

Technological Institute of the Philippines	Quezon City - Computer Engineering
Course Code:	CPE 312
Code Title:	Predictive Analytics using Machine Learning
1st Semester	AY 2024-2025
<u>ACTIVITY NO.</u>	Activity 2
Name	Pisalbon, Ery Jay P.
Section	CPE31S31
Date Performed:	9/6/2024
Date Submitted:	9/6/2024
Instructor:	Dr. Alonica Villanueva / Engr. Roman M. Richard

```
In [ ]: from google.colab import drive
drive.mount('/content/drive')
```

1. Objectives

This activity aims to:

- Perform regression analysis using linear regression and polynomial regression; and
- Solve classification problem using logistic regression.

2. Intended Learning Outcomes (ILOs)

After this activity, the students should be able to:

- Demonstrate how to use Python to predic outcome using linear and polynomial regression;
- Demonstrate how to use single and multiple features to predict the outcome using linear regression;
- Demonstrate how to train and predict classification model using logistic regression;
- Evaluate and visualize the performance of different regression models.

3. Procedures and Outputs

3.1 Setup

These are configurations for running on your local machine. No problem with running it on your colab. This project requires Python 3.7 or above:

```
In [1]: import sys

assert sys.version_info >= (3, 7)
```

It also requires Scikit-Learn \geq 1.0.1:

```
In [2]: from packaging import version
import sklearn

assert version.parse(sklearn.__version__) >= version.parse("1.0.1")
```

As we did in previous chapters, let's define the default font sizes to make the figures prettier:

```
In [3]: import matplotlib.pyplot as plt

plt.rc('font', size=14)
plt.rc('axes', labelsz=14, titlesz=14)
plt.rc('legend', fontsize=14)
plt.rc('xtick', labelsz=10)
plt.rc('ytick', labelsz=10)
```

And let's create the `images/training_linear_models` folder (if it doesn't already exist), and define the `save_fig()` function which is used through this notebook to save the figures in high-res for the book:

```
In [4]: from pathlib import Path

IMAGES_PATH = Path() / "images" / "training_linear_models"
IMAGES_PATH.mkdir(parents=True, exist_ok=True)

def save_fig(fig_id, tight_layout=True, fig_extension="png", resolution=300):
    path = IMAGES_PATH / f"{fig_id}.{fig_extension}"
    if tight_layout:
        plt.tight_layout()
    plt.savefig(path, format=fig_extension, dpi=resolution)
```

3.2 Linear Regression

A linear model makes a prediction by simply computing a weighted sum of the input features, plus a constant called the *bias term* or the *intercept term*.

3.2.1 The Normal Equation

```
In [5]: import numpy as np

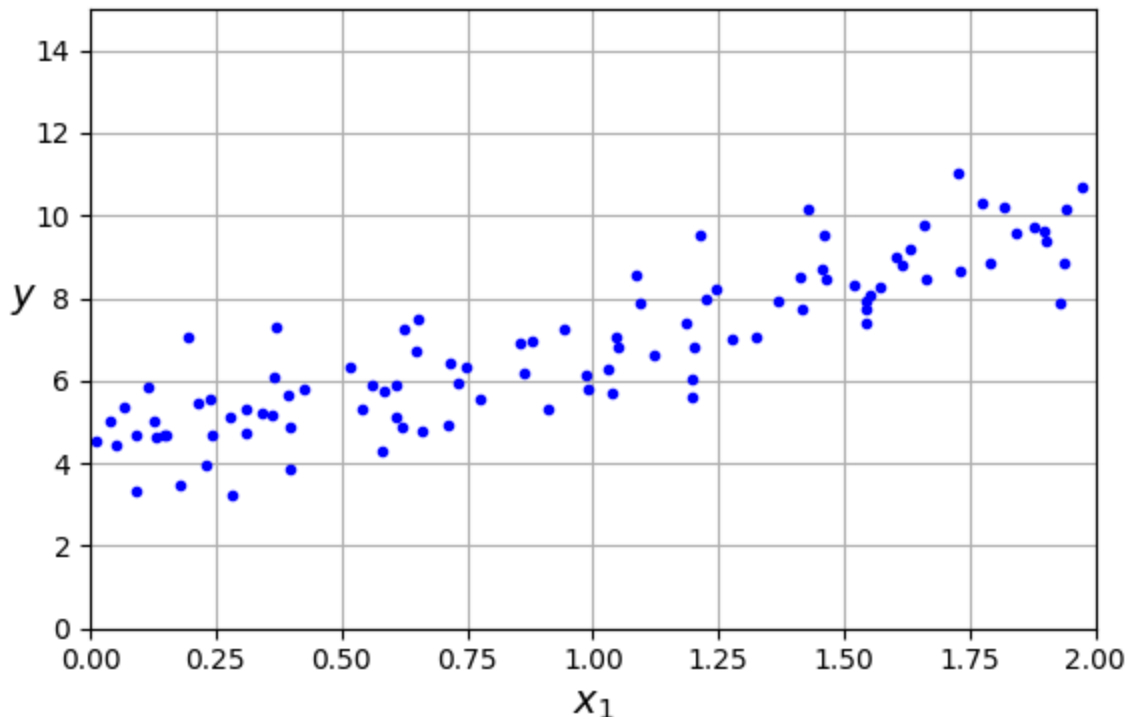
np.random.seed(42) # to make this code example reproducible
m = 100 # number of instances
```

```
X = 2 * np.random.rand(m, 1) # column vector
y = 4 + 3 * X + np.random.randn(m, 1) # column vector
```

