Tema 2: Aprendizaje Supervisado

Regresión Lineal Regularizada I

Prof. Wladimir Rodríguez

wladimir@ula.ve

Departamento de Computación

Demostración de sobreajuste sobre datos sintéticos

Crear un conjunto de datos basado en una función sinusoidal

Veamos un conjunto de datos sintéticos que consta de 30 puntos extraídos de la sinusoide $y = \sin(4x)$:

```
import math
import random
import numpy as np
import pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.utils import shuffle
from sklearn.metrics import mean_squared_error
from matplotlib import pyplot as plt
%matplotlib inline
```

Crear valores aleatorios para x en el intervalo [0,1]

```
In [2]: np.random.seed(98103)
    n = 30
    x = np.random.uniform(0, 1, size=n)
    x = np.sort(x)
```

Calcular y. Y agregar ruido gaussiano aleatorio a y

```
In [3]: def f(x):
    return np.sin(4 * x)
y = f(x) + np.random.normal(scale=0.3, size=n)
```

Crear un dataframe de pandas.

```
In [4]: data = pd.DataFrame({'x':x, 'y':y})
   data.head()
```

```
        0ut[4]:
        x
        y

        0
        0.007306
        0.168648

        1
        0.031659
        0.291408

        2
        0.036159
        0.700509

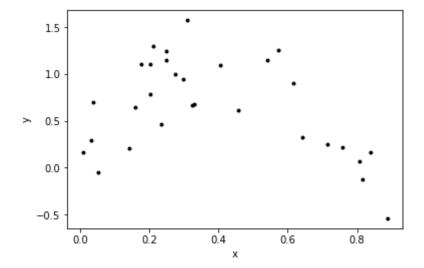
        3
        0.052886
        -0.049021

        4
        0.140328
        0.207561
```

Crea una función para graficar los datos, ya que lo haremos muchas veces

```
In [5]: def graficar_data(data):
    plt.plot(data['x'],data['y'],'k.')
    plt.xlabel('x')
    plt.ylabel('y')

graficar_data(data)
```



Definir algunas funciones útiles de regresión polinomial

Función para generar los dataframe polinomiales

```
In [6]:
        def polinomial_dataframe(atributo, grado):
            # asumir que grado es >= 1
            # inicializar el DataFrame:
            poli dataframe = pd.DataFrame()
            # fijar poli_dataframe['potencia_1'] igual al atributo pasado
            poli_dataframe['potencia_1'] = atributo
            # chequear si grado > 1
            if grado > 1:
                # realizar un lazo con los grados restantes:
                for potencia in range(2, grado+1):
                    # primero le damos el nombre a la columna:
                    nombre = 'potencia_' + str(potencia)
                    # luego asignamos a poli_dataframe[nombre] la potencia del atributo apropiada
                    poli_dataframe[nombre] = atributo**potencia
            return poli_dataframe
```

Definir una función para ajustar un modelo de regresión lineal polinomial de grado "grado" a los datos en "data":

```
In [7]: def regresion_polinomial(data, grado):
    poli_data_X = polinomial_dataframe(data.x, grado)
    modelo = LinearRegression()
    modelo.fit(poli_data_X, y)
    return modelo
```

Definir la función para graficar los datos y las predicciones hechas, ya que vamos a usarlo muchas veces.

```
def graficar_predicciones_poly(data, modelo):
In [8]:
            graficar_data(data)
            # Graficar la verdadera relación entre X e y
            x_v = np.random.uniform(0, 1, size=200)
            x_v = np.sort(x_v)
            y_v = f(x_v)
            plt.plot(x_v, y_v, 'r-')
            # Obtener el grado del polinomio
            grado = len(modelo.coef_)
            # Crear 200 puntos en el eje x axis y calcular la predicción para cada punto
            x = np.random.uniform(0, 1, size=200)
            x = np.sort(x)
            x pred = pd.DataFrame({'x': x})
            y pred = modelo.predict(polinomial dataframe(x pred.x,grado))
            # graficar predicciones
            plt.plot(x_pred.x, y_pred, 'g-', label='ajuste de grado ' + str(grado))
            plt.legend(loc='upper left')
            plt.axis([0,1,-1.5,2])
```

