Relazione finale: Gli animali e gli alberi.

Chiara Camerota

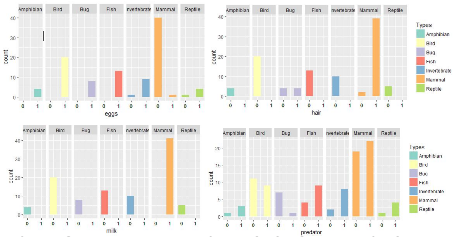
Università degli studi di Firenze

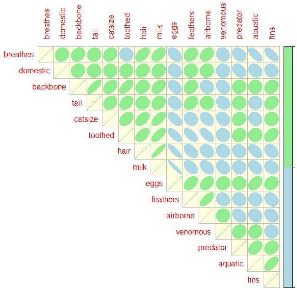
14 gennaio 2018

Gli animali

Descrizione data set e problema

- ▷ 101 animali da classificare
- ▷ 7 gruppi
- ▷ 16 predittori (buona parte Booleani)
- ▷ Problema di classificazione multinomiale
- ▷ Train set: 70 % dei dati originali

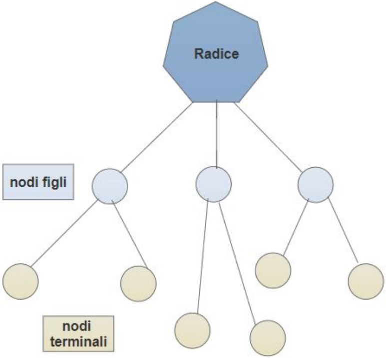


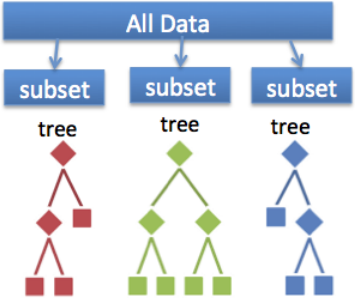


Gli alberi

Metodi applicati

- CART
- Random Forest
- Stochastic gradient boosting





CART

Indici di split usati per creare i due alberi e pacchetto

- **Classification error rate:**

$$E = 1| - max_k(\hat{p}_{mk})$$

- **Indice di Gini:**

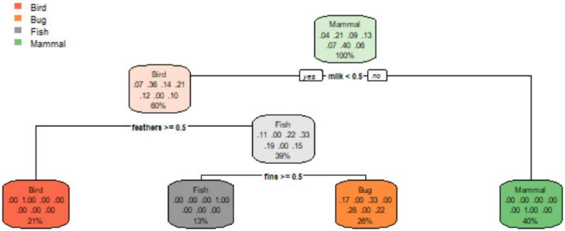
$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

  Dove $\hat{p}_{mk}$ indica la proporzione di osservazioni del train set nella regione m appartenente alla classe k.

