「データ分析プロセス」　正誤表

- p.9 本文 最終行

  (誤) 適宣活用してほしい.

  (正) 適宜活用してほしい.

- p.58 (3.6) 式

  (誤)

$$f(\boldsymbol{x}_i|\boldsymbol{\mu},\, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

  (正)

$$f(\boldsymbol{x}_i|\boldsymbol{\mu},\, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})\right)$$

- p.61 本文 下から 4 行

  (誤) ここで，箱ひげ図の見方は，図 3.19 のとおりである．この箱ひげ図の見方に従うと，Sepal.Width が 4.4, 4.1, 4.2 のデータは，中央値+1.5（第三四分位点− 中央値）よりも大きく，Sepal.Width が 2.0 の点は，中央値− 1.5（中央値− 第一四分位点）よりも小さいことがわかる．これらは，`boxplot` 関数の返り値 out に格納されていることが確認できる．

  (正) ここで，箱ひげ図の見方は，図 3.19 のとおりである．この箱ひげ図の見方に従うと，Sepal.Width が 4.4, 4.1, 4.2 のデータは，第三四分位点+1.5（第三四分位点−第一四分位点）よりも大きく，Sepal.Width が 2.0 の点は，第一四分位点− 1.5（第三四分位点−第一四分位点）よりも小さいことがわかる．これらは，`boxplot` 関数の返り値 out に格納されていることが確認できる．

