

Lec 13 - scikit-learn

Statistical Computing and Computation

Sta 663 | Spring 2022

Dr. Colin Rundel

scikit-learn

Scikit-learn is an open source machine learning library that supports supervised and unsupervised learning. It also provides various tools for model fitting, data preprocessing, model selection, model evaluation, and many other utilities.

- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

This is one of several other "scikits" (e.g. scikit-image) which are scientific toolboxes built on top of scipy.

Submodules

The `sklearn` package contains a large number of submodules which are specialized for different tasks / models,

- `sklearn.base` - Base classes and utility functions
- `sklearn.calibration` - Probability Calibration
- `sklearn.cluster` - Clustering
- `sklearn.compose` - Composite Estimators
- `sklearn.covariance` - Covariance Estimators
- `sklearn.cross_decomposition` - Cross decomposition
- `sklearn.datasets` - Datasets
- `sklearn.decomposition` - Matrix Decomposition
- `sklearn.discriminant_analysis` - Discriminant Analysis
- `sklearn.ensemble` - Ensemble Methods
- `sklearn.exceptions` - Exceptions and warnings
- `sklearn.experimental` - Experimental
- `sklearn.feature_extraction` - Feature Extraction
- `sklearn.feature_selection` - Feature Selection
- `sklearn.gaussian_process` - Gaussian Processes
- `sklearn.kernel_ridge` - Kernel Ridge Regression
- `sklearn.linear_model` - Linear Models
- `sklearn.manifold` - Manifold Learning
- `sklearn.metrics` - Metrics
- `sklearn.mixture` - Gaussian Mixture Models
- `sklearn.model_selection` - Model Selection
- `sklearn.multiclass` - Multiclass classification
- `sklearn.multioutput` - Multioutput regression and classification
- `sklearn.naive_bayes` - Naive Bayes
- `sklearn.neighbors` - Nearest Neighbors
- `sklearn.neural_network` - Neural network models
- `sklearn.pipeline` - Pipeline
- `sklearn.preprocessing` - Preprocessing and Normalization
- `sklearn.random_projection` - Random projection

Model Fitting

Sample data

To begin, we will examine a simple data set on the size and weight of a number of books. The goal is to model the weight of a book using some combination of the other features in the data.

The included columns are:

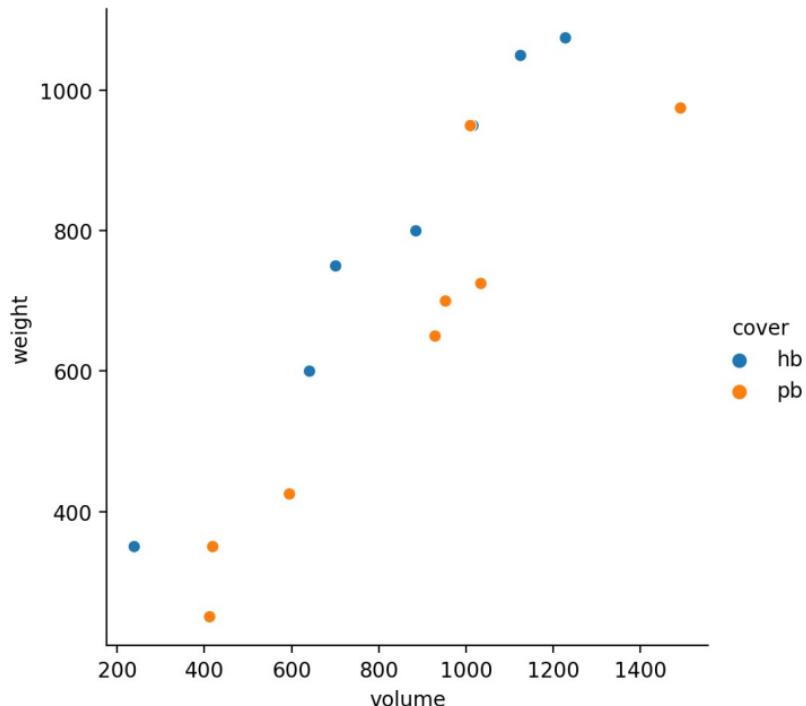
- volume - book volumes in cubic centimeters
- weight - book weights in grams
- cover - a categorical variable with levels "

```
books = pd.read_csv("data/daag_books.csv")  
books
```

```
##      volume    weight  cover  
## 0       885     800    hb  
## 1      1016     950    hb  
## 2      1125    1050    hb  
## 3       239     350    hb  
## 4       701     750    hb  
## 5       641     600    hb  
## 6      1228    1075    hb  
## 7       412     250    pb  
## 8      953      700    pb  
## 9      929      650    pb  
## 10     1492     975    pb  
## 11     419      350    pb  
## 12     1010     950    pb  
## 13      595     425    pb  
## 14     1034     725    pb
```

