

Machine learning for regression problems

Filippo Biscarini (CNR, Milan, Italy)

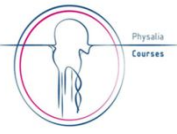
filippo.biscarini@cnr.it



Introducing the dataset



DNA methylation and age in European bats



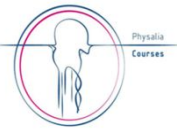
- Wild Bechstein's bats (*Myotis bechsteinii*)
- N = 62
- seven CpG sites ([here](#))
- % methylation
- age: [0-14] years



Source: https://en.wikipedia.org/wiki/Bechstein%27s_bat



DNA methylation and age in European bats



Received: 26 February 2018 | Revised: 21 April 2018 | Accepted: 2 May 2018

DOI: 10.1111/1755-0998.12925

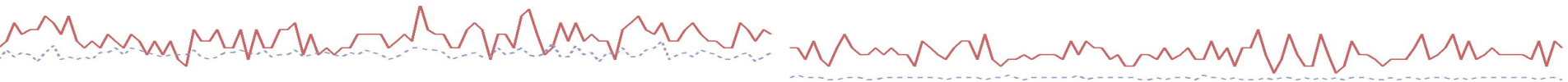
RESOURCE ARTICLE

WILEY

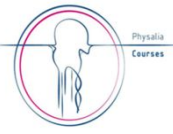
MOLECULAR ECOLOGY
RESOURCES

Application of a novel molecular method to age free-living wild Bechstein's bats

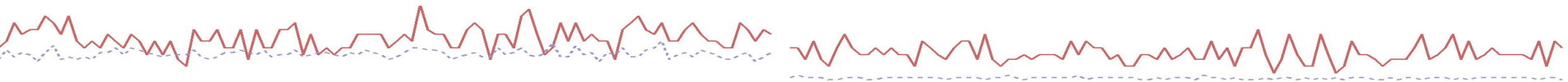
Patrick G. R. Wright¹  | Fiona Mathews²  | Henry Schofield³ | Colin Morris³ |
Joe Burrage⁴ | Adam Smith⁴ | Emma L. Dempster⁴ | Patrick B. Hamilton¹ 



DNA methylation and age in European bats



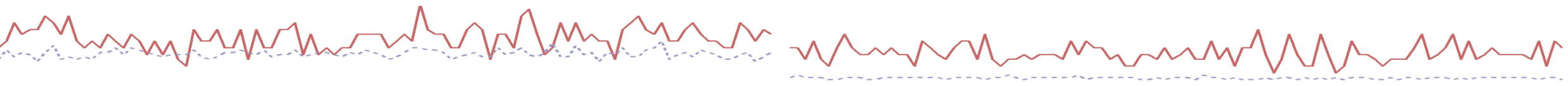
Sample	Age	Age category	CpG 1 TET2	CpG 2 TET2	CpG 3 TET2	CpG 4 TET2	CpG GRIA2 1	CpG GRIA2 2	ASPA 1
BabyBechs_SHW	0	Age 0-3	29	21	26	31	2	2	61
Dd_Juv_Hamgreen	0	Age 0-3	30	21	24	32	1	2	59
A2402	1	Age 0-3	44	38	51	53	1	5	48
A2414-2014	1	Age 0-3	48	36	50	46	1	2	53
...



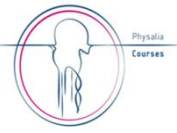
Regression problems

Demonstration 2.1

→ `linear_regression.Rmd`



Normalized discounted cumulative gain (NDCG)



NDCG is a **ranking metric** developed in information theory which has been applied to evaluation of genomic selection models

NDCG evaluates the **top** (e.g. 20%) **individuals in the ranking**, which are supposed to be the most relevant when comparing models

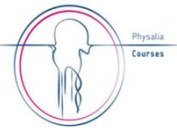
$$DCG@k = \sum_{i=1}^k y[\pi(\hat{y}_i)] \cdot d(i)$$

gain function

discount



Normalized discounted cumulative gain (NDCG)



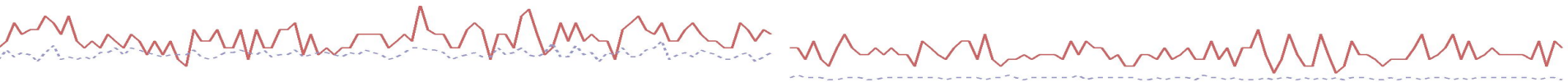
NDCG is a **ranking metric** developed in information theory which has been applied to evaluation of genomic selection models

NDCG evaluates the **top** (e.g. 20%) **individuals in the ranking**, which are supposed to be the most relevant when comparing models

$$DCG@k = \sum_{i=1}^k y[\pi(\hat{y}_i)] \cdot d(i)$$

The higher the DCG, the better

However, DCG is difficult to interpret (unbounded)  **NDCG**



Normalized discounted cumulative gain (NDCG)



NDCG is a **ranking metric** developed in information theory which has been applied to evaluation of genomic selection models

NDCG evaluates the **top** (e.g. 20%) **individuals in the ranking**, which are supposed to be the most relevant when comparing models

$$NDCG(y, \hat{y}) = \frac{\sum_{i=1}^k (y[\pi(\hat{y})]_i \cdot d(i))}{\sum_{i=1}^k (y[\pi(y)]_i \cdot d(i))}$$

NDCG values lie in [0,1]



Regression problems

Exercise 2.1

→ `linear_regression.Rmd`