- p.62 図 3.19
  以下の図に差し替え

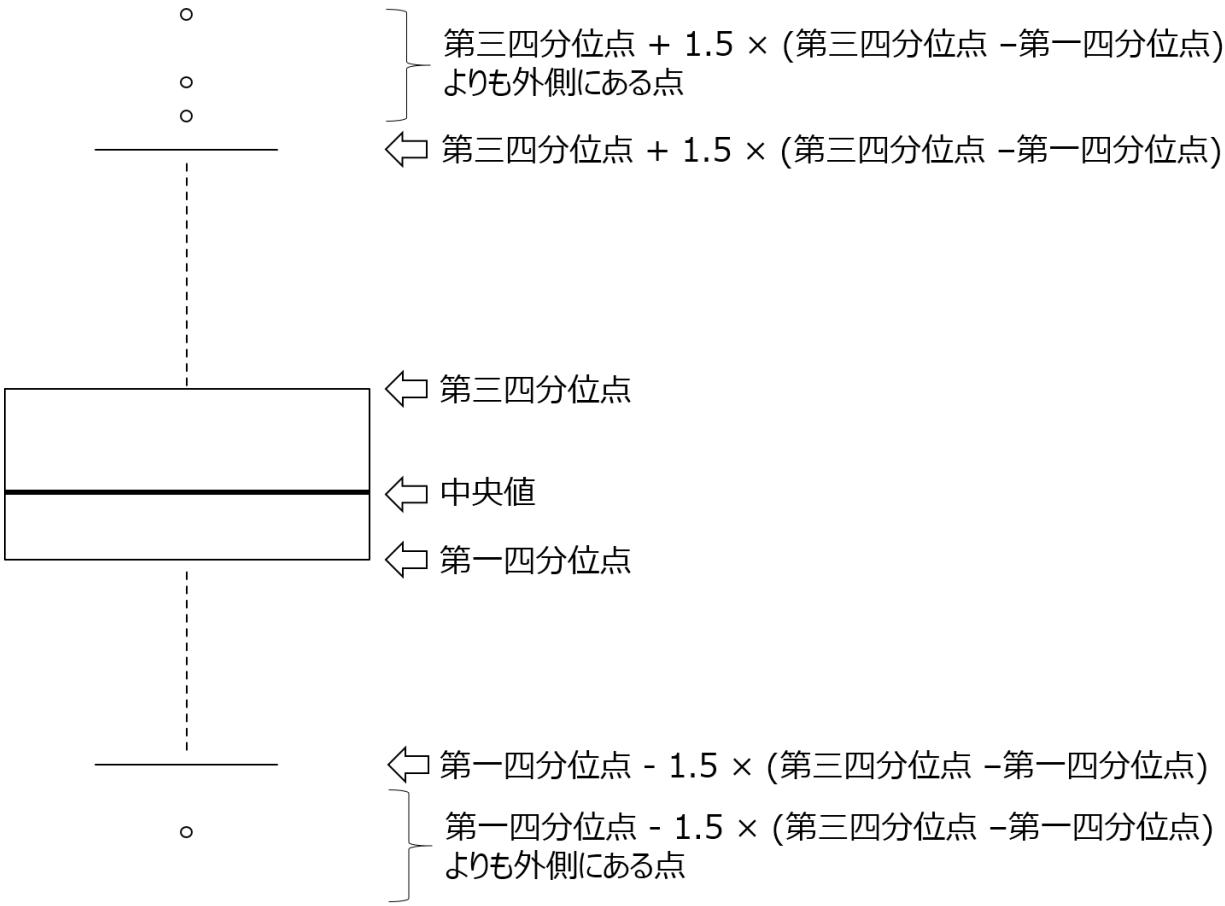

図 3.19: 箱ひげ図の見方

- p.103 2 ブロック目のプログラム

(誤)

```
> library(pROC)
> # テストデータのクラスラベル・クラス確率の予測
> pred <- predict(fit.rf, churnTest)
> prob <- predict(fit.rf, churnTest, type = "prob")
> # ROC曲線のプロット（roc関数のlevels引数は興味のあるクラスが2番目にあるという前提のもとに実装されているため，逆順に並び替え）
> rocCurve <- roc(response = pred, predictor = prob$yes, levels = rev(levels(pred)))
> plot(rocCurve, legacy.axes = TRUE)


Call:
roc.default(response = pred, predictor = prob$yes, levels = rev(levels(pred)))

Data: prob$yes in 1562 controls (pred no) < 105 cases (pred yes).
Area under the curve: 1

> # AUC
> auc(rocCurve)

Area under the curve: 1

> # 信頼区間
> ci.auc(rocCurve)

95% CI: 1-1 (DeLong)
```

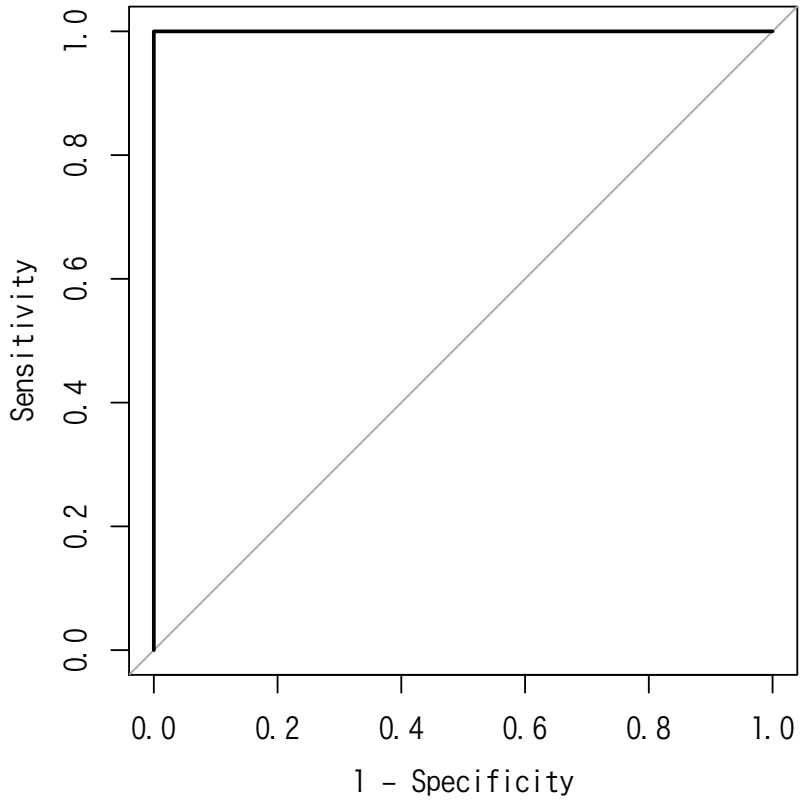

図 4.6: ROC 曲線のプロット

(正)

```
> library(pROC)
> # テストデータのクラスラベル・クラス確率の予測
> label <- churnTest$churn
> prob <- predict(fit.rf, churnTest, type = "prob")
> # ROC曲線のプロット（roc関数のlevels引数は興味のあるクラスが2番目にあるという前提のもとに実装されているため，逆順に並び替え）
> rocCurve <- roc(response = label, predictor = prob$yes, levels = rev(levels(label)))
> plot(rocCurve, legacy.axes = TRUE)


Call:
roc.default(response = label, predictor = prob$yes, levels = rev(levels(label)))

Data: prob$yes in 1443 controls (label no) < 224 cases (label yes).
Area under the curve: 0.9249

> # AUC
> auc(rocCurve)

Area under the curve: 0.9249

> # 信頼区間
> ci.auc(rocCurve)

95% CI: 0.8983-0.9515 (DeLong)
```