These data come from the `abucks` data set from the DAAG package in R

```
sns.relplot(data=books, x="volume", y="weight", hue="cover")
```



Linear regression

scikit-learn uses an object oriented system for implementing the various modeling approaches, the class for LinearRegression is part of the linear_model submodule.

```
from sklearn.linear_model import LinearRegression
```

Each modeling class needs to be constructed (potentially with options) and then the resulting object will provide attributes and methods.

```
lm = LinearRegression()  
  
m = lm.fit(  
    X = books[["volume"]],  
    y = books.weight  
)  
  
m.coef_  
  
## array([0.70863714])
```

```
m.intercept_
```

Note lm and m are labels for the same object,

```
lm.coef_
```

```
## array([0.70863714])
```

```
lm.intercept_
```

```
## 107.679310613766
```

A couple of considerations

When fitting a model, scikit-learn expects `x` to be a 2d array-like object (e.g. a `np.array` or `pd.DataFrame`) but will not accept a `pd.Series` or **1d np.array**.

```
lm.fit(  
    X = books.volume,  
    y = books.weight  
)
```

```
## ValueError: Expected 2D array, got 1D array instead:  
## array=[ 885 1016 1125  239  701  641 1228  412  
##  1034].  
## Reshape your data either using array.reshape(-1
```

```
lm.fit(  
    X = np.array(books.volume),  
    y = books.weight  
)
```

```
## ValueError: Expected 2D array, got 1D array instead:  
## array=[ 885 1016 1125  239  701  641 1228  412  953  
##  1034].  
## Reshape your data either using array.reshape(-1, 1)
```

```
lm.fit(  
    X = np.array(books.volume).reshape(-1,1),  
    y = books.weight  
)
```

```
## LinearRegression()
```

Model parameters

Depending on the model being used, there will be a number of parameters that can be configured when creating the model object or via the `set_params()` method.

```
lm.get_params()  
  
## {'copy_X': True, 'fit_intercept': True, 'n_jobs': None, 'normalize': 'deprecated', 'positive': False}  
  
lm.set_params(fit_intercept = False)  
  
## LinearRegression(fit_intercept=False)  
  
lm = lm.fit(X = books[["volume"]], y = books.weight)  
lm.intercept_  
  
## 0.0  
  
lm.coef_  
  
## array([0.81932487])
```

Model prediction

Once the model coefficients have been fit, it is possible to predict using the model via the `predict()` method, this method requires a matrix-like `x` as input and in the case of `LinearRegression` returns an array of predicted `y` values.

```
lm.predict(X = books[["volume"]])  
  
## [1] 725.10251417 832.43407276 921.74048411 195.81864507  
## [5] 574.34673721 525.18724472 1006.13094621 337.5618484 ,  
## [9] 780.81660565 761.15280865 1222.43271315 343.29712253,  
## [13] 827.51812351 487.49830048 847.1819205 ])
```

```
books[ "weight_lm_pred" ] = lm.predict(X = books[["volume"]])  
books
```

```
##   volume weight cover weight_lm_pred  
## 0     885    800   hb      725.102514  
## 1    1016    950   hb      832.434073  
## 2    1125   1050   hb      921.740484  
## 3     239    350   hb      195.818645  
## 4     701    750   hb      574.346737  
## 5     641    600   hb      525.187245  
## 6    1228   1075   hb     1006.130946
```

```
plt.figure()  
sns.scatterplot(data=books, x="volume", y="weight", hue="cover")  
sns.lineplot(data=books, x="volume", y="weight_lm_pred", color="green")  
plt.show()
```

Residuals?

