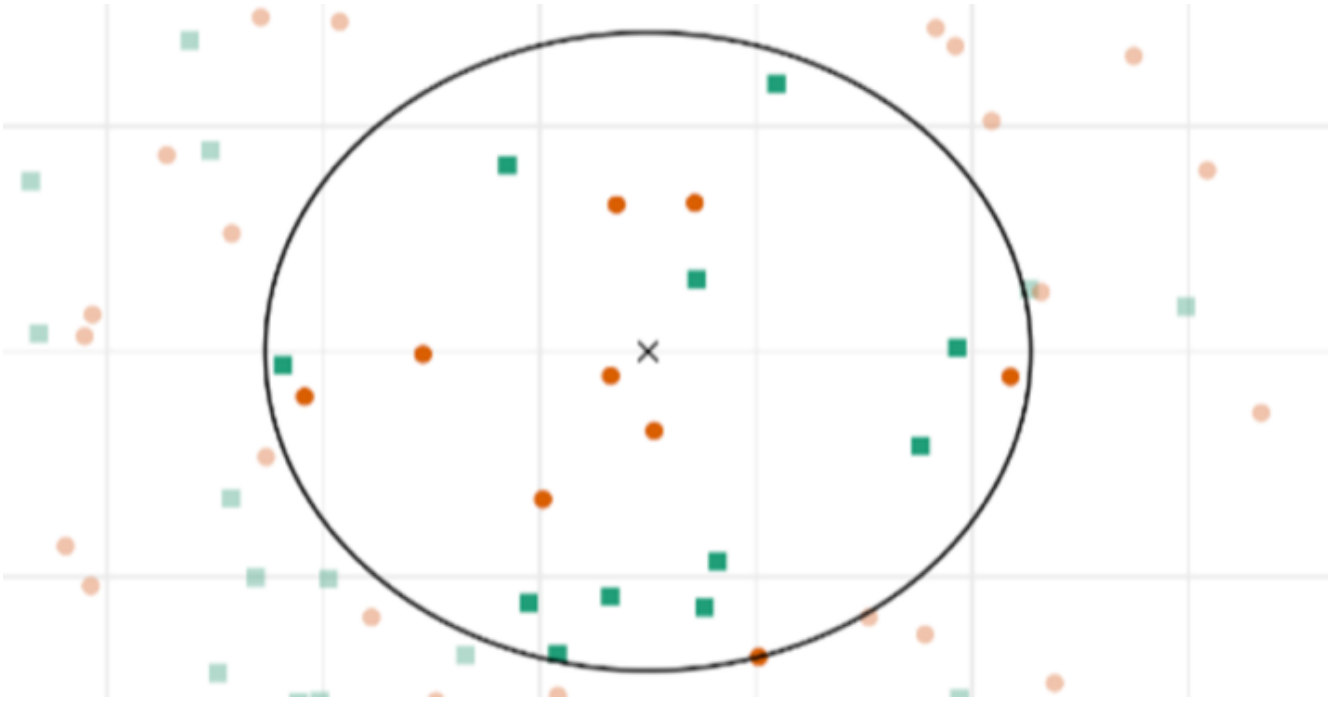

統計的機械学習

Yutaka MOTEKI
2023年2月16日 15:36

「収入に対する（抵当権を除いた）債務支払い比dtiと、収入に対するローン支払い比payment_inc_ratioの2つの予測変数だけの非常に単純なモデルを考える。」と書いてある（**注1**）。これでは意味が分からない。「抵当権を除いた」ということがどう意味なのだろうか思って原文を調べてみると、抵当権ではなく「住宅ローン」であった。「収入に対する（住宅ローンを除いた）債務支払い比」ということであれば理解できる。

```
newloan <- loan200[1, 2:3, drop=FALSE]

knn_pred <- knn(train=loan200[-1, 2:3], test=newloan,
 cl=loan200[-1, 1], k=20)

knn_pred == 'paid off'

knn_pred == 'default'
```

これを実行すると、以下のようになる。

```
> newloan <- loan200[1, 2:3, drop=FALSE]
> knn_pred <- knn(train=loan200[-1, 2:3], test=newloan, cl=loan200
[-1, 1], k=20)
> knn_pred == 'paid off'
[1] TRUE
> knn_pred == 'default'
[1] FALSE
```

本では、「knn_pred=="paid off"」を実行していて、「ローンに対するKNN予測は返済不能だ」と書いてある。これはおかしい。結果が「TRUE」というのだから「完済（paid off）」という予測になるのではないか。この誤りは、原文に原因があるようだ。

k=20ではなく、k=10と設定してみると次のようになる。この結果についてであるならば、「返済不能だ」という説明は正しい。

```
> newloan <- loan200[1, 2:3, drop=FALSE]
> knn_pred <- knn(train=loan200[-1, 2:3], test=newloan, cl=loan200[-1, 1],
```

```
k=10)
> knn_pred == 'paid off'
[1] FALSE
> knn_pred == 'default'
[1] TRUE
```

「kの選択」について、「通常、k値は1から20の範囲になる。同順位を避けるため、kを奇数にとることも多い」(**注2**)と書いてある。 奇数でも偶数でも同順位の事態は生じるのではないだろうか。この部分は、原文の誤訳ではない。

「数値を予測しているので、（**KNN回帰という**）**多数決の代わりに**、k近傍の平均値が使われる」という文がある(**注3**)。内容を考えてみると、「多数決の代わりにk近傍の平均値が使われる(KNN回帰と呼ばれる)」と訳した方がよいだろう。「k近傍の平均値が使われる」というところはかまわないが、KNN回帰が多数決を使うということではない。「（KNN回帰という）多数決の代わりに、」という部分がおかしい。訳者は、「KNN回帰と呼ばれている(known as KNN regression)」という言葉を、文全体の意味を考えずに直前の「多数決(majority vote)」という言葉に結びつけてしまったのだろう。訳文では、「KNN回帰という」は、すぐ後の「多数決」を修飾しているとしか受け取れないので誤訳であろう。

[注]

(1) 『データサイエンスのための統計学入門・第2版』（オライリー・ジャパン）

の250ページ。

(2) 『同上書』、257ページ。

(3) 『同上書』、258ページ。

[補足]

『データサイエンスのための統計学入門・第2版』（オライリー・ジャパン）の第6章「統計的機械学習」は、k近傍法、木モデル、バギングとランダムフォレスト、ブースティングを扱っている。最初の2項目は何とか理解できたが、その後の部分は今の私には手に負えないようなので、検討を中止する。負け惜しみのようなものだが、誰がローンの返済不能になるかを予測するプログラムとかいうのが「機械学習」の応用分野だとしたら関わりたくない気もする。（半分冗談）