In [6]: # extra code - generates and saves Figure 4-1

```
import matplotlib.pyplot as plt

plt.figure(figsize=(6, 4))
plt.plot(X, y, "b.")
plt.xlabel("$x_1$")
plt.ylabel("$y$", rotation=0)
plt.axis([0, 2, 0, 15])
plt.grid()
save_fig("generated_data_plot")
plt.show()
```



In [7]: from sklearn.preprocessing import add_dummy_feature

```
X_b = add_dummy_feature(X) # add x0 = 1 to each instance
theta_best = np.linalg.inv(X_b.T @ X_b) @ X_b.T @ y
```

In [8]: theta_best

Out[8]: array([[4.21509616],
[2.77011339]])

```
In [9]: X_new = np.array([[0], [2]])
X_new_b = add_dummy_feature(X_new) # add x0 = 1 to each instance
y_predict = X_new_b @ theta_best
y_predict
```

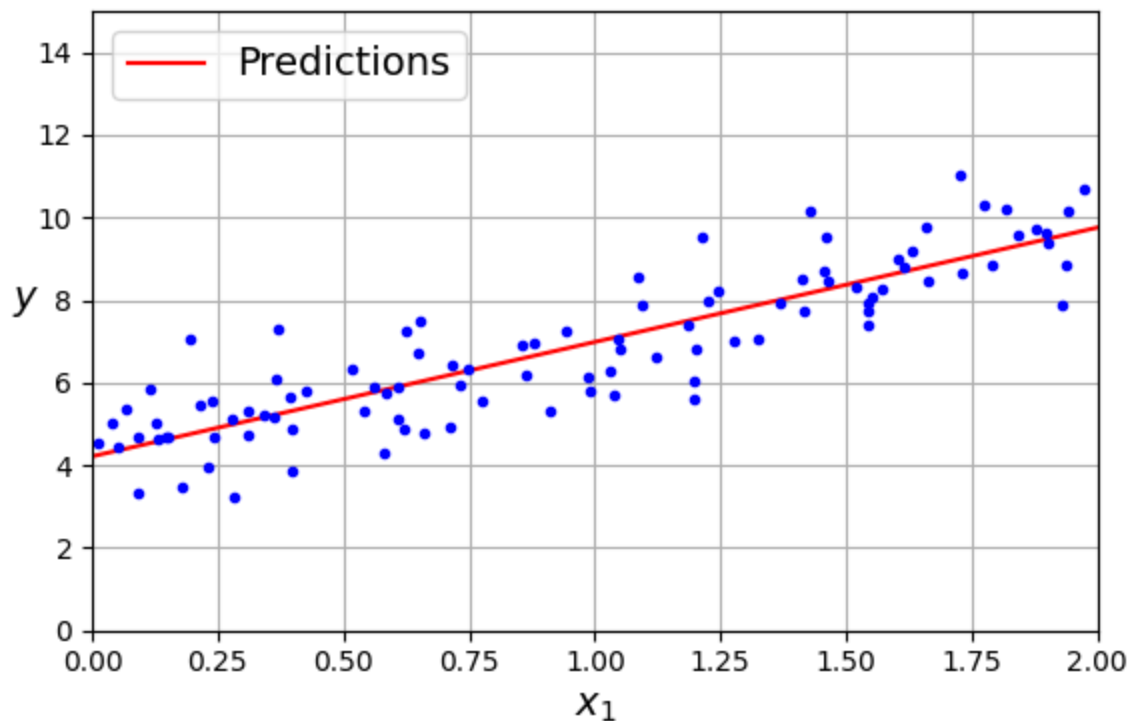
```
Out[9]: array([[4.21509616],
               [9.75532293]])
```

```
In [10]: import matplotlib.pyplot as plt

plt.figure(figsize=(6, 4)) # extra code - not needed, just formatting
plt.plot(X_new, y_predict, "r-", label="Predictions")
plt.plot(X, y, "b.")

# extra code - beautifies and saves Figure 4-2
plt.xlabel("$x_1$")
plt.ylabel("$y$", rotation=0)
plt.axis([0, 2, 0, 15])
plt.grid()
plt.legend(loc="upper left")
save_fig("linear_model_predictions_plot")

plt.show()
```



```
In [11]: from sklearn.linear_model import LinearRegression

lin_reg = LinearRegression()
lin_reg.fit(X, y)
lin_reg.intercept_, lin_reg.coef_
```

```
Out[11]: (array([4.21509616]), array([[2.77011339]]))
```

```
In [12]: lin_reg.predict(X_new)
```

```
Out[12]: array([[4.21509616],
               [9.75532293]])
```