There is no built in functionality for calculating residuals, so this needs to be done by hand.

```
books["resid_lm_pred"] = books["weight"] - books["weight_lm_pred"]

plt.figure(layout="constrained")
ax = sns.scatterplot(data=books, x="volume", y="resid_lm_pred", hue="cover")
ax.axhline(c="k", ls="--", lw=1)
plt.show()
```

Categorical variables?

Scikit-learn expects that the model matrix be numeric before fitting,

```
lm = lm.fit(  
    X = books[["volume", "cover"]],  
    y = books.weight  
)
```

```
## ValueError: could not convert string to float: 'hb'
```

the obvious solution here is dummy coding the categorical variables - this can be done with pandas via `pd.get_dummies()` or with a scikit-learn preprocessor, we'll demo the former first.

```
pd.get_dummies(books[["volume", "cover"]])
```

```
##      volume  cover_hb  cover_pb  
## 0      885        1        0  
## 1     1016        1        0  
## 2     1125        1        0  
## 3      239        1        0  
## 4      701        1        0  
## 5      641        1        0  
## 6    1228        1        0  
## 7      412        0        1
```

```
lm = LinearRegression().fit(  
    X = pd.get_dummies(books[["volume", "cover"]]),  
    y = books.weight  
)
```

```
lm.intercept_
```

```
## 105.93920788192202
```

```
lm.coef_
```

```
## array([ 0.71795374,  92.02363569, -92.02363569])
```

Do these results look reasonable? What went wrong?

Quick comparison with R

```
d = read.csv('data/daag_books.csv')
d['cover_hb'] = ifelse(d$cover == "hb", 1, 0)
d['cover_pb'] = ifelse(d$cover == "pb", 1, 0)
(lm = lm(weight~volume+cover_hb+cover_pb, data=d))

##
## Call:
## lm(formula = weight ~ volume + cover_hb + cover_pb, data = d)
##
## Coefficients:
## (Intercept)      volume    cover_hb    cover_pb
##       13.916       0.718     184.047        NA

summary(lm)
```

```
##
## Call:
## lm(formula = weight ~ volume + cover_hb + cover_pb, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -110.10  -32.32  -16.10   28.93  210.95
##
## Coefficients: (1 not defined because of singularities)
```

Avoiding co-linearity

```
lm = LinearRegression(fit_intercept = False).fit(  
    X = pd.get_dummies(books[["volume", "cover"]]),  
    y = books.weight  
)  
  
lm.intercept_  
## 0.0  
  
lm.coef_  
## array([ 0.71795374, 197.96284357, 13.91557219  
  
lm.feature_names_in_  
## array(['volume', 'cover_hb', 'cover_pb'], dtype=object)
```

```
lm = LinearRegression().fit(  
    X = pd.get_dummies(books[["volume", "cover"]]),  
    y = books.weight  
)  
  
lm.intercept_  
## 197.96284357271753  
  
lm.coef_  
## array([ 0.71795374, -184.04727138])  
  
lm.feature_names_in_  
## array(['volume', 'cover_pb'], dtype=object)
```

Preprocessors

These are a set of transformer classes present in the `sklearn.preprocessing` submodule that are designed to help with the preparation of raw feature data into quantities more suitable for downstream modeling tools.

Like the modeling classes, they have an object oriented design that shares a common interface (methods and attributes) for bringing in data, transforming it, and returning it.

OneHotEncoder

For dummy coding we can use the OneHotEncoder preprocessor, the default is to use one hot encoding but standard dummy coding can be achieved via the drop parameter.

```
from sklearn.preprocessing import OneHotEncoder
```

```
enc = OneHotEncoder(sparse=False)
enc.fit(X = books[["cover"]])
```

```
## OneHotEncoder(sparse=False)
enc.transform(X = books[["cover"]])
```

```
## array([[1., 0.],
##         [1., 0.],
##         [1., 0.],
##         [1., 0.],
##         [1., 0.],
##         [1., 0.],
##         [1., 0.],
##         [1., 0.],
##         [0., 1.],
##         [0., 1.],
##         [0., 1.]])
```

```
enc = OneHotEncoder(sparse=False, drop="first")
enc.fit_transform(X = books[["cover"]])
```

```
## array([[0.],
##         [0.],
##         [0.],
##         [0.],
##         [0.],
##         [0.],
##         [0.],
##         [0.],
##         [1.],
##         [1.],
##         [1.],
##         [1.],
##         [1.],
##         [1.],
##         [1.],
##         [1.],
##         [1.],
##         [1.],
##         [1.]])
```