Cree una función que imprima los coeficientes polinomiales de una manera bonita:)

```
In [9]: def imprimir_coefficientes(modelo):
    # Obtener el grado del polinomio
    grado = len(modelo.coef_)

# Obtener los parámetros aprendidos como una lista
    w = [modelo.intercept_]
    w += (modelo.coef_).tolist()
    # Numpy tiene un a función para imprimir polinomios de manera elegante
    # (La usaremos, pero necesita los parámetros en orden inverso)
    print ('Polinomio de grado ' + str(grado) + ':')
    w.reverse()
    print (np.poly1d(w))
```

Ajustar un polinomio de grado 2

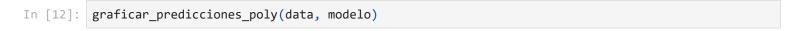
```
In [10]: modelo = regresion_polinomial(data, grado=2)
```

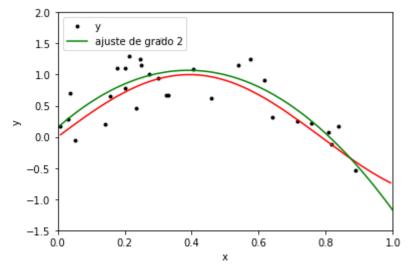
Inspeccionar los parámetros aprendidos

```
In [11]: imprimir_coefficientes(modelo)

Polinomio de grado 2:
2
-6.047 x + 4.717 x + 0.1494
```

Formar y graficar nuestras predicciones a lo largo de una cuadrícula de valores x:





Calcular la media del error al cuadrado (MSE = mean squared error)

```
In [ ]: print('MSE = ', mean_squared_error(data.y, modelo.predict(polinomial_dataframe(data.x,2))))
```

Ajustar un polinomio de grado 4

```
modelo = regresion_polinomial(data, grado=4)
In [13]:
          imprimir_coefficientes(modelo)
          graficar_predicciones_poly(data, modelo)
          print('MSE = ', mean squared error(data.y, modelo.predict(polinomial dataframe(data.x,4))))
          Polinomio de grado 4:
                              3
          -1.366 \times + 3.125 \times - 8.299 \times + 5.259 \times + 0.1204
                  0.0845842722706
              2.0
                       ajuste de grado 4
              1.5
              1.0
              0.5
              0.0
             -0.5
             -1.0
```

0.8

1.0

Ajustar un polinomio de grado 8

0.4

0.6

0.2

-1.5

0.0

```
In [14]:
          modelo = regresion_polinomial(data, grado=8)
           imprimir_coefficientes(modelo)
           graficar_predicciones_poly(data, modelo)
           print('MSE = ', mean_squared_error(data.y, modelo.predict(polinomial_dataframe(data.x,8))))
           Polinomio de grado 8:
           -1.192e+04 \times + 3.974e+04 \times - 5.286e+04 \times + 3.558e+04 \times - 1.261e+04 \times
            + 2176 \times - 135.5 \times + 1.609 \times + 0.2985
          MSE = 0.0682028308328
               2.0
                        ajuste de grado 8
               1.5
               1.0
               0.5
               0.0
             -0.5
             -1.0
             -1.5
                 0.0
                            0.2
                                       0.4
                                                  0.6
                                                             0.8
                                                                        1.0
```