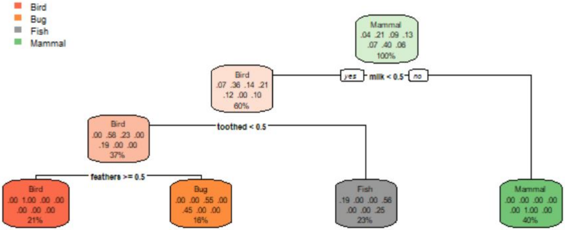
- **Pacchetto:** *rpart*

CART

CART usando l'indice di Gini

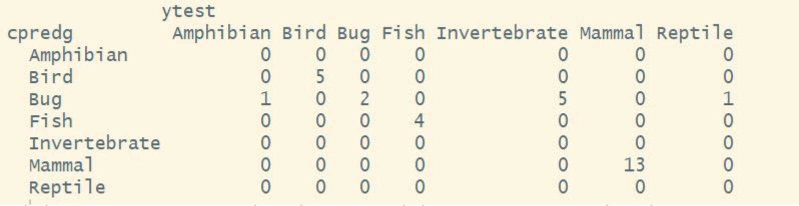


CART usando l'error rate



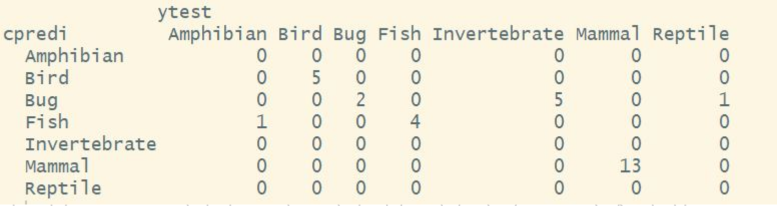
CART

Matrici di confondimento



```
              ytest
cpredg          Amphibian Bird Bug Fish Invertebrate Mammal Reptile
  Amphibian             0    0   0    0            0      0       0
  Bird                  0    5   0    0            0      0       0
  Bug                   1    0   2    0            5      0       1
  Fish                  0    0   0    4            0      0       0
  Invertebrate          0    0   0    0            0      0       0
  Mammal                0    0   0    0            0     13       0
  Reptile               0    0   0    0            0      0       0
```

Figura: Matrice di confondimento, usando l'indice di Gini.

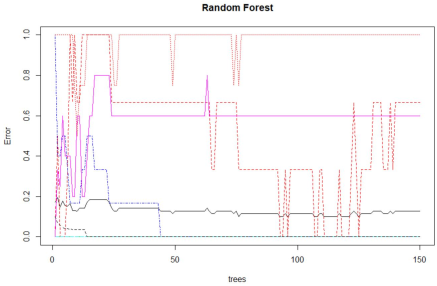


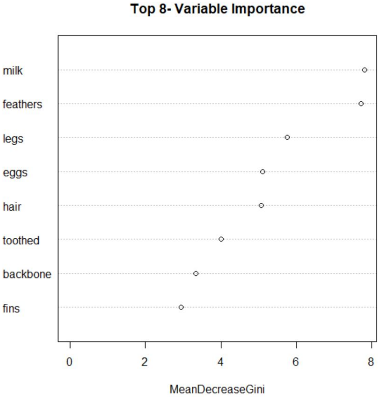
```
              ytest
cpredi          Amphibian Bird Bug Fish Invertebrate Mammal Reptile
  Amphibian             0    0   0    0            0      0       0
  Bird                  0    5   0    0            0      0       0
  Bug                   0    0   2    0            5      0       1
  Fish                  1    0   0    4            0      0       0
  Invertebrate          0    0   0    0            0      0       0
  Mammal                0    0   0    0            0     13       0
  Reptile               0    0   0    0            0      0       0
```

Figura: Matrice di confondimento, usando l'error rate.

Random Forest

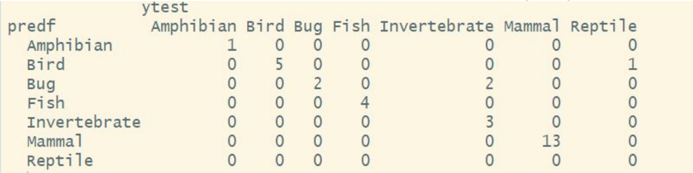
- ▷ Pacchetto: *RandomForest*
- ▷ Usa l'**indice di Gini** per split e pruning
- ▷ Due modelli nested





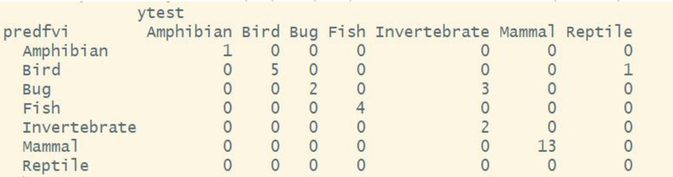
Random Forest

Matrici di confondimento



```
               ytest
predf            Amphibian Bird Bug Fish Invertebrate Mammal Reptile
  Amphibian              1    0   0    0            0      0       0
  Bird                   0    5   0    0            0      0       1
  Bug                    0    0   2    0            2      0       0
  Fish                   0    0   0    4            0      0       0
  Invertebrate           0    0   0    0            3      0       0
  Mammal                 0    0   0    0            0     13       0
  Reptile                0    0   0    0            0      0       0
```

