House Price Regression

Sebastian Öhman

April 10, 2024

1 Introduktion

I denna uppgift ska patienter klassificeras om. Klasserna är om Patienten har fått en hjärtsjukdom eller inte. Notera att detta inte är en inlämning och det kommer ej ges någon rättning. Däremot kan du fråga om hjälp, om du tänkt rätt samt dela ditt resultat. Uppgiften är uppdelad i en godkänt del och en väl godkänt del.

De obligatoriska inlämningarna som kommer framöver kommer vara uppdelade på liknande vis. Första delen kommer mestadels vara att implementera algoritmen och andra delen kommer handla om djupare förståelse.

För hela uppgiften skall hearth.csv användas. Datasetet innehåller en hel del medicinska termer. Nästan aldrig har man full kunskap om vad ens data innehåller från början, utan man får undersöka vad olika attribut faktiskt betyder. Antingen från person med domänkunskap eller med hjälp av en sökmotor.

2 Godkänt:

Använd datasetet area_price.csv för att förutsäga huspriser. Följande punkter skall behandlas och besvaras:

- Analysera datan
- Se till att du hyfsat förstår vad de olika attributen innebär
- Dela upp datan i träning och testset
- Träna en logistisk regression
- Vad får du för träningsscore?
- Vad får du för testscore?

3 Väl Godkänt:

Följande punkter skall behandlas och besvaras

- Djupare dataanalys
- Ålder och kön brukar ses om en vanlig koppling till hjärtsjukdomar. Hur ser fördelningen på hjärtsjukdomar ut baserat på kön och ålder?
- Vilka attribut påverkar modellen mest?
- Ett mått för modellen är Accuarcy, men en annan viktig del är vilka fel modellen gör.
- Skapa en confusion matrix för att se vilken typ av fel modellen gör
- Finns det något du kan ändra för att det ska ändras vilken typ av fel modellen gör? (Testa att göra ändringarna och se vad som händer)

4 Dataset Info

This data set dates from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 of them. The "target" field refers to the presence of heart disease in the patient. It is integer valued 0 = no disease and 1 = disease.

Attribut

1. age - age in years
2. sex - (1 = male; 0 = female)
3. cp - chest pain type
4. trestbps - resting blood pressure (in mm Hg on admission to the hospital)
5. chol - serum cholestoral in mg/dl
6. fbs - fasting blood sugar ¿ 120 mg/dl (1 = true; 0 = false)
7. restecg - resting electrocardiographic results
8. thalach - maximum heart rate achieved
9. exang - exercise induced angina (1 = yes; 0 = no)
10. oldpeak = ST depression induced by exercise relative to rest
11. slope - the slope of the peak exercise ST segment
12. number of major vessels (0-3) colored by flourosopy
13. thal: 0 = normal; 1 = fixed defect; 2 = reversable defect