Other useful bits

```
enc.get_feature_names_out()

## array(['cover_hb', 'cover_pb'], dtype=object)

f = enc.transform(X = books[["cover"]])
enc.inverse_transform(f)

## array([[['hb'],
##          ['hb'],
##          ['pb'],
##          ['pb']],
##         dtype=object)]
```

A cautionary note

Unlike `pd.get_dummies()` it is not safe to use `OneHotEncoder` with both numerical and categorical features, as the former will also be transformed.

```
enc = OneHotEncoder(sparse=False)
X = enc.fit_transform(
    X = books[["volume", "cover"]]
)

pd.DataFrame(
    data=X,
    columns = enc.get_feature_names_out()
)
```

```
##      volume_239  volume_412  volume_419  ...  volume_1492  cover_hb  cover_pb
## 0          0.0      0.0      0.0  ...        0.0      1.0      0.0
## 1          0.0      0.0      0.0  ...        0.0      1.0      0.0
## 2          0.0      0.0      0.0  ...        0.0      1.0      0.0
## 3          1.0      0.0      0.0  ...        0.0      1.0      0.0
## 4          0.0      0.0      0.0  ...        0.0      1.0      0.0
## 5          0.0      0.0      0.0  ...        0.0      1.0      0.0
## 6          0.0      0.0      0.0  ...        0.0      1.0      0.0
## 7          0.0      1.0      0.0  ...        0.0      0.0      1.0
## 8          0.0      0.0      0.0  ...        0.0      0.0      1.0
## 9          0.0      0.0      0.0  ...        0.0      0.0      1.0
## 10         0.0      0.0      0.0  ...        1.0      0.0      1.0
## 11         0.0      0.0      1.0  ...        0.0      0.0      1.0
## 12         0.0      0.0      0.0  ...        0.0      0.0      1.0
## 13         0.0      0.0      0.0  ...        0.0      0.0      1.0
## 14         0.0      0.0      0.0  ...        0.0      0.0      1.0
```

Putting it together

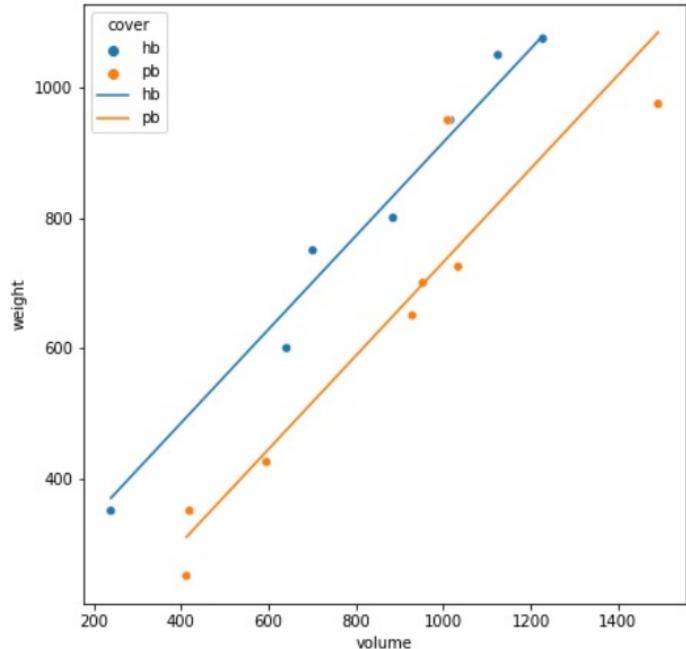
```
cover = OneHotEncoder(  
    sparse=False  
).fit_transform(  
    books[["cover"]]  
)  
X = np.c_[books.volume, cover]  
  
lm2 = LinearRegression(fit_intercept=False).fit(  
    X = X,  
    y = books.weight  
)  
  
lm2.coef_  
  
## array([ 0.71795374, 197.96284357, 13.91557219
```

```
books["weight_lm2_pred"] = lm2.predict(X=X)  
books.drop(["weight_lm_pred", "resid_lm_pred"], axis=1)
```

