Assignment 2

In this assignment you'll explore the relationship between model complexity and generalization performance, by adjusting key parameters of various supervised learning models. Part 1 of this assignment will look at regression and Part 2 will look at classification.

Part 1 - Regression

First, run the following block to set up the variables needed for later sections.

In [70]:

```python
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split


np.random.seed(0)
n = 15
x = np.linspace(0,10,n) + np.random.randn(n)/5
y = np.sin(x)+x/6 + np.random.randn(n)/10


X_train, X_test, y_train, y_test = train_test_split(x, y, random_state=0)

# You can use this function to help you visualize the dataset by
# plotting a scatterplot of the data points
# in the training and test sets.
def part1_scatter():
    import matplotlib.pyplot as plt
    %matplotlib notebook
    plt.figure()
    plt.scatter(X_train, y_train, label='training data')
    plt.scatter(X_test, y_test, label='test data')
    plt.legend(loc=4);


# NOTE: Uncomment the function below to visualize the data, but be sure
# to **re-comment it before submitting this assignment to the autograder**.
#part1_scatter()
```

Question 1

Write a function that fits a polynomial LinearRegression model on the *training data* X_train for degrees 1, 3, 6, and 9. (Use PolynomialFeatures in sklearn.preprocessing to create the polynomial features and then fit a linear regression model) For each model, find 100 predicted values over the interval x = 0 to 10 (e.g. np.linspace(0,10,100)) and store this in a numpy array. The first row of this array should correspond to the output from the model trained on degree 1, the second row degree 3, the third row degree 6, and the fourth row degree 9.

The figure above shows the fitted models plotted on top of the original data (using plot_one()).

This function should return a numpy array with shape (4, 100)

In [71]:

```python
def answer_one():
```

```python
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import PolynomialFeatures
degrees = [1,3,6,9]
#reference variable outside of function
global X_train
#reshape from (11,) to (11,1) so poly transformation works
X_train=X_train.reshape(-1,1)
#new x values you will plug in, reshape so it works below
new_X=np.linspace(0,10,100).reshape(-1,1)
#create array to which you will add predictions
predictions=np.array([])
for i in degrees:
    poly = PolynomialFeatures(degree=i)
    X_train_poly = poly.fit_transform(X_train)
    linreg = LinearRegression().fit(X_train_poly, y_train)
    poly_prediction=linreg.predict(poly.fit_transform(new_X))
    predictions=np.append(predictions, poly_prediction)
predictions.shape=(4, 100)
return predictions
#answer_one()
```

In [72]:

```python
# feel free to use the function plot_one() to replicate the figure
# from the prompt once you have completed question one
def plot_one(degree_predictions):
    import matplotlib.pyplot as plt
    %matplotlib notebook
    plt.figure(figsize=(10,5))
    plt.plot(X_train, y_train, 'o', label='training data', markersize=10)
    plt.plot(X_test, y_test, 'o', label='test data', markersize=10)
    #for index, actual value in an index, value pair (which is what enumerate returns)
    for i,degree in enumerate([1,3,6,9]):
        plt.plot(np.linspace(0,10,100), degree_predictions[i], alpha=0.8, lw=2, label='degree={}'.f
ormat(degree))
    plt.ylim(-1,2.5)
    plt.legend(loc=4)

#plot_one(answer_one())
```

Question 2

Write a function that fits a polynomial LinearRegression model on the training data `X_train` for degrees 0 through 9. For each model compute the R^2 (coefficient of determination) regression score on the training data as well as the the test data, and return both of these arrays in a tuple.