The `LinearRegression` class is based on the `scipy.linalg.lstsq()` function (the name stands for "least squares"), which you could call directly:

```
In [13]: theta_best_svd, residuals, rank, s = np.linalg.lstsq(X_b, y, rcond=1e-6)
         theta_best_svd
```

```
Out[13]: array([[4.21509616],
                [2.77011339]])
```

The pseudoinverse is computed using a standard matrix factorization technique called *Singular Value Decomposition*. What does SVD do? Quickly research and answer. SVD is decomposes a matrix into three matrices, those are orthogonal matrix containing left singular vectors, diagonal matrix containing the singular values, orthogonal matrix containing the right singular vectors.

This function computes $\mathbf{X}^+ \mathbf{y}$, where \mathbf{X}^+ is the *pseudoinverse* of \mathbf{X} (specifically the Moore-Penrose inverse). You can use `np.linalg.pinv()` to compute the pseudoinverse directly:

```
In [14]: np.linalg.pinv(X_b) @ y
```

```
Out[14]: array([[4.21509616],
                [2.77011339]])
```

You have multiple methods used to implement the for linear regression. Provide a comparison between the *Singular Value Decomposition* and the Normal Equation Computation. Which is more efficient?

3.2 Gradient Descent

Gradient Descent is a very generic optimization algorithm capable of finding optimal solutions to a wide range of problems. Generally, the idea is to tweak parameters iteratively to minimize a cost function.

3.2.1 Batch Gradient Descent

To implement Gradient Descent, you need to compute the gradient of the cost function with regards to each model parameter θ_j . In other words, you need to calculate how much the cost function will change if you change θ_j just a little bit. This is called a partial derivative. It is like asking "what is the slope of the mountain under my feet if I face east?" and then asking the same question facing north (and so on for all other dimensions, if you can imagine a universe with more than three dimensions).

```
In [15]: eta = 0.1 # Learning rate
         n_epochs = 1000
         m = len(X_b) # number of instances

         np.random.seed(42)
```

```
theta = np.random.randn(2, 1) # randomly initialized model parameters

for epoch in range(n_epochs):
    gradients = 2 / m * X_b.T @ (X_b @ theta - y)
    theta = theta - eta * gradients
```

The trained model parameters:

In [16]: theta

Out[16]: array([[4.21509616],
[2.77011339]])