	volume	weight	cover	weight_lm2_pred
## 0	885	800	hb	833.351907
## 1	1016	950	hb	927.403847
## 2	1125	1050	hb	1005.660805
## 3	239	350	hb	369.553788
## 4	701	750	hb	701.248418
## 5	641	600	hb	658.171193
## 6	1228	1075	hb	1079.610041
## 7	412	250	pb	309.712515
## 8	953	700	pb	698.125490
## 9	929	650	pb	680.894600
## 10	1492	975	pb	1085.102558
## 11	419	350	pb	314.738191
## 12	1010	950	pb	739.048853
## 13	595	425	pb	441.098050
## 14	1034	725	pb	756.279743

We'll see a more elegant way of doing this in the near future

```
plt.figure()  
sns.scatterplot(data=books, x="volume", y="weight", hue="cover"  
sns.lineplot(data=books, x="volume", y="weight_lm2_pred", hue=""  
plt.show()
```



```
books["resid_lm2_pred"] = books["weight"] - books["weight_lm2_p  
plt.figure(layout="constrained")  
ax = sns.scatterplot(data=books, x="volume", y="resid_lm2_pred"  
ax.axhline(c="k", ls="--", lw=1)  
plt.show()
```

Model performance

Scikit-learn comes with a number of builtin functions for measuring model performance in the `sklearn.metrics` submodule - these are generally just functions that take the vectors `y_true` and `y_pred` and return the score as a scalar.

```
from sklearn.metrics import mean_squared_error, r2_score
```

```
r2_score(books.weight, books.weight_lm_pred)
```

```
## 0.7800969547785038
```

```
mean_squared_error(books.weight, books.weight_lm_
```

```
## 14833.68208377448
```

```
mean_squared_error(books.weight, books.weight_lm_
```

```
## 121.79360444528473
```

```
r2_score(books.weight, books.weight_lm2_pred)
```

```
## 0.9274775756821679
```

```
mean_squared_error(books.weight, books.weight_lm2
```

```
## 4892.040422595093
```

```
mean_squared_error(books.weight, books.weight_lm2
```

```
## 69.94312276839727
```

See API Docs for a list of available metrics

Exercise 1

Create and fit a model for the books data that includes an interaction effect between volume and cover.

You will need to do this manually with `pd.getdummies()` or `OneHotEncoder()`.

Polynomial regression

We will now look at another flavor of regression model, that involves preprocessing and a hyperparameter - namely polynomial regression.

```
df = pd.read_csv("data/gp.csv")
sns.relplot(data=df, x="x", y="y")
```

By hand

It is certainly possible to construct the necessary model matrix by hand (or even use a function to automate the process), but this is less then desirable generally - particularly if we want to do anything fancy (e.g. cross validation)

```
X = np.c_[  
    np.ones(df.shape[0]),  
    df.x,  
    df.x**2,  
    df.x**3  
]  
  
plm = LinearRegression(fit_intercept = False).fit(X=X, y=df.y)  
  
plm.coef_  
  
## array([ 2.36985684, -8.49429068, 13.95066369, -8.39215284])
```

```
df["y_pred"] = plm.predict(X=X)  
plt.figure(layout="constrained")  
sns.scatterplot(data=df, x="x", y="y")  
sns.lineplot(data=df, x="x", y="y_pred", color="k")  
plt.show()
```

PolynomialFeatures

This is another transformer class from `sklearn.preprocessing` that simplifies the process of constructing polynomial features for your model matrix. Usage is similar to that of `OneHotEncoder`.

```
from sklearn.preprocessing import PolynomialFeatures  
X = np.array(range(6)).reshape(-1,1)
```

```
pf = PolynomialFeatures(degree=3)  
pf.fit(X)
```

```
## PolynomialFeatures(degree=3)
```

```
pf.transform(X)
```

```
## array([[ 1.,  0.,  0.,  0.],  
##        [ 1.,  1.,  1.,  1.],  
##        [ 1.,  2.,  4.,  8.],  
##        [ 1.,  3.,  9., 27.],  
##        [ 1.,  4., 16., 64.],  
##        [ 1.,  5., 25., 125.]])
```

```
pf = PolynomialFeatures(degree=2, include_bias=False)  
pf.fit_transform(X)
```

```
## array([[ 0.,  0.],  
##        [ 1.,  1.],  
##        [ 2.,  4.],  
##        [ 3.,  9.],  
##        [ 4., 16.],  
##        [ 5., 25.]])
```

```
pf.get_feature_names_out()
```

```
## array(['x0', 'x0^2'], dtype=object)
```