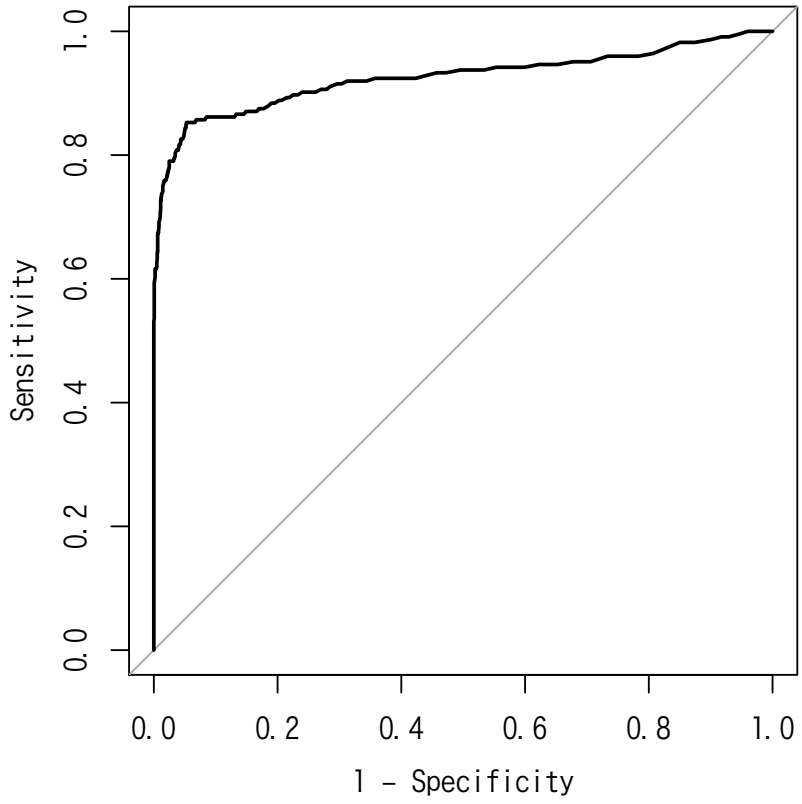

図 4.6: ROC 曲線のプロット

- p.104 本文 1–2 行目

  (誤) 図 4.6 を見ると，ROC 曲線がプロットされていることを確認できる．この場合は予測が完全に当たっているため ROC 曲線は直角になり，AUC は 1 となる．

  (正) 図 4.6 を見ると，ROC 曲線がプロットされていることを確認できる．この場合は，AUC は 0.9249 となっている．

- p.125 1 番目のソースコードブロック 1–2 行目

  (誤)

```
> # KDDCUP2009 のサイトのアドレス
> url <- "http://www.sigkdd.org/sites/default/files/kddcup/site/2009/files"
```

  (正)

```
> # KDDCUP2009 のサイトのアドレス
> url <- "http://kdd.org/cupfiles/KDDCupData/2009"
```

  SIGKDD のウェブアドレスが執筆時点から変更になったことに伴う修正になります．

- p.140 本文 4 行目

  (誤) 使用するデータは，中国の食料品店において 4 カ月に渡って収集した POS データである `Tafeng` データセットとする．

  (正) 使用するデータは，台湾のスーパーマーケットにおいて 4 カ月に渡って収集した POS データである `Tafeng` データセットとする．

- p.209 参考文献

  (誤) [4] KDD Cup 2009: Customer relationship prediction, http://www.sigkdd.org/kdd-cup-2009-customer-relationship-prediction 2009.

  (正) [4] KDD Cup 2009: Customer relationship prediction, http://kdd.org/kdd-cup/view/kdd-cup-2009 2009.

  SIGKDD のウェブアドレスが執筆時点から変更になったことに伴う修正になります．