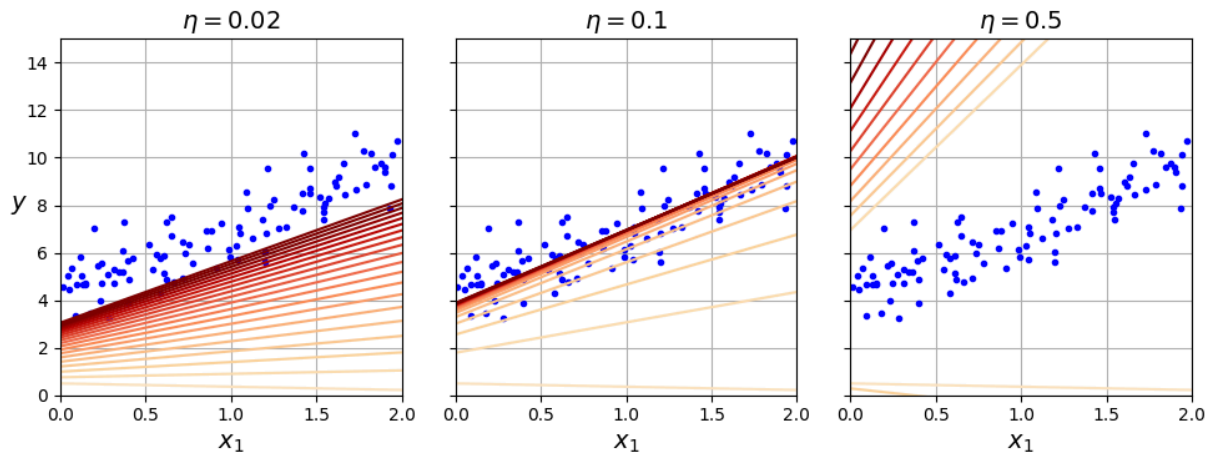
In [17]: # extra code - generates and saves Figure 4-8

```
import matplotlib as mpl

def plot_gradient_descent(theta, eta):
    m = len(X_b)
    plt.plot(X, y, "b.")
    n_epochs = 1000
    n_shown = 20
    theta_path = []
    for epoch in range(n_epochs):
        if epoch < n_shown:
            y_predict = X_new_b @ theta
            color = mpl.colors.rgb2hex(plt.cm.OrRd(epoch / n_shown + 0.15))
            plt.plot(X_new, y_predict, linestyle="solid", color=color)
            gradients = 2 / m * X_b.T @ (X_b @ theta - y)
            theta = theta - eta * gradients
            theta_path.append(theta)
    plt.xlabel("$x_1$")
    plt.axis([0, 2, 0, 15])
    plt.grid()
    plt.title(fr"$\eta = \{eta\}$")
    return theta_path

np.random.seed(42)
theta = np.random.randn(2, 1) # random initialization

plt.figure(figsize=(10, 4))
plt.subplot(131)
plot_gradient_descent(theta, eta=0.02)
plt.ylabel("$y$", rotation=0)
plt.subplot(132)
theta_path_bgd = plot_gradient_descent(theta, eta=0.1)
plt.gca().axes.yaxis.set_ticklabels([])
plt.subplot(133)
plt.gca().axes.yaxis.set_ticklabels([])
plot_gradient_descent(theta, eta=0.5)
save_fig("gradient_descent_plot")
plt.show()
```



Compare the learning rate provided and shown in the graph above. Provide a quick discussion on how learning rate affects model learning based on the given figure.