Figura: Matrice di confondimento Random Forest

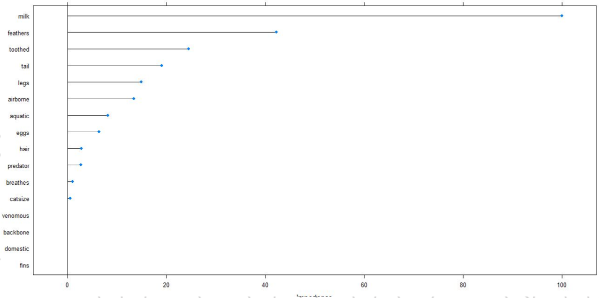


```
               ytest
predfvi          Amphibian Bird Bug Fish Invertebrate Mammal Reptile
  Amphibian              1    0   0    0            0      0       0
  Bird                   0    5   0    0            0      0       1
  Bug                    0    0   2    0            3      0       0
  Fish                   0    0   0    4            0      0       0
  Invertebrate           0    0   0    0            2      0       0
  Mammal                 0    0   0    0            0     13       0
  Reptile                0    0   0    0            0      0       0
```

Figura: Matrice di confondimento Random Forest con selezione di variabili

Stochastic Gradient Boosting

- Pacchetto: *caret*
- Funzione perdita : $\Psi_{i,k} = \sum_{i=1}^{N} \sum_{k=1}^{K} -log(p_{\pi(i),k}) 1_{\hat{y}_i=k}$ dove $\hat{y}_{i,k} = 1$ *se* $\hat{y}_i = k$, altrimenti $\hat{y}_{i,k} = 0$

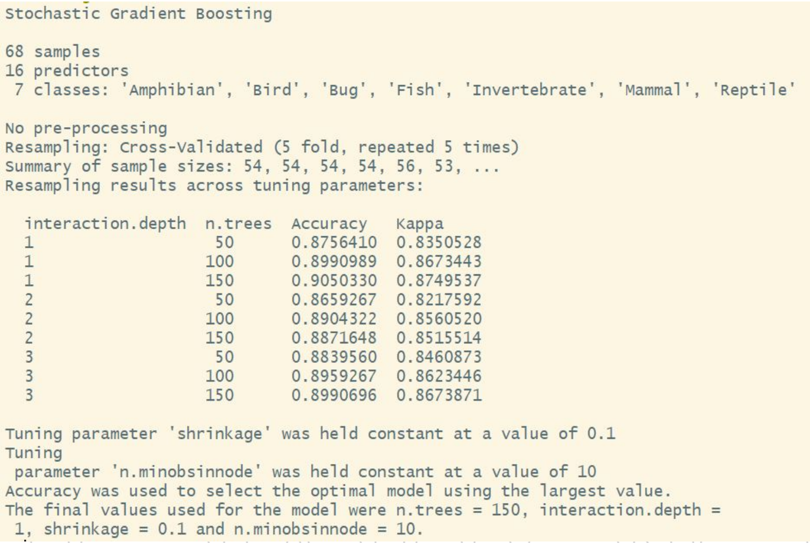


Stochastic Gradient Boosting

Algoritmo

```
  ▷ set: $\Psi_{i,k} = 0, k = 1toK, i = 1toN.$
▷▷ for m=1 to M do:
▷▷ for k=1 to K do:
▷▷ $\{\pi(i)_1^N = random(i)\}$
▷▷ $p_{\pi(i),k} = \frac{exp(\Psi_{\pi(i),k})}{\sum_{s=1}^{K} exp(\Psi_{\pi(i),s})}$
▷▷ $\{R_{j,k,m}\}_{j=1}^{J} = \{\hat{y}_{\pi(i),k} - p_{i,k}, \mathbf{x}_i\}_{i=1}^{N}$ per il nodo terminale $J$
▷▷ $\beta_{j,k,m} = \frac{K-1}{K} \frac{\sum_{\mathbf{x}_i \in R_{j,k,m}} \hat{y}_{i,k} - p_{i,k}}{(1 - p_{i,k})p_{i,k}}$
▷▷ $\Psi_{i,k} = \Psi_{i,k} + \lambda\beta_{j,k,m} 1_{\mathbf{x}_{i \in R_{j,k,m}}}$
 ▷ end both for
```