This function should return one tuple of numpy arrays `(r2_train, r2_test)`. *Both arrays should have shape* `(10,)`

In [73]:

```python
def answer_two():
    from sklearn.linear_model import LinearRegression
    from sklearn.preprocessing import PolynomialFeatures
    from sklearn.metrics.regression import r2_score
    degrees=list(range(0,10))
    global X_train
    global X_test
    global y_train
    global y_test
    #reshape from (11,) to (11,1) so poly transformation works
    X_train=X_train.reshape(-1,1)
    X_test=X_test.reshape(-1,1)
    degrees=list(range(0,10))
    #create array to which you will add predictions
    train_r2=np.array([])
    test_r2=np.array([])
    for i in degrees:
        poly = PolynomialFeatures(degree=i)
        X_train_poly = poly.fit_transform(X_train)
        linreg = LinearRegression().fit(X_train_poly, y_train)
        train_poly_prediction=linreg.predict(poly.fit_transform(X_train))
        test_poly_prediction=linreg.predict(poly.fit_transform(X_test))
        r2_train=r2_score(y_train, train_poly_prediction)
```

```
        r2_test=r2_score(y_test, test_poly_prediction)
        train_r2=np.append(train_r2, r2_train)
        test_r2=np.append(test_r2, r2_test)
    return (train_r2, test_r2)
#answer_two()
```

Question 3

Based on the R^2 scores from question 2 (degree levels 0 through 9), what degree level corresponds to a model that is underfitting? What degree level corresponds to a model that is overfitting? What choice of degree level would provide a model with good generalization performance on this dataset?

Hint: Try plotting the R^2 scores from question 2 to visualize the relationship between degree level and R^2. Remember to comment out the import matplotlib line before submission.

This function should return one tuple with the degree values in this order: `(Underfitting, Overfitting,`
`Good_Generalization)`. *There might be multiple correct solutions, however, you only need to return one possible solution, for example, (1,2,3).*

In [74]:

```
#def answer_three():
    #degrees=list(range(0,10))
    #x=answer_two()
    #import matplotlib.pyplot as plt
    #%matplotlib notebook
    #plt.figure(figsize=(10,5))
    #plt.plot(degrees, x[0], label='Train R Sq.')
    #plt.plot(degrees, x[1], label='Test R Sq.')
    #plt.legend(loc=2)
#answer_three()
```

In [75]:

```
def answer_three():

    degrees=list(range(0,10))
    x=answer_two()
    train=list(x[0])
    test=list(x[1])
    #empty lists for degrees which are underfitting, overfitting
    underfitting_degrees=[]
    overfitting_degrees=[]
    #what is the highest R Squared on test data
    highest_test_r2=max(test)
    #what is the index (degree) corresponding to the highest R Squared
    highest_test_r2_degree=test.index(highest_test_r2)
    for i in degrees:
        #what degrees have R squared less than 0.5 on training data (underfitting)
        if train[i]<.5:
            underfitting_degrees.append(i)
        #what degrees of the test data are past the degree with the highest R Squared
(overfitting)
        if i>highest_test_r2_degree:
            overfitting_degrees.append(i)
    #tuple of one of the underfitting degrees, one of the overfitting degrees, and the degree that
generalizes best
    return (underfitting_degrees[1], overfitting_degrees[1], highest_test_r2_degree)
#answer_three()
```

Question 4

Training models on high degree polynomial features can result in overly complex models that overfit, so we often use regularized versions of the model to constrain model complexity, as we saw with Ridge and Lasso linear regression.

For this question, train two models: a non-regularized LinearRegression model (default parameters) and a regularized Lasso Regression model (with parameters `alpha=0.01, max_iter=10000`) both on polynomial features of degree 12. Return the R^2 score for both the LinearRegression and Lasso model's test sets.

This function should return one tuple `(LinearRegression_R2_test_score, Lasso_R2_test_score)`

```
def answer_four():
    from sklearn.preprocessing import PolynomialFeatures
    from sklearn.linear_model import Lasso, LinearRegression
    from sklearn.metrics.regression import r2_score
    global X_train
    global X_test
    global y_train
    global y_test
    poly = PolynomialFeatures(degree=12)
    X_train_poly = poly.fit_transform(X_train)
    linreg = LinearRegression().fit(X_train_poly, y_train)
    non_reg_prediction=linreg.predict(poly.fit_transform(X_test))
    #the R squared will be terrible because the predictions are worse than just using the mean
    LinearRegression_R2_test_score=r2_score(y_test, non_reg_prediction)
    linlasso = Lasso(alpha=0.01, max_iter = 10000).fit(X_train_poly, y_train)
    lasso_prediction=linlasso.predict(poly.fit_transform(X_test))
    Lasso_R2_test_score=r2_score(y_test, lasso_prediction)

    return (LinearRegression_R2_test_score, Lasso_R2_test_score)
#answer_four()
```

