Data Mining - Prova d'esame del 18.11.2019

Laboratorio: punteggi e soluzione

Punteggi

files	RMSE	Percentuale	Punti
779608.txt	2.691290	59.9%	4.8
779824.txt	2.882319	50.4%	4.0
788966.txt	2.852276	51.9%	4.2
790677.txt	2.984072	45.3%	3.6
791517.txt	2.756433	56.7%	4.5
796724.txt	2.925061	48.2%	3.9
800691.txt	1.828194	100%	8.0
800695.txt	1.773628	100%	8.0
801768.txt	2.917458	48.6%	3.9
803335.txt	2.885299	50.2%	4.0
804450.txt	3.767332	6.1%	0.5
805830.txt	2.890889	50%	4.0
805913.txt	2.810467	54%	4.3
807406.txt	2.524294	68.3%	5.5
807699.txt	2.532645	67.9%	5.4
807842.txt	2.771483	55.9%	4.5
808167.txt	2.752851	56.9%	4.5
808644.txt	3.446487	22.2%	1.8
823968.txt	2.754757	56.8%	4.5
836695.txt	6.685364	0%	0.0
848786.txt	2.688093	60.1%	4.8
849323.txt	2.265654	81.2%	6.5

La percentuale è calcolata come

$$\min\left(\frac{3.89-x}{3.89-1.89}, 100\%\right)$$

dove x rappresenta il proprio RMSE_{Te}. Se $\text{RMSE}_{\text{Te}} > 3.89$, la percentuale è 0%.

Per il modello di *benchmark*, $\text{RMSE}_{\text{Te}} = 2.89$, quindi la percentuale è 50%. Questo significa che percentuali inferiori al 50% peggiorano il modello di *benchmark*.

Soluzione

I $p = 10$ predittori $X = (X_1, \ldots, X_{10})^\mathsf{T}$ sono stati generati da una distribuzioen Uniforme(0,1).

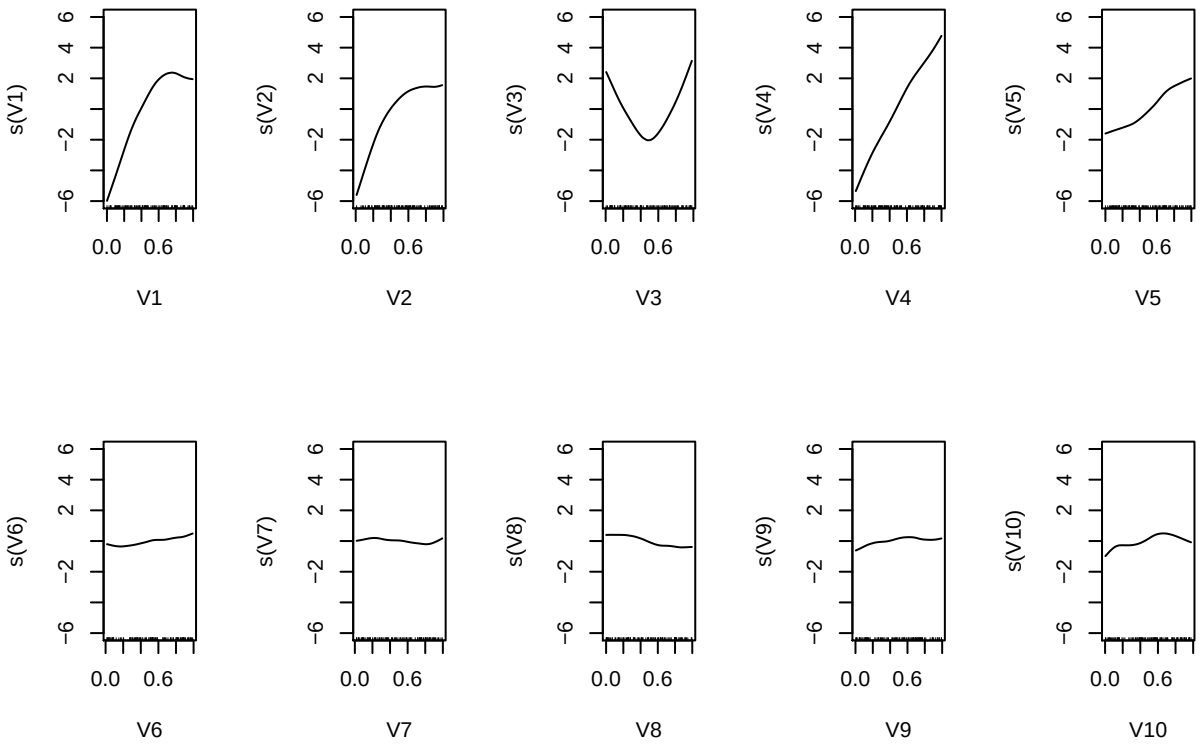
Solo 5 predittori sono stati utilizzati per determinare la variabile risposta $Y = f(X) + \varepsilon$ dove

$$f(X) = 10sin(\pi X_1 X_2) + 20(X_3 - 0.5)^2 + 10X_4 + 5X_5$$

e $\varepsilon \sim N(0, 1)$.