Stochastic Gradient Boosting

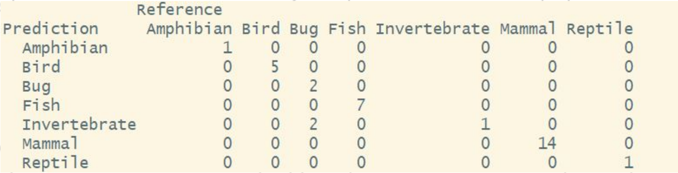


```
Stochastic Gradient Boosting

68 samples
16 predictors
 7 classes: 'Amphibian', 'Bird', 'Bug', 'Fish', 'Invertebrate', 'Mammal', 'Reptile'

No pre-processing
Resampling: Cross-Validated (5 fold, repeated 5 times)
Summary of sample sizes: 54, 54, 54, 54, 56, 53, ...
Resampling results across tuning parameters:

  interaction.depth  n.trees  Accuracy   Kappa
  1                   50      0.8756410  0.8350528
  1                  100      0.8990989  0.8673443
  1                  150      0.9050330  0.8749537
  2                   50      0.8659267  0.8217592
  2                  100      0.8904322  0.8560520
  2                  150      0.8871648  0.8515514
  3                   50      0.8839560  0.8460873
  3                  100      0.8959267  0.8623446
  3                  150      0.8990696  0.8673871

Tuning parameter 'shrinkage' was held constant at a value of 0.1
Tuning
 parameter 'n.minobsinnode' was held constant at a value of 10
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were n.trees = 150, interaction.depth =
 1, shrinkage = 0.1 and n.minobsinnode = 10.
```

Figura: Descrizione del modello

Stochastic Gradient Boosting

Matrice di confondimento

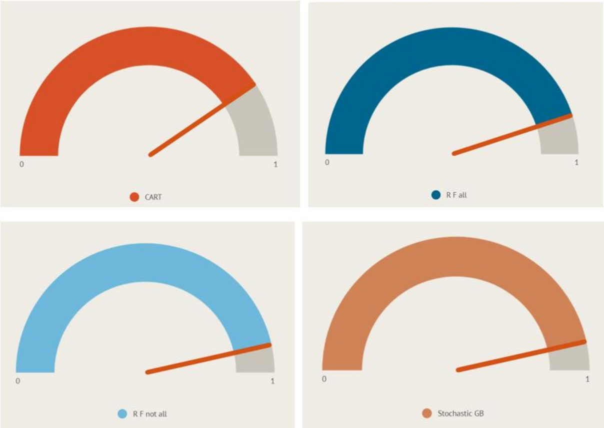


```
              Reference
Prediction     Amphibian Bird Bug Fish Invertebrate Mammal Reptile
  Amphibian            1    0   0    0            0      0       0
  Bird                 0    5   0    0            0      0       0
  Bug                  0    0   2    0            0      0       0
  Fish                 0    0   0    7            0      0       0
  Invertebrate         0    0   2    0            1      0       0
  Mammal               0    0   0    0            0     14       0
  Reptile              0    0   0    0            0      0       1
```

Figura: Matrice di confondimento di entrambi i modelli

Conclusione

Classificatore	Accuratezza
CART (entrambi)	0.81
Random Forest (senza selezione di variabili)	0.90
Random Forest (con selezione di variabili)	0.93
Stochastic GB (entrambi)	0.93



Bibliografia

- Johnson, R.A., & Wichern, D.W., *Applied multivariate statistical analysis*, Pearson 2014

- T. M. Mitchell, *Machine Learning*, McGraw-Hill, 1997

- T. Hastie, R. Tibshirani, & J. Friedman. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction. 2nd edition*, Springer, 2009

- Frank, Eibe, et al. *Using model trees for classification, Machine Learning 32.1* 1998: 63-76

- Ho, Tin Kam *Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC,* 14?16 August 1995. pp. 278?282

- Friedman, J. H., *Stochastic Gradient Boosting* , 1999

Bibliografia

Mason, L.; Baxter, J.; Bartlett, P. L.; Frean, Marcus *"Boosting Algorithms as Gradient Descent"; In S.A. Solla and T.K. Leen and K. Müller.*, Advances in Neural Information Processing Systems 12. MIT Press. pp. 512?518.

Hastie, T.; Tibshirani, R.; Friedman, J. H. *The Elements of Statistical Learning (2nd ed.)*, New York: Springer, 2009.