3.2.2 Stochastic Gradient Descent

The main problem with Batch Gradient Descent is the fact that it uses the whole training set to compute the gradients at every step, which makes it very slow when the training set is large. At the opposite extreme, Stochastic Gradient Descent just picks a random instance in the training set at every step and computes the gradients based only on that single instance. Obviously this makes the algorithm much faster since it has very little data to manipulate at every iteration. It also makes it possible to train on huge training sets, since only one instance needs to be in memory at each iteration (SGD can be implemented as an out-of-core algorithm.)

```
In [18]: theta_path_sgd = [] # extra code - we need to store the path of theta in the
#                               parameter space to plot the next figure
```

```
In [19]: n_epochs = 50
t0, t1 = 5, 50 # learning schedule hyperparameters

def learning_schedule(t):
    return t0 / (t + t1)

np.random.seed(42)
theta = np.random.randn(2, 1) # random initialization

n_shown = 20 # extra code - just needed to generate the figure below
plt.figure(figsize=(6, 4)) # extra code - not needed, just formatting

for epoch in range(n_epochs):
    for iteration in range(m):

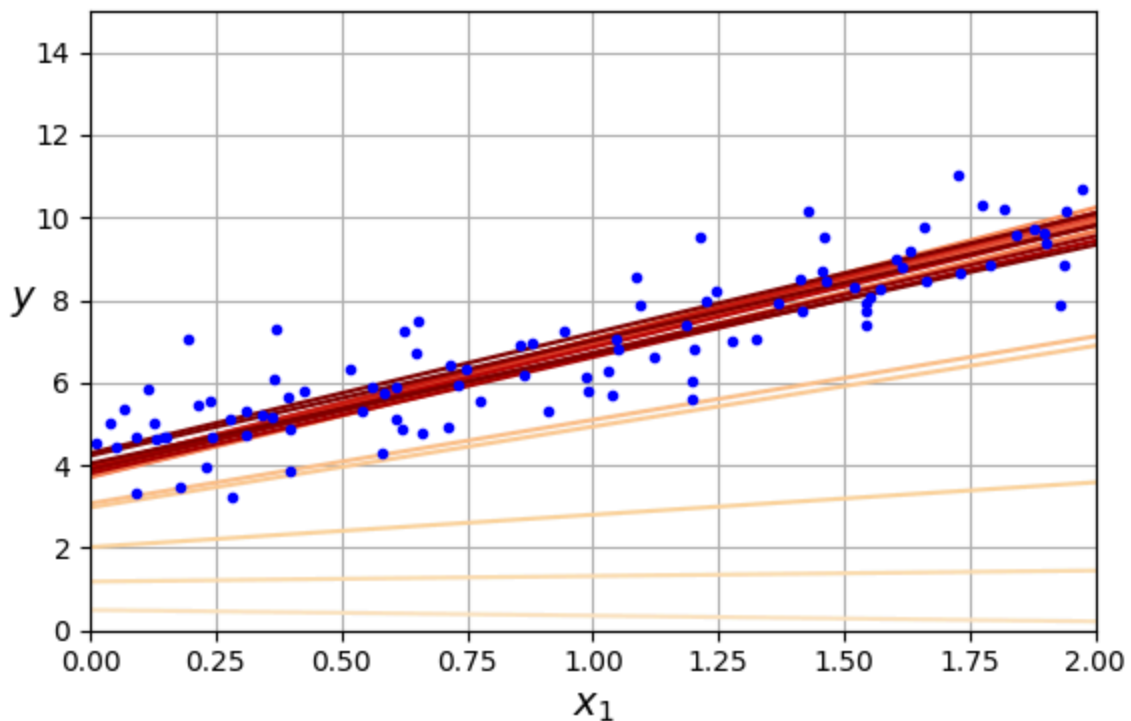
        # extra code - these 4 lines are used to generate the figure
        if epoch == 0 and iteration < n_shown:
            y_predict = X_new_b @ theta
            color = mpl.colors.rgb2hex(plt.cm.OrRd(iteration / n_shown + 0.15))
            plt.plot(X_new, y_predict, color=color)
```

```

random_index = np.random.randint(m)
xi = X_b[random_index : random_index + 1]
yi = y[random_index : random_index + 1]
gradients = 2 * xi.T @ (xi @ theta - yi) # for SGD, do not divide by m
eta = learning_schedule(epoch * m + iteration)
theta = theta - eta * gradients
theta_path_sgd.append(theta) # extra code - to generate the figure

# extra code - this section beautifies and saves Figure 4-10
plt.plot(X, y, "b.")
plt.xlabel("$x_1$")
plt.ylabel("$y$", rotation=0)
plt.axis([0, 2, 0, 15])
plt.grid()
save_fig("sgd_plot")
plt.show()

```



In [20]: theta

Out[20]: array([[4.21076011],
[2.74856079]])

In [21]: from sklearn.linear_model import SGDRegressor

```

sgd_reg = SGDRegressor(max_iter=1000, tol=1e-5, penalty=None, eta0=0.01,
                        n_iter_no_change=100, random_state=42)
sgd_reg.fit(X, y.ravel()) # y.ravel() because fit() expects 1D targets

```



```
Out[21]: SGDRegressor
SGDRegressor(n_iter_no_change=100, penalty=None, random_state=42, tol=1e-05)
```

```
In [22]: sgd_reg.intercept_, sgd_reg.coef_
```

```
Out[22]: (array([4.21278812]), array([2.77270267]))
```

Compare the Stochastic Gradient Descent used in this section to the gradient descent algorithm used previously. What are the benefits to using Stochastic Gradient Descent (SGD)?

