

Introduction to Data Science

Data Science Essentials



Goals for today

- Review last session coding tasks
- Intro to Machine Learning
- A look ahead



Review last session coding tasks

week4_review notebook




Supervised Learning

You **supervise** your computer as it learns by giving it known outcomes for a sample of explanatory variables. This involves creating **labeled training data**.

Some common supervised learning algorithms are **linear regression**, **logistic regression**, **classification**, **support vector machines**, and **decision trees**.

	-----predictor variables-----					
	name	hair	cold-blooded	land/sea	flies	class
0	dog	Y	N	land	N	mammal
1	snake	N	Y	land	N	reptile
2	trout	N	Y	sea	N	fish
...						
499	dove	N	N	land	Y	bird

target variable



Unsupervised Learning

Your computer makes predictions ***based on anomalies, patterns, and relationships it finds in data.***

Some unsupervised machine learning algorithms are k-means clustering, principal component analysis (PCA).

	name	hair	cold-blooded	land/sea	flies	class
0	dog	Y	N	land	N	mammal
1	snake	N	Y	land	N	reptile
2	trout	N	Y	sea	N	fish
...						
499	dove	N	N	land	Y	bird



Choose Explanatory Variables

- What variables do you think might predict some target given your exploratory data analysis?
- Choose variables that ***represent the variance in the data*** (not highly correlated variables!)
- A pairplot will let you see if any of your predictors are highly correlated with one another

```
sns.set(style = 'ticks', color_codes = True)  
sns.pairplot(tn_ha_costs2);
```



Transform categorical predictor variables

recent_5year_trend

- falling
- rising
- stable

If we encode these as numerical values, say,

Falling = 1

Rising = 2

Stable = 3

Higher numbers will be weighted more!



pd.get_dummies()

```
In [79]: one_hot_encoded_X = pd.get_dummies(X)
```

```
In [100]: one_hot_encoded_X.head()
```

```
Out[100]:
```

id	percent_pop_over_50_2017	avg_annual_count	recent_5year_trend_falling	recent_5year_trend_rising	recent_5year_trend_stable
751	0.412762	195	1	0	0
325	0.339755	92	1	0	0
955	0.464281	59	0	0	1
271	0.410546	31	1	0	0
860	0.414311	283	1	0	0



Scikit-learn

Basic Recipe for Building a Model

1. Encode categorical data
2. Split into train and test
3. Construct model
4. Fit model with training data
5. Use fitted model to predict target for test data
6. Compare predictions to actual target values in the test data to evaluate performance
7. Iterate the model building process to improve performance

<https://scikit-learn.org/stable/>

<https://scikit-learn.org/stable/modules/classes.html#sklearn-metrics-metrics>



Today we will look at these modules from scikit-learn

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
```



Model building and evaluation

model_evaluation notebook



Next Steps:

Your turn! Build a logistic regression model to predict whether a county's cost-income ratio is above or below the mean (hint: first create a label for the data that answers that).

- **October 7 – random forest classifier example**
- **October 14 - data storytelling and presentation; work in teams to create a 7-10 minute presentation of your findings**
- **October 21 – Team presentations**



Questions?