Part 2 - Classification

Here's an application of machine learning that could save your life! For this section of the assignment we will be working with the UCI Mushroom Data Set stored in `mushrooms.csv`. The data will be used to train a model to predict whether or not a mushroom is poisonous. The following attributes are provided:

Attribute Information:

1. cap-shape: bell=b, conical=c, convex=x, flat=f, knobbed=k, sunken=s
2. cap-surface: fibrous=f, grooves=g, scaly=y, smooth=s
3. cap-color: brown=n, buff=b, cinnamon=c, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y
4. bruises?: bruises=t, no=f
5. odor: almond=a, anise=l, creosote=c, fishy=y, foul=f, musty=m, none=n, pungent=p, spicy=s
6. gill-attachment: attached=a, descending=d, free=f, notched=n
7. gill-spacing: close=c, crowded=w, distant=d
8. gill-size: broad=b, narrow=n
9. gill-color: black=k, brown=n, buff=b, chocolate=h, gray=g, green=r, orange=o, pink=p, purple=u, red=e, white=w, yellow=y
10. stalk-shape: enlarging=e, tapering=t
11. stalk-root: bulbous=b, club=c, cup=u, equal=e, rhizomorphs=z, rooted=r, missing=?
12. stalk-surface-above-ring: fibrous=f, scaly=y, silky=k, smooth=s
13. stalk-surface-below-ring: fibrous=f, scaly=y, silky=k, smooth=s
14. stalk-color-above-ring: brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
15. stalk-color-below-ring: brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
16. veil-type: partial=p, universal=u
17. veil-color: brown=n, orange=o, white=w, yellow=y
18. ring-number: none=n, one=o, two=t
19. ring-type: cobwebby=c, evanescent=e, flaring=f, large=l, none=n, pendant=p, sheathing=s, zone=z
20. spore-print-color: black=k, brown=n, buff=b, chocolate=h, green=r, orange=o, purple=u, white=w, yellow=y
21. population: abundant=a, clustered=c, numerous=n, scattered=s, several=v, solitary=y
22. habitat: grasses=g, leaves=l, meadows=m, paths=p, urban=u, waste=w, woods=d

The data in the mushrooms dataset is currently encoded with strings. These values will need to be encoded to numeric to work with sklearn. We'll use pd.get_dummies to convert the categorical variables into indicator variables.

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split


mush_df = pd.read_csv('mushrooms.csv')
mush_df2 = pd.get_dummies(mush_df)
```

```
X_mush = mush_df2.iloc[:,2:]
y_mush = mush_df2.iloc[:,1]

# use the variables X_train2, y_train2 for Question 5
X_train2, X_test2, y_train2, y_test2 = train_test_split(X_mush, y_mush, random_state=0)

