CSIC 5011 Project 1 – How do features of animals correlated?

Richard, Li Yan Chak[1]

[1]Department of Chemical and Biological Engineering, HKUST

Aims

Find out **Weights of well-known features (Columns without NA)**:
Body, Brain, Predation, Sleep Exposure and Danger
correlating other features (Columns with NA):
Slow Wave Sleep, Dream Sleep, Sleep, Life, Gestation

Methodology

Employ Different Linear Regression Model from sklearn library, find out **w**:

y = **Xw**

Ordinary Least Square (OLS):

$$\min_w ||Xw - y||_2{}^2$$

Lasso:

$$\min_w \frac{1}{2n_{samples}}||Xw - y||_2^2 + \alpha||w||_1$$

Orthogonal Matching Pursuit (Omp):

$$min_w\ ||w||_0\ subject\ to\ ||Xw - y||_2^2 \leq tol$$

BayesianRidge: Assuming **y** is Gaussian distributed around **Xw**

$$p(y|X, w, \alpha) = \mathcal{N}(y|Xw, \alpha)$$

ARDRegression: Similar to Bayesian Ridge, but with zero-mean Gaussian prior of **w**

Training and Evaluation Data

Training Data (total 42 animal species):
Animals without any unknown (no NA)

Evaluation Data:
Animals with some unknown but known with target feature (like African elephant will be used for evaluating sleep, life, gestation but not for Slow Wave Sleep, Dream Sleep)

Score of Evaluation:
R2 Score (Coefficient of determination) is used

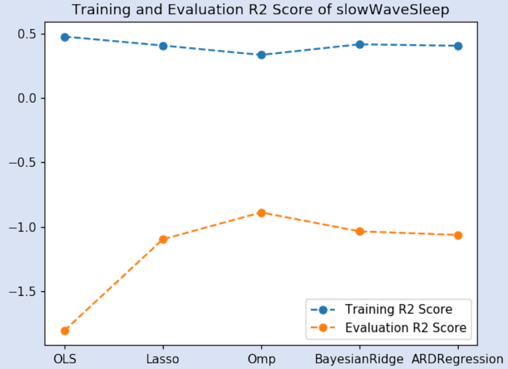
Data Preprocessing

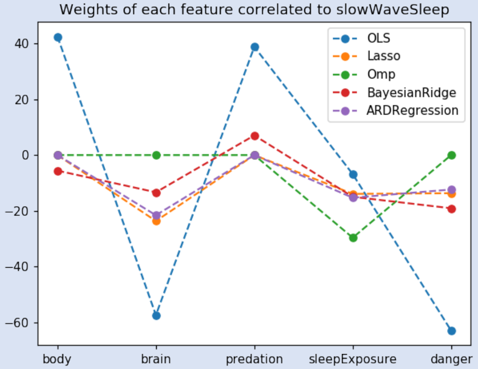
Ranges of well-known features **X** are very diverge like, Body = [0.01, 6654] and danger = [1, 5].

Lasso will tends to suppress the weight of higher value, Therefore, Sigmoid based normalisation is performed, which is similar to z-score which is bounded by [0, 1]:

$$w_{normalised} = \frac{1}{1 + e^{\frac{-(w - w_{median})}{w_{max}}}}$$

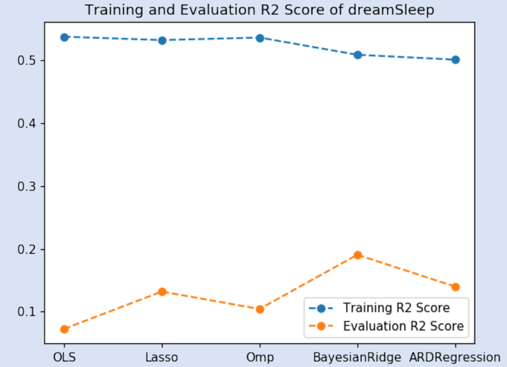
Evaluation of Regression on Slow Wave Sleep

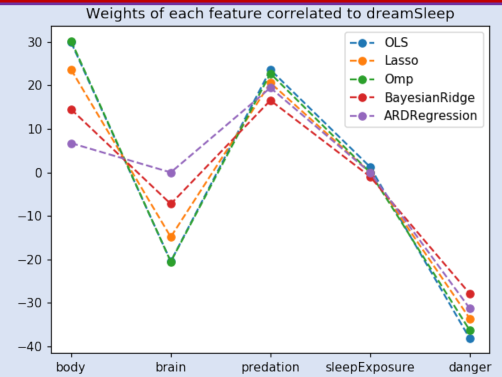

Training and Evaluation R2 Score of slowWaveSleep


Weights of each feature correlated to slowWaveSleep

Although R2 looks not bad in training data set, R2 of evaluation data is extremely low. It is failed to draw out a linear relationship of slow wave sleep and other well-known features.

