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**.

target variable

----------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

NASHVILLE SOFTWARE SCHOOL

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

NASHVILLE SOFTWARE SCHOOL

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);
```

NASHVILLE SOFTWARE SCHOOL

Scale variables if needed

- by hand – variables on same scale with same proportion within the column (look at range divide by a factor to get values between 0 and 1)
- with standard scaler from scikit-learn

http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

Variable 1	Variable 2	Variable 3	Variable 4
123	2.7	190	.092
614	3.9	84	.245

V1_scaled	V2_scaled	V3_scaled	Variable 4
.123	.27	.190	.092
.614	.39	.084	.245

NASHVILLE SOFTWARE SCHOOL

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!

NASHVILLE SOFTWARE SCHOOL

pd.get_dummies()

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

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

Out[100]:

014	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

Still may need to scale avg_annual_count!

NASHVILLE SOFTWARE SCHOOL

Scikit-learn

Basic Recipe for Building a Model

1. Split into train and test
2. Standardize/scale/normalize and encode data as needed
3. Construct model
4. Fit model with training data
5. Use fitted model to predict target for test data
    a. don't forget to apply the modifications in step 2 to predictors in the test set
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.linear_model import LinearRegression
from sklearn import metrics
```

NASHVILLE SOFTWARE SCHOOL

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**

NASHVILLE SOFTWARE SCHOOL

Questions?