# For performance reasons in Questions 6 and 7, we will create a smaller version of the
# entire mushroom dataset for use in those questions.  For simplicity we'll just re-use
# the 25% test split created above as the representative subset.
#
# Use the variables X_subset, y_subset for Questions 6 and 7.
X_subset = X_test2
y_subset = y_test2
```

Question 5

Using `X_train2` and `y_train2` from the preceeding cell, train a DecisionTreeClassifier with default parameters and random_state=0. What are the 5 most important features found by the decision tree?

As a reminder, the feature names are available in the `X_train2.columns` property, and the order of the features in `X_train2.columns` matches the order of the feature importance values in the classifier's `feature_importances_` property.

This function should return a list of length 5 containing the feature names in descending order of importance.

Note: remember that you also need to set random_state in the DecisionTreeClassifier.

In [87]:
```
def answer_five():
    from sklearn.tree import DecisionTreeClassifier
    clf = DecisionTreeClassifier(random_state=0).fit(X_train2, y_train2)
    #put feature importances into new variable
    ind=clf.feature_importances_
    #sort feature importances by value but return the indexes.
    #Because argsort only does ascending, use negated values so you are actually going by
descending
    #Take top 5
    top5=list(np.argsort(-ind)[:5])
    #return list of columns that correspond to the indexes of those top 5
    return list(X_train2.columns[top5])
#answer_five()
```

Question 6

For this question, we're going to use the `validation_curve` function in `sklearn.model_selection` to determine training and test scores for a Support Vector Classifier (`SVC`) with varying parameter values. Recall that the validation_curve function, in addition to taking an initialized unfitted classifier object, takes a dataset as input and does its own internal train-test splits to compute results.

Because creating a validation curve requires fitting multiple models, for performance reasons this question will use just a subset of the original mushroom dataset: please use the variables X_subset and y_subset as input to the validation curve function (instead of X_mush and y_mush) to reduce computation time.

The initialized unfitted classifier object we'll be using is a Support Vector Classifier with radial basis kernel. So your first step is to create an `SVC` object with default parameters (i.e. `kernel='rbf', C=1`) and `random_state=0`. Recall that the kernel width of the RBF kernel is controlled using the `gamma` parameter.

With this classifier, and the dataset in X_subset, y_subset, explore the effect of `gamma` on classifier accuracy by using the `validation_curve` function to find the training and test scores for 6 values of `gamma` from `0.0001` to `10` (i.e. `np.logspace(-4,1,6)`). Recall that you can specify what scoring metric you want validation_curve to use by setting the "scoring" parameter. In this case, we want to use "accuracy" as the scoring metric.

For each level of `gamma`, `validation_curve` will fit 3 models on different subsets of the data, returning two 6x3 (6 levels of gamma x 3 fits per level) arrays of the scores for the training and test sets.

Find the mean score across the three models for each level of `gamma` for both arrays, creating two arrays of length 6, and return a tuple with the two arrays.

e.g.

if one of your array of scores is

```
array([[ 0.5,   0.4,   0.6],
       [ 0.7,   0.8,   0.7],
       [ 0.9,   0.8,   0.8],
       [ 0.8,   0.7,   0.8],
       [ 0.7,   0.6,   0.6],
       [ 0.4,   0.6,   0.5]])
```

it should then become

```
array([ 0.5,  0.73333333,  0.83333333,  0.76666667,  0.63333333, 0.5])
```

This function should return one tuple of numpy arrays `(training_scores, test_scores)` *where each array in the tuple has shape* `(6,)`.

In [79]:

```python
def answer_six():
    from sklearn.svm import SVC
    from sklearn.model_selection import validation_curve
    #because it is logspace: -4=0.0001 and 1=10
    param_range = np.logspace(-4, 1, 6)
    train_scores, test_scores = validation_curve(SVC(), X_subset, y_subset, param_name='gamma', par
am_range=param_range, cv=3)
    train_scores=np.mean(train_scores, axis=1)
    test_scores=np.mean(test_scores, axis=1)
    return (train_scores, test_scores)
#answer_six()
```

Question 7

Based on the scores from question 6, what gamma value corresponds to a model that is underfitting (and has the worst test set accuracy)? What gamma value corresponds to a model that is overfitting (and has the worst test set accuracy)? What choice of gamma would be the best choice for a model with good generalization performance on this dataset (high accuracy on both training and test set)?

Hint: Try plotting the scores from question 6 to visualize the relationship between gamma and accuracy. Remember to comment out the import matplotlib line before submission.

This function should return one tuple with the degree values in this order: `(Underfitting, Overfitting, Good_Generalization)` *Please note there is only one correct solution.*

In [80]:

```python
def answer_seven():
    x=answer_six()
    degrees=np.logspace(-4, 1, 6)
    train=list(x[0])
    test=list(x[1])
    underfitting=train.index(min(train))
    overfitting=5
    good_generalization=3
    return (degrees[underfitting], degrees[overfitting], degrees[good_generalization])
#answer_seven()
```