Problemes APA
Problema 11: Pràctica amb la xarxa MLP

Lluc Bové

Q1 2016-17

Es disposa de 48 mesures de roques d'un dipòsit de petroli. L'objectiu és modelar la permeabilitat en funció de l'àrea, el perímetre i la forma. En primer lloc transformem les dades per ajudar a l'ajust del model:

```
library(datasets)
data(rock)
?rock

rock.x <- data.frame(area = scale(rock$area), peri = scale(rock$peri),shape = scale(rock$shape))
rock.y <- log(rock$perm)
```

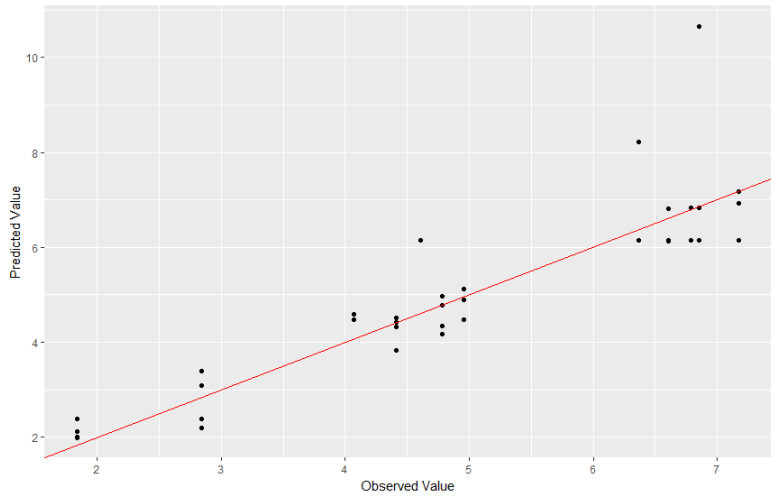
Entreneu una xarxa *MLP* per aprendre la tasca. Donat el baix número d'exemples, useu *leave-one-out cross-validation* i regularització per trobar la millor xarxa. Per avaluar el model, feu una gràfica de resposta predita vs. observbada i guieu-vos per l'error quadràtic normalitzat (normalized root MSE) del laboratori 3.

Primer hem de decidir quantes dades seràn de test i quantes de training. Decidim usar un 75% de les dades per training i la resta per test. Per a selaccionar una mostra aleatòria fem servir la funció sample. També definim la funció norm.mse que és l'error quadràtic normalitzat que es defineix com a:

$$\frac{\sum_{i=0}^{N} t_i - y(x_i)}{(N-1)Var(t)}$$

on t són les mostres de la variable que volem predir, x la resta de variables, N la mida de les dades, y el model en qüestio i Var és la variança mostral.

Primer creem una xarxa neuronal sense regularització i amb 4 neurones a la capa oculta. Veiem que obtenim valors del 15% per l'error de training i 61% pel de test. Probablament estiguem sobreajustant. Així doncs apliquem regularització, provem per valors de λ de 10^{-5} fins a 0, en increments de 0.1. Ho fem aplicant *LOOCV*. Trobem que la millor λ és de 0.063 i calculem l'error amb un model que la utilitzi. Trobem un error de 9% i 17% per els errors de training i test respectivament. Veiem com hem aconseguit reduir l'error de test. Finalment representem el valor de permeabilitat observat respecte el que predim amb el model, i obtenim el següent gràfic:



La línia vermella representa la recta on haurien d'estar situats els punts si la predicció fos perfecta. En tot cas veiem que alguns punts no es desvien gaire i tenen poca variabilitat però alguns punts la predicció no és acurada. El model és limitat per el baix nombre de dades.