Ajustar un polinomio de grado 16

-1.0

-1.5

0.0

0.2

0.4

0.6

0.8

1.0

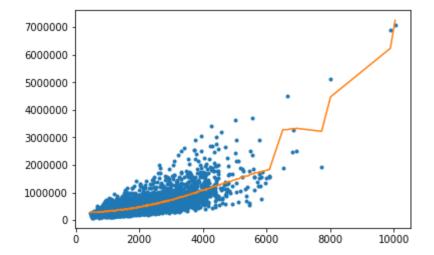
```
In [15]:
          modelo = regresion_polinomial(data, grado=16)
          imprimir coefficientes(modelo)
          graficar_predicciones_poly(data, modelo)
          print('MSE = ', mean_squared_error(data.y, modelo.predict(polinomial_dataframe(data.x,16))))
          Polinomio de grado 16:
                                                                                       12
                      16
                                       15
                                                      14
                                                                      13
          7.098e+09 x
                        -4.998e+10 \times +1.6e+11 \times
                                                        - 3.079e+11 x
                                                                        + 3.977e + 11 x
                                         10
                                           -1.201e+11 x + 4.404e+10 x
           -3.64e+11 \times + 2.431e+11 \times
           -1.191e+10 \times + 2.339e+09 \times -3.238e+08 \times +3.017e+07 \times -1.756e+06 \times
           + 5.663e+04 x - 837.4 x + 3.866
                  0.0405643635189
          MSE =
              2.0
                       ajuste de grado 16
              1.5
              1.0
              0.5
              0.0
             -0.5
```

Demostración de sobreajuste sobre datos reales

En primer lugar, dividir los datos de ventas en cuatro subconjuntos de aproximadamente el mismo tamaño y llamarlos ventas_1, ventas_2, ventas_3 y ventas_4

Ajustar un modelo polinomial de grado 15 al conjunto ventas_1

```
In [18]: X_1 = ventas_1.sort_values(['sqft_living', 'price'])
          poli_data_X = polinomial_dataframe(X_1.sqft_living, 15)
          y_1 = X_1.price
In [19]: modelo = LinearRegression()
          modelo.fit(poli data X, y 1)
          imprimir_coefficientes(modelo)
          print()
          mse = mean_squared_error(y_1, modelo.predict(poli_data_X))
          print('MSE = ', mse)
          print('RMSE = ', math.sqrt(mse))
          plt.plot(poli_data_X['potencia_1'],y_1,'.',
                  poli_data_X['potencia_1'], modelo.predict(poli_data_X),'-')
          Polinomio de grado 15:
                     15
                                     14
                                                     13
          6.453e-16 \times - 8.197e-16 \times + 6.661e-16 \times + 1.151e-15 \times
                                        10
                        11
           - 2.229e-16 x - 6.141e-16 x - 3.903e-17 x - 5.627e-16 x
                                      6
                                                    5
           - 1.15e-15 \times + 4.649e-16 \times - 7.328e-14 \times - 6.789e-11 \times + 2.166e-06 \times
          + 0.04793 \times + 2.197e-05 \times + 2.703e+05
          MSE = 57776932071.8
          RMSE = 240368.32584972732
          [<matplotlib.lines.Line2D at 0x11496a518>,
           <matplotlib.lines.Line2D at 0x11496a160>]
```



Ajustar un modelo polinomial de grado 15 al conjunto ventas_2

```
In [20]: X_2 = ventas_2.sort_values(['sqft_living', 'price'])
          poli_data_X_2 = polinomial_dataframe(X_2.sqft_living, 15)
          y_2 = X_2.price
          modelo = LinearRegression()
In [21]:
          modelo.fit(poli_data_X_2, y_2)
          imprimir_coefficientes(modelo)
          print()
          mse = mean_squared_error(y_2, modelo.predict(poli_data_X_2))
          print('MSE = ', mse)
          print('RMSE = ', math.sqrt(mse))
          plt.plot(poli_data_X_2['potencia_1'],y_2,'.',
                   poli_data_X_2['potencia_1'], modelo.predict(poli_data_X_2),'-')
          Polinomio de grado 15:
                       15
          -1.145e-15 x
                        -7.633e-16 \times +8.048e-16 \times +1.171e-15 \times
                         11
                                       10
           - 1.561e-16 x
                           - 1.7e-16 x
                                         -9.394e-16 x + 1.034e-15 x + 8.066e-16 x
           -1.693e-15 \times -3.123e-14 \times -3.513e-10 \times +2.508e-06 \times +0.04928 \times +1.982e-05 \times +2.723e+05
          MSE = 62718455928.6
          RMSE = 250436.530738969
          [<matplotlib.lines.Line2D at 0x11261a748>,
Out[21]:
           <matplotlib.lines.Line2D at 0x11469a9b0>]
           8000000
           7000000
           6000000
           5000000
           4000000
           3000000
           2000000
          1000000
                0
                        2000
                                40<u>0</u>0
                                      6000
                                             8000
                                                    10000
                                                           12000
                                                                  14000
```