Interactions

If the feature matrix x has more than one column then `PolynomialFeatures` transformer will include interaction terms with total degree up to `degree`.

```
X.reshape(-1, 2)
```

```
## array([[0, 1],  
##         [2, 3],  
##         [4, 5]])
```

```
pf = PolynomialFeatures(degree=3, include_bias=False)  
pf.fit_transform(X.reshape(-1, 2))
```

```
## array([[ 0.,   1.,   0.,   0.,   1.,   0.,   0.  
##         [ 2.,   3.,   4.,   6.,   9.,   8.,  12  
##         [ 4.,   5.,  16.,  20.,  25.,  64.,  80]
```

```
pf.get_feature_names_out()
```

```
## array(['x0', 'x1', 'x0^2', 'x0 x1', 'x1^2', 'x0  
##        'x1^3'], dtype=object)
```

```
X.reshape(-1, 3)
```

```
## array([[0, 1, 2],  
##         [3, 4, 5]])
```

```
pf = PolynomialFeatures(degree=2, include_bias=False)  
pf.fit_transform(X.reshape(-1, 3))
```

```
## array([[ 0.,   1.,   2.,   0.,   0.,   0.,   1.,   2.,   4.]  
##         [ 3.,   4.,   5.,   9.,  12.,  15.,  16.,  20.,  25.]])
```

```
pf.get_feature_names_out()
```

```
## array(['x0', 'x1', 'x2', 'x0^2', 'x0 x1', 'x0 x2', 'x1^2'],  
##        dtype=object)
```

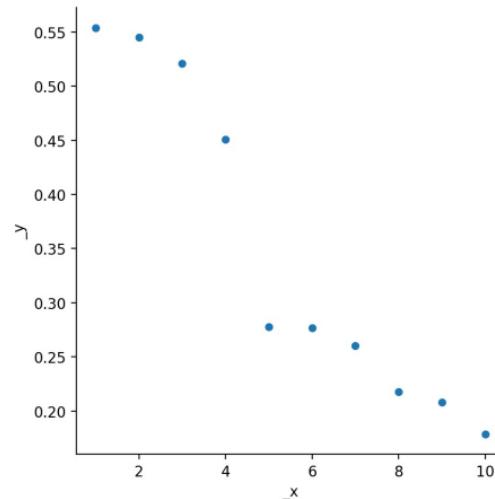
Modeling with PolynomialFeatures

```
def poly_model(X, y, degree):
    X = PolynomialFeatures(
        degree=degree, include_bias=False
    ).fit_transform(
        X=X
    )
    y_pred = LinearRegression().fit(X=X, y=y).predict(X)
    return mean_squared_error(y, y_pred, squared=False)

poly_model(X = df[["x"]], y = df.y, degree = 2)
## 0.5449418707295371

poly_model(X = df[["x"]], y = df.y, degree = 3)
## 0.5208157900621085
```

```
degrees = range(1,11)
rmses = [poly_model(X=df[["x"]], y=df.y, degree=d)
         for d in degrees]
sns.relplot(x=degrees, y=rmses)
```



Pipelines

You may have noticed that `PolynomialFeatures` takes a model matrix as input and returns a new model matrix as output which is then used as the input for `LinearRegression`. This is not an accident, and by structuring the library in this way `sklearn` is designed to enable the connection of these steps together, into what `sklearn` calls a pipeline.

```
from sklearn.pipeline import make_pipeline

p = make_pipeline(
    PolynomialFeatures(degree=4),
    LinearRegression()
)
p

## Pipeline(steps=[('polynomialfeatures', PolynomialFeatures(degree=4)),
##                  ('linearregression', LinearRegression())])
```