L'assenza di relazione tra Y e X_6, X_7, X_8, X_9, X_{10} e la la relazione non-lineare tra Y e X_1, X_2 e X_3 si poteva notare con l'esplorazione grafica utilizzando un modello addittivo (package gam):

```
library(gam)
fml = as.formula(paste("y ~ ", paste("s(",paste("V",1:10, sep=""),")", collapse = "+")))
fit = gam::gam(fml, data=train, family="gaussian")
op <- par(mfrow = c(2, 5))
plot(fit, ylim=c(-6,6))
```



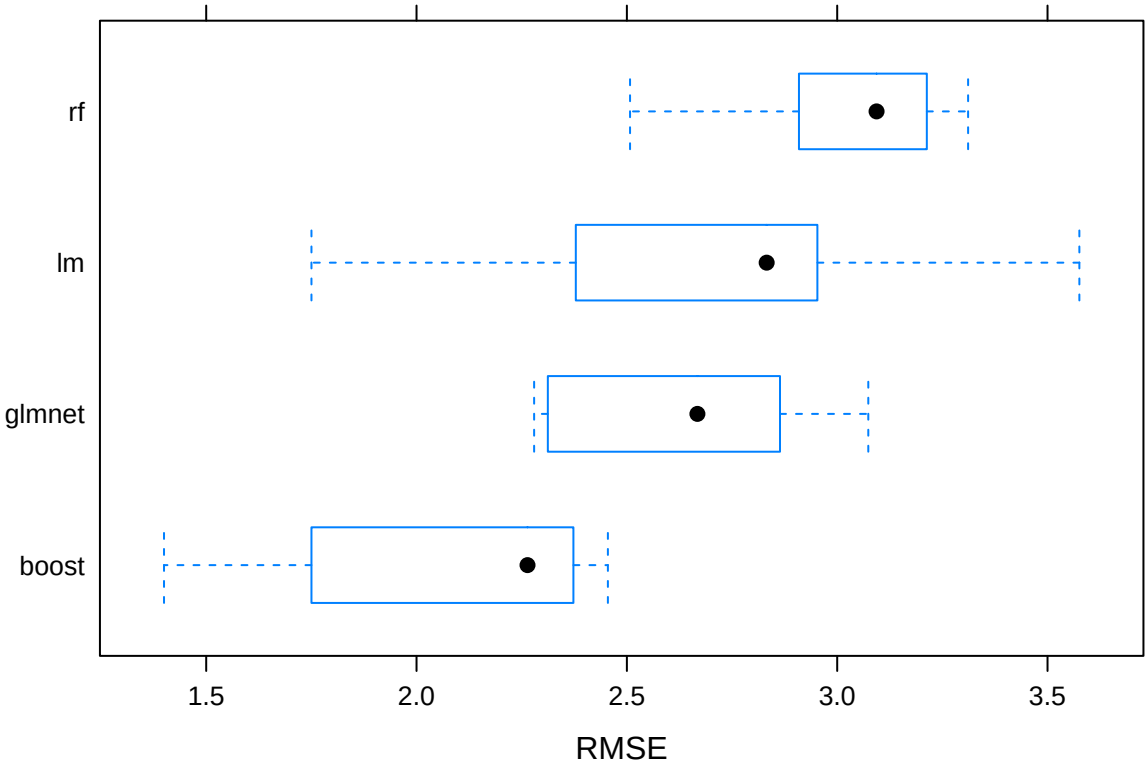

```
par(op)
```

Un buon modello si poteva quindi ottenere utilizzando un GAM

```
fit.gam = gam::gam(fml, data=train)
yhat = predict(fit.gam, newdata = test)
RMSE = sqrt( mean( (yhat - testy)^2) )
RMSE
```

```
[1] 1.894177
```

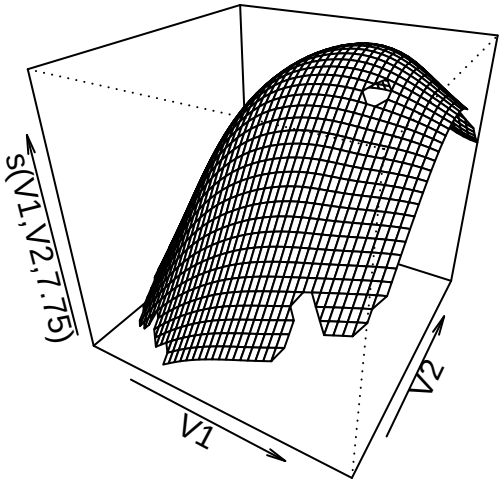
Alternativamente, si potevano valutare diversi modelli come segue (libreria caret)




```
$lm
[1] 2.890889

$glmnet
[1] 2.831696

$rf
[1] 2.903982

$boost
[1] 2.251615
```

Infine, si poteva individuare l'effetto di interazione della coppia X_1, X_2 con Y con il seguente grafico (libreria mgcv):

```
fit2 = mgcv::gam(y ~ s(V1,V2), data=train)
plot(fit2, scheme = TRUE)
```


s(V1,V2,7.75)
V1
V2