- sgd is more faster (it is ideal on larger datasets), sgd are compatible with regularization techniques like L1 or L2 regularization so that we can prevent overfitting

3.2.3 Mini-batch gradient descent

The last Gradient Descent algorithm we will look at is called Mini-batch Gradient Descent. It is quite simple to understand once you know Batch and Stochastic Gradient Descent: at each step, instead of computing the gradients based on the full training set (as in Batch GD) or based on just one instance (as in Stochastic GD), Minibatch GD computes the gradients on small random sets of instances called minibatches. The main advantage of Mini-batch GD over Stochastic GD is that you can get a performance boost from hardware optimization of matrix operations, especially when using GPUs.

```
In [23]: # extra code - this cell generates and saves Figure 4-11

from math import ceil

n_epochs = 50
minibatch_size = 20
n_batches_per_epoch = ceil(m / minibatch_size)

np.random.seed(42)
theta = np.random.randn(2, 1) # random initialization

t0, t1 = 200, 1000 # Learning schedule hyperparameters

def learning_schedule(t):
    return t0 / (t + t1)

theta_path_mgd = []
for epoch in range(n_epochs):
    shuffled_indices = np.random.permutation(m)
    X_b_shuffled = X_b[shuffled_indices]
    y_shuffled = y[shuffled_indices]
    for iteration in range(0, n_batches_per_epoch):
        idx = iteration * minibatch_size
        xi = X_b_shuffled[idx : idx + minibatch_size]
```

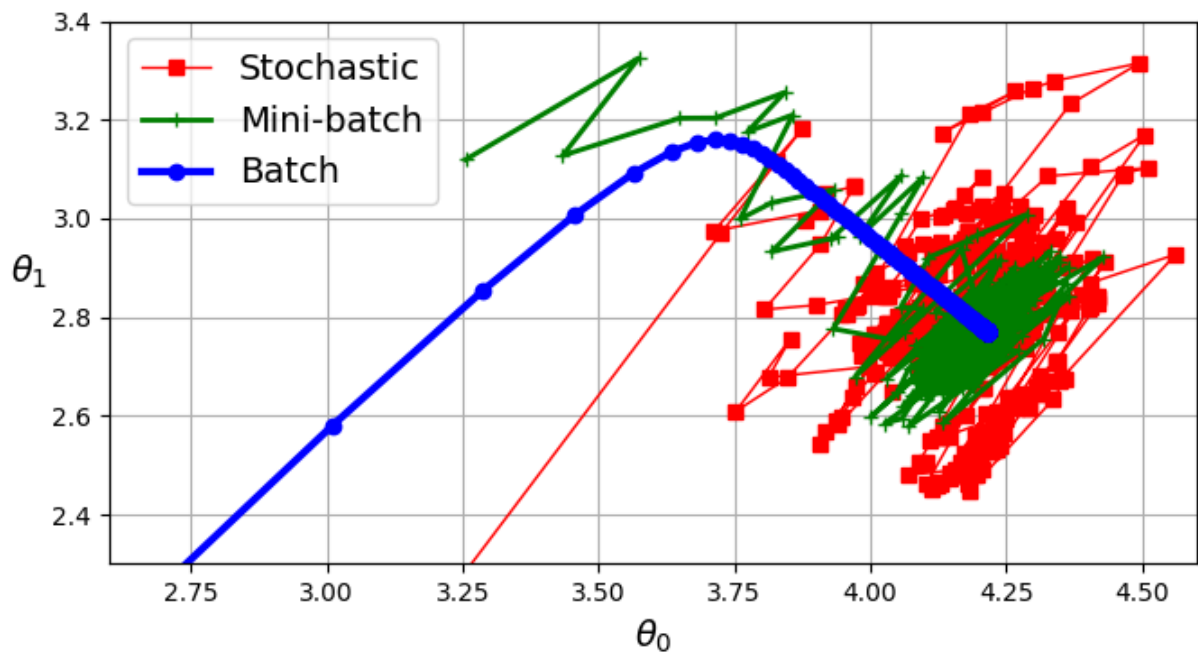
```

yi = y_shuffled[idx : idx + minibatch_size]
gradients = 2 / minibatch_size * xi.T @ (xi @ theta - yi)
eta = learning_schedule(iteration)
theta = theta - eta * gradients
theta_path_mgd.append(theta)

theta_path_bgd = np.array(theta_path_bgd)
theta_path_sgd = np.array(theta_path_sgd)
theta_path_mgd = np.array(theta_path_mgd)

plt.figure(figsize=(7, 4))
plt.plot(theta_path_sgd[:, 0], theta_path_sgd[:, 1], "r-s", linewidth=1,
         label="Stochastic")
plt.plot(theta_path_mgd[:, 0], theta_path_mgd[:, 1], "g-+", linewidth=2,
         label="Mini-batch")
plt.plot(theta_path_bgd[:, 0], theta_path_bgd[:, 1], "b-o", linewidth=3,
         label="Batch")
plt.legend(loc="upper left")
plt.xlabel(r"$\theta_0$")
plt.ylabel(r"$\theta_