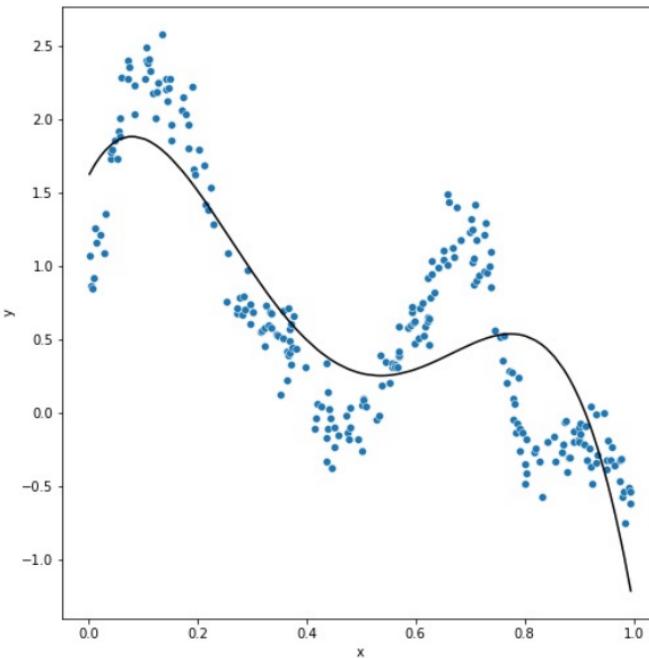
Using Pipelines

Once constructed, this object can be used just like our previous LinearRegression model (i.e. fit to our data and then used for prediction)

```
p = p.fit(X = df[["x"]], y = df.y)
p
## Pipeline(steps=[('polynomialfeatures', PolynomialFeatures(degree=4)),
##                  ('linearregression', LinearRegression())])
```

```
p.predict(X = df[["x"]])
## array([ 1.6295693 ,  1.65734929,  1.6610466 ,  1.67779767,  1.69667491,
##        1.70475286,  1.75280126,  1.78471392,  1.79049912,  1.82690007,
##        1.82966357,  1.83376043,  1.84494343,  1.86002819,  1.86228095,
##        1.86619112,  1.86837909,  1.87065283,  1.88417882,  1.8844024 ,
##        1.88527174,  1.88577463,  1.88544367,  1.86890805,  1.86365035,
##        1.86252922,  1.86047349,  1.85377801,  1.84937708,  1.83754576,
##        1.82623453,  1.82024199,  1.81799793,  1.79767794,  1.77255319,
##        1.77034143,  1.76574288,  1.75371272,  1.74389585,  1.73804309,
##        1.73356954,  1.65527727,  1.64812184,  1.61867613,  1.6041325 ,
##        1.5960389 ,  1.56080881,  1.55036459,  1.54004364,  1.50903953,
##        1.45096594,  1.43589836,  1.41886389,  1.39423307,  1.36180712,
##        1.23072992,  1.21355164,  1.11776117,  1.11522002,  1.09595388,
```

```
plt.figure(layout="constrained")
sns.scatterplot(data=df, x="x", y="y")
sns.lineplot(x=df.x, y=p.predict(X = df[["x"]]), color="k")
plt.show()
```



Model coefficients (or other attributes)

The attributes of steps are not directly accessible, but can be accessed via steps or named_steps attributes,

```
p.coef_
## AttributeError: 'Pipeline' object has no attribute 'coef_'

p.named_steps["linearregression"].intercept_
## 1.6136636604768615

p.steps[1][1].coef_
## array([ 0.          ,  7.39051417, -57.67175293, 102.72227443,
##        -55.38181361])

p.steps
## [('polynomialfeatures', PolynomialFeatures(degree=4)), ('linearregression', LinearRegression())]

p.steps[0][1].get_feature_names_out()
## array(['1', 'x', 'x^2', 'x^3', 'x^4'], dtype=object)

Anyone notice a problem?
```

What about step parameters?

