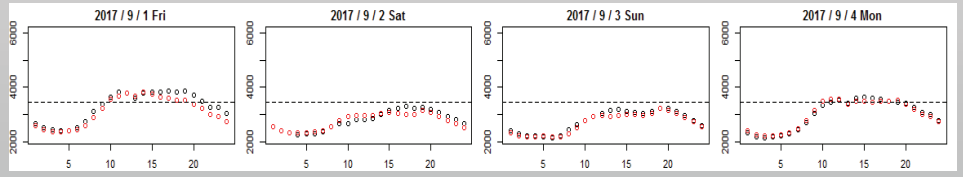
予測に使う環境変数

統計モデル

$$p(x) = \frac{\exp\{\sum_{j=1}^{m} \lambda_j f_j(x)\}}{Z_\lambda},$$

$$\hat{\lambda} = (3.46, 0, 0, 0.94, 0.14, -0.35, -4.01, 0, 0, 0, 0, 0)^\top$$

電力需要予測モデルの構築



赤丸：実測値，黒丸：予測値

医療データへの機械学習法の応用

モデル構築アルゴリズム

LogitBoost(二値判別)

1. 重み w の初期値を $w_i = \frac{1}{n}$ と置く．強分類器 $F(x)$ の初期値は 0，及び確率 $p(x)$ は要素すべて $\frac{1}{2}$ と置く．

2. $m = 1, 2, \ldots, M$ に対して，以下の (a), (b), (c) を繰り返す．

(a) 作業目的変数 z_i と重み w_i を以下の式に従い計算する．

$$z_i^{(m)} = \frac{y_i - p^{(m-1)}(x_i)}{p^{(m-1)}(x_i)(1 - p^{(m-1)}(x_i))} \quad (3.1)$$

$$w_i^{(m)} = p^{(m-1)}(x_i)(1 - p^{(m-1)}(x_i)) \quad (3.2)$$

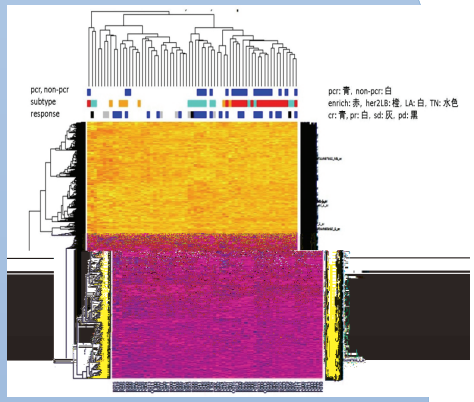
(b) 弱分類器 $f^{(m)}$ を z_i に対する重み付き最小二乗回帰でフィッティングする．

$$f^{(m)} = \underset{f}{\operatorname{argmin}} \sum_{i=1}^{n} w_i^{(m)} (z_i^{(m)} - f(x_i))^2 \quad (3.3)$$

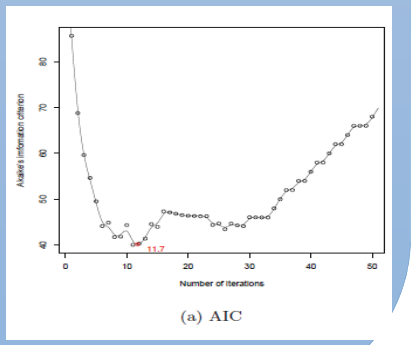
(c) $F^{(m)}(x_i)$ を $F^{(m)}(x_i) \leftarrow F^{(m-1)}(x_i) + \frac{1}{2} f^{(m)}(x_i)$ に従って更新する．続いて $p^{(m)}(x_i)$ も $p^{(m)}(x_i) \leftarrow \frac{\exp(F^{(m)}(x_i))}{\exp(F^{(m)}(x_i)) + \exp(-F^{(m)}(x_i))}$ と更新する．

3. 結果を以下の式に従い出力する．

$$\hat{F}(x_i) = \sum_{m=1}^{M} f^{(m)}(x_i)$$



遺伝子発現量のクラスタリング解析



(a) AIC

情報量基準によるモデル選択

准教授　小森　理