Ajustar un modelo polinomial de grado 15 al conjunto ventas_3

```
In [22]: X_3 = ventas_3.sort_values(['sqft_living', 'price'])
                                         poli_data_X_3 = polinomial_dataframe(X_3.sqft_living, 15)
                                        y_3 = X_3.price
                                        modelo = LinearRegression()
 In [23]:
                                        modelo.fit(poli_data_X_3, y_3)
                                         imprimir_coefficientes(modelo)
                                        print()
                                        mse = mean_squared_error(y_3, modelo.predict(poli_data_X_3))
                                         print('MSE = ', mse)
                                        print('RMSE = ', math.sqrt(mse))
                                         plt.plot(poli_data_X_3['potencia_1'],y_3,'.',
                                                                          poli_data_X_3['potencia_1'], modelo.predict(poli_data_X_3),'-')
                                        Polinomio de grado 15:
                                                                                      15
                                                                                                                                                    14
                                                                                                                                                                                                                  13
                                                                                                                                                                                                                                                                                 12
                                        1.027e-15 \times + 5.109e-16 \times + 1.287e-15 \times + 8.049e-16 \times
                                                                                                                                                               10
                                            + 8.916e-16 \times - 2.585e-16 \times + 1.344e-16 \times + 8.483e-16 \times
                                             -9.541e-16 \times +1.037e-15 \times -5.893e-15 \times -1.112e-09 \times +8.921e-06 \times -1.037e-15 \times -1.
                                            + 0.03712 x + 1.852e-05 x + 2.769e+05
                                        MSE = 60135048052.3
                                        RMSE = 245224.4850178539
                                        [<matplotlib.lines.Line2D at 0x114777080>,
Out[23]:
                                            <matplotlib.lines.Line2D at 0x114777be0>]
                                          4000000
                                          3000000
                                          2000000
                                         1000000
                                                                                                            2000
                                                                                                                                                      4000
                                                                                                                                                                                                 6000
                                                                                                                                                                                                                                           8000
```

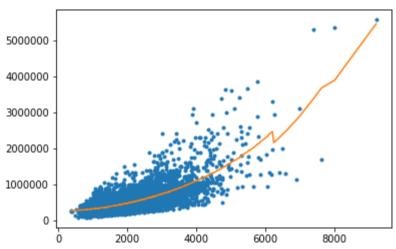
Ajustar un modelo polinomial de grado 15 al conjunto ventas_4

Polinomio de grado 15:

15
14
13
12
-3.331e-16 x - 2.845e-16 x - 2.22e-16 x - 1.362e-16 x
11
10
9
8
+ 1.596e-16 x + 3.521e-16 x + 1.908e-17 x + 8.153e-17 x
7
6
5
4
- 3.227e-16 x - 5.065e-16 x + 1.849e-14 x + 4.909e-10 x - 4.145e-06 x
2
+ 0.05959 x + 2.906e-05 x + 2.663e+05

MSE = 61395360069.8
RMSE = 247780.87107322144

Out[25]:



Regresión Ridge

La Regresión Ridge tiene como objetivo evitar el sobreajuste añadiendo un coste al término RSS de mínimos cuadrados estándar que depende de la norma 2 de los coeficientes ||w||. El resultado es penalizar ajustes con grandes coeficientes. La fuerza de esta penalización, y por lo tanto el balance de complejidad vs. complejidad del modelo, se controla mediante un parámetro α (aquí llamado "Penalidad_L2").

$$J(oldsymbol{eta}) = rac{1}{n} \sum_{i=0}^n (y_i - oldsymbol{eta}^T \mathbf{x}_i')^2 + lpha \|oldsymbol{eta}\|_2$$

• α = 0:

El objetivo se vuelve igual que la simple regresión lineal. Obtendremos los mismos coeficientes que la regresión lineal simple.

• $\alpha = \infty$:

Los coeficientes serán cero. ¿Por qué? Debido a una ponderación infinita del cuadrado de coeficientes, cualquier cosa menos de cero hará que el objetivo sea infinito.

• $0 < \alpha < \infty$:

La magnitud de \$\alpha\$ determinará la ponderación dada a diferentes partes del objetivo.

Los coeficientes estarán entre 0 y unos para la regresión lineal simple.

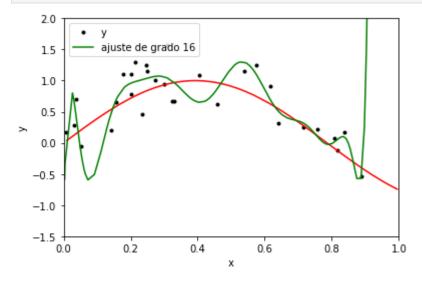
Definir nuestra función para resolver el objetivo de la Regresión Ridge para un modelo de regresión polinomial de cualquier grado:

```
In [27]: def regresion_ridge_polinomial(data, grado, penalidad_12):
    poli_data_X = polinomial_dataframe(data.x, grado)
    modelo = Ridge(alpha=penalidad_12)
    modelo.fit(poli_data_X, y)
    return modelo
```

Realizar un ajuste de Regresion Ridge de un polinomio de grado 16 usando una penalidad muy pequeña

```
modelo = regresion_ridge_polinomial(data, grado=16, penalidad_12=1e-25)
In [28]:
          imprimir_coefficientes(modelo)
         Polinomio de grado 16:
                                    15
                                                                   13
         6.223e+06 x - 1.258e+07 x - 2.801e+07 x + 1.278e+08 x
                                       11
           - 2.056e+08 x + 2.158e+08 x - 2.072e+08 x + 1.995e+08 x
                                      7
                                                     6
           -1.597e+08 \times + 9.217e+07 \times -3.652e+07 \times + 9.712e+06 \times -1.681e+06 \times
          + 1.784e+05 x - 1.028e+04 x + 250.2 x - 1.215
         /Users/wladimir/anaconda/lib/python3.5/site-packages/scipy/linalg/basic.py:223: RuntimeWarning:
         scipy.linalg.solve
         Ill-conditioned matrix detected. Result is not guaranteed to be accurate.
         Reciprocal condition number: 6.24183070750001e-20
            ' condition number: {}'.format(rcond), RuntimeWarning)
```

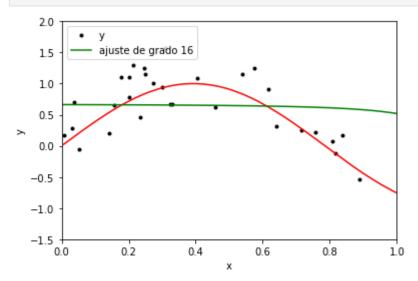
In [29]: graficar_predicciones_poly(data, modelo)



Realizar un ajuste de Regresion Ridge de un polinomio de grado 16 usando una penalidad alta

```
In [30]: modelo = regresion_ridge_polinomial(data, grado=16, penalidad_12=100)
   imprimir_coefficientes(modelo)
```

```
In [31]: graficar_predicciones_poly(data, modelo)
```