By accessing each step we can adjust their parameters if necessary,

```
p.named_steps["linearregression"].get_params()  
  
## {'copy_X': True, 'fit_intercept': True, 'n_jobs': None, 'normalize': 'deprecated', 'positive': False}  
  
p.named_steps["linearregression"].set_params(fit_intercept=False)  
  
## LinearRegression(fit_intercept=False)  
  
p.fit(X = df[["x"]], y = df.y)  
  
## Pipeline(steps=[('polynomialfeatures', PolynomialFeatures(degree=4)),  
##                  ('linearregression', LinearRegression(fit_intercept=False))])  
  
p.named_steps["linearregression"].intercept_  
  
## 0.0  
  
p.named_steps["linearregression"].coef_  
  
## array([ 1.61366366,  7.39051417, -57.67175293, 102.72227443,  
##        -55.38181361])
```

These parameters can also be directly accessed at the pipeline level, note how the names are constructed:

```
p.get_params()  
  
## {'memory': None, 'steps': [('polynomialfeatures', PolynomialFeatures(degree=4)), ('linearregression', LinearRegression())]}  
  
p.set_params(linearregression__fit_intercept=True, polynomialfeatures__include_bias=False)  
  
## Pipeline(steps=[('polynomialfeatures',  
##                  PolynomialFeatures(degree=4, include_bias=False)),  
##                  ('linearregression', LinearRegression())])  
  
p.fit(X = df[["x"]], y = df.y)  
  
## Pipeline(steps=[('polynomialfeatures',  
##                  PolynomialFeatures(degree=4, include_bias=False)),  
##                  ('linearregression', LinearRegression())])  
  
p.named_steps["linearregression"].intercept_  
  
## 1.6136636604768375  
  
p.named_steps["linearregression"].coef_  
  
## array([-7.39051417, -57.67175293, 102.72227443, -55.38181361])
```

Tuning parameters

We've already seen a manual approach to tuning models over the degree parameter, scikit-learn also has built in tools to aide with this process. Here we will leverage GridSearchCV to tune the degree parameter in our pipeline.

```
from sklearn.model_selection import GridSearchCV, KFold  
  
p = make_pipeline(  
    PolynomialFeatures(include_bias=True),  
    LinearRegression(fit_intercept=False)  
)  
  
grid_search = GridSearchCV(  
    estimator = p,  
    param_grid = {"polynomialfeatures__degree": range(1,11)},  
    scoring = "neg_root_mean_squared_error",  
    cv = KFold(shuffle=True)  
)  
  
grid_search  
  
## GridSearchCV(cv=KFold(n_splits=5, random_state=None, shuffle=True),  
##                 estimator=Pipeline(steps=[('polynomialfeatures'  
##                                         PolynomialFeatures())],  
##                                         )  
## Much more detail on this next time - including the proper way to do cross-validation
```

Preview - Performing a grid search

```
grid_search.fit(X = df[["x"]], y = df.y)

## GridSearchCV(cv=KFold(n_splits=5, random_state=None, shuffle=True),
##               estimator=Pipeline(steps=[('polynomialfeatures',
##                                         PolynomialFeatures()),
##                                         ('linearregression',
##                                         LinearRegression(fit_intercept=False))]),
##               param_grid={'polynomialfeatures__degree': range(1, 11)},
##               scoring='neg_root_mean_squared_error')

grid_search.best_index_

## 9

grid_search.best_params_

## {'polynomialfeatures__degree': 10}

grid_search.best_score_

## -0.18606542718881763
```

cv_results_

```
grid_search.cv_results_["mean_test_score"]
```

```
## array([-0.56055647, -0.55414851, -0.54662235, -0.4660496 , -0.28446434,
##        -0.28374668, -0.26616867, -0.23083559, -0.22367841, -0.18606543])
```

```
grid_search.cv_results_["rank_test_score"]
```

```
## array([10,  9,  8,  7,  6,  5,  4,  3,  2,  1], dtype=int32)
```

```
grid_search.cv_results_["mean_fit_time"]
```

```
## array([0.0012239 , 0.00113297, 0.00105224, 0.00109162, 0.00105658,
##        0.00168719, 0.00103779, 0.00115523, 0.00105133, 0.00106945])
```

```
grid_search.cv_results_.keys()
```

```
## dict_keys(['mean_fit_time', 'std_fit_time', 'mean_score_time', 'std_score_time', 'param_polyomialfeatures__'])
```