Evaluation of Regression on Dream Sleep


Training and Evaluation R2 Score of dreamSleep


Weights of each feature correlated to dreamSleep

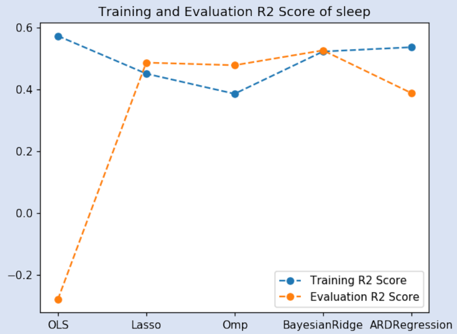
Same as slow wave sleep, good training R2 score but bad in evaluation score. It failed to draw out good linear relationship, but weight of feature among different regressors are similar, like heavier body need longer dream sleep, endangerous species have a shorter dream sleep.

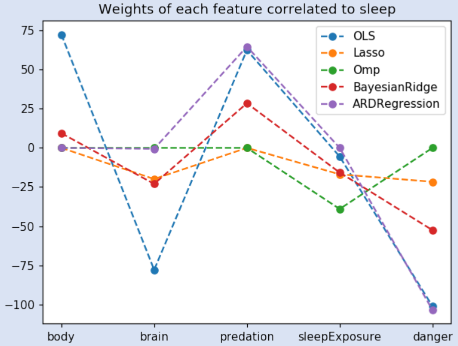
CSIC 5011 Project 1 – How do features of animals correlated?

Richard, Li Yan Chak[1]

[1]Department of Chemical and Biological Engineering, HKUST

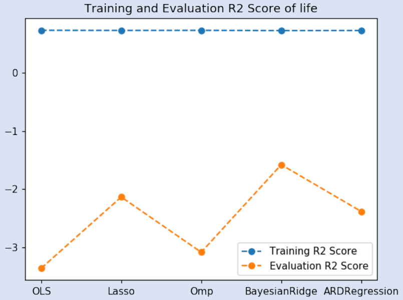
Evaluation of Regression on (total) Sleep

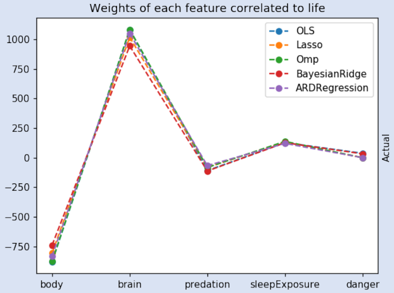


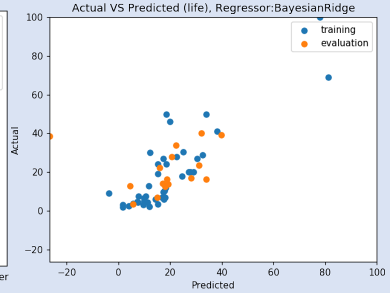


Only OLS performs poorly in evaluation, while others R2 score looks not bad. Weights of different regressors are not consist as dream sleep, but still reflect total sleep having positive correlation to predation and negative correlation to its endangerous level.

Evaluation of Regression on Life

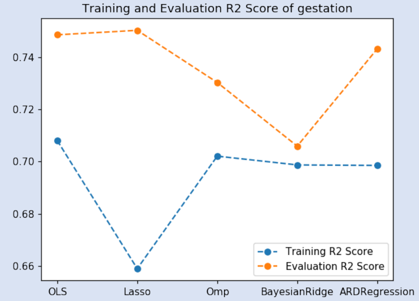


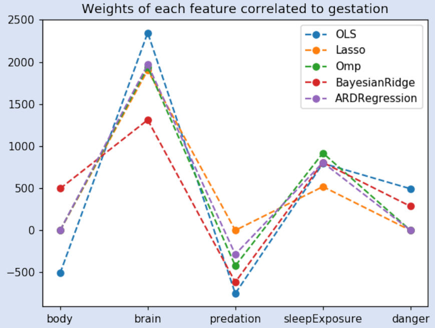




Although training score is not bad, there is outliers in evaluation, which make a predicted Life time as -25 while its actual value is 40. Other than this, Life time show a strong positive relation with brain size and negative relation with body weight.

Evaluation of Regression on Gestation





All regressors perform well in Gestation. It show gestation have a strong positive correlation brain size and its endangerous level.

Conclusion

Not all features have a strong linear relationship with well-known features, but gestation is sharply shown a good linear relationship with well-known features. Hope it can be a good research direction to ecologist.

References

sklearn.linear_model: http://scikit-learn.org/stable/modules/linear_model.html

Animal Sleep Data: http://math.stanford.edu/~yuany/course/data/sleep1.csv