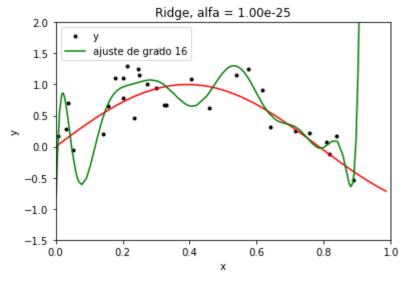


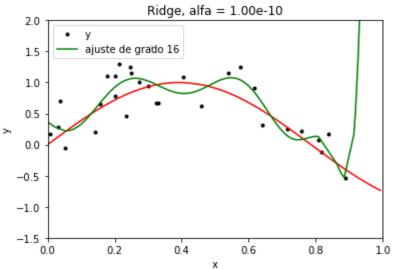
Veamos los ajustes para una secuencia de valores alfa en aumento

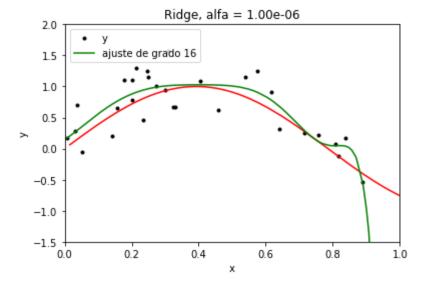
```
Polinomio de grado 16:
                                                 15
                                                                                 14
6.223e+06 x - 1.258e+07 x - 2.801e+07 x + 1.278e+08 x
                      12 11 10 9
  - 2.056e+08 x + 2.158e+08 x - 2.072e+08 x + 1.995e+08 x
                8 7 6 5
  -1.597e+08 \times +9.217e+07 \times -3.652e+07 \times +9.712e+06 \times -1.681e+06 \times -1.081e+06 \times -1.
 + 1.784e+05 x - 1.028e+04 x + 250.2 x - 1.215
alpha = 1.00e-10
Polinomio de grado 16:
  16 15 14 13 12 11
509.6 x + 2322 x - 9.445 x - 3071 x - 3804 x - 932.9 x + 3782 x
               9 8 7 6 5 4 3
  + 5767 x + 1044 x - 7020 x - 4838 x + 1.063e+04 x - 5191 x + 887.6 x
  -11.8 \times -3.863 \times +0.3649
alpha = 1.00e-06
Polinomio de grado 16:
      16 15 14 13 12 11
-94.93 \times -51.19 \times -6.794 \times +33.06 \times +61.43 \times +70.52 \times
                 10 9 8 7 6 5
 + 54.01 x + 11.73 x - 43.59 x - 80.63 x - 56.42 x + 40.11 x
                 4 3
 + 98.38 \times - 76.4 \times + 10.76 \times + 3.758 \times + 0.1418
alpha = 1.00e-03
Polinomio de grado 16:
                                                           14 13
       16 15
                                                                                                            12
-1.462 \times -1.203 \times -0.8922 \times -0.5391 \times -0.162 \times +0.2113 \times
                  10 9 8 7 6
 + 0.5438 \times + 0.7927 \times + 0.9163 \times + 0.8822 \times + 0.6616 \times + 0.1788 \times
               4 3 2
  -0.8056 \times -2.607 \times -3.759 \times +4.227 \times +0.1715
alpha = 1.00e+01
Polinomio de grado 16:
                16 15 14 13 12
-0.01697 x - 0.01962 x - 0.02273 x - 0.02639 x - 0.03069 x
                                                10 9 8 7
  - 0.03576 x - 0.04175 x - 0.04882 x - 0.05712 x - 0.0668 x
                     6 5 4 3 2
  -0.07792 \times -0.09022 \times -0.1026 \times -0.1117 \times -0.1064 \times -0.055 \times +0.7415
alpha = 1.00e+02
Polinomio de grado 16:
                                                15
                                                                             14
                                                                                                          13
-0.002467 x - 0.002856 x - 0.003314 x - 0.003853 x - 0.00449 x
          11 10 9 8 7
  - 0.005242 x - 0.006133 x - 0.00719 x - 0.008441 x - 0.009917 x
            6 5 4 3
```

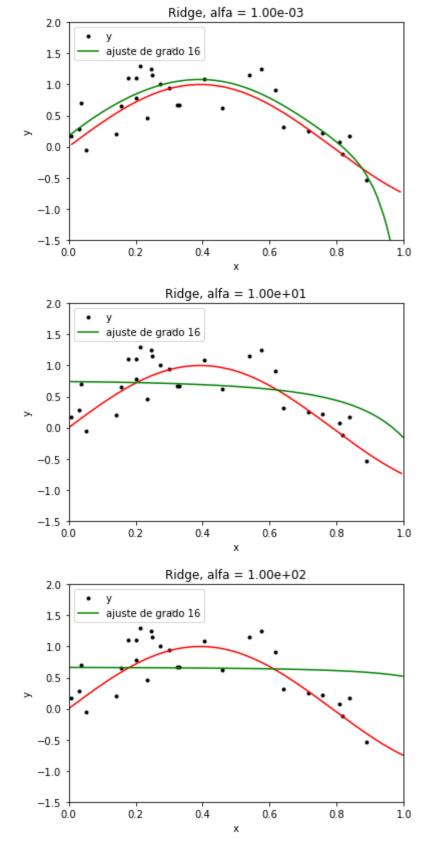
- $0.01164 \times - 0.01362 \times - 0.01575 \times - 0.01769 \times - 0.01816 \times - 0.01309 \times + 0.663$

alpha = 1.00e-25









Regresión Lasso

La Regresión Lasso reduce conjuntamente los coeficientes para evitar el ajuste excesivo, e implica implícitamente la selección de los atributos estableciendo algunos coeficientes exactamente a 0 para una fuerza de penalidad suficientemente grande alfa (aquí llamada "penalidad_L1"). En particular, Lasso toma el término RSS de los mínimos cuadrados estándar y añade un coste de norma 1 de los coeficientes $\|w\|$.

```
In [33]: from sklearn.linear_model import Lasso
```

Definir nuestra función para resolver el objetivo de la Regresión Lasso para un modelo de regresión polinomial de cualquier grado:

```
In [34]: def regresion_lasso_polinomial(data, grado, penalidad_l1):
    poli_data_X = polinomial_dataframe(data.x, grado)
    modelo = Lasso(alpha=penalidad_l1)
    modelo.fit(poli_data_X, y)
    return modelo
```

Explore la solución de la Regresión Lasso en función de diferentes valores de alfa

```
In [35]:
          for alfa in [0.0001, 0.001, 0.01, 0.1, 10]:
              modelo = regresion_lasso_polinomial(data, grado=16, penalidad_l1=alfa)
              print ('alpha = %.2e' % alfa)
              imprimir coefficientes(modelo)
              print ('\n')
              plt.figure()
              graficar_predicciones_poly(data, modelo)
              plt.title('Lasso, alfa = %.2e' % alfa)
          /Users/wladimir/anaconda/lib/python3.5/site-packages/sklearn/linear_model/coordinate_descent.py:
          484: ConvergenceWarning: Objective did not converge. You might want to increase the number of it
          erations. Fitting data with very small alpha may cause precision problems.
            ConvergenceWarning)
          alpha = 1.00e-04
          Polinomio de grado 16:
          0.02452 \times - 0 \times - 6.019 \times + 4.681 \times + 0.1556
          alpha = 1.00e-03
          Polinomio de grado 16:
                  3
          -2.471 \times -2.37 \times +3.207 \times +0.2909
          alpha = 1.00e-02
          Polinomio de grado 16:
               4 3 2
          -1.808 \times -0 \times -0 \times +0.2161 \times +0.7568
          alpha = 1.00e-01
          Polinomio de grado 16:
          0.6461
          alpha = 1.00e+01
          Polinomio de grado 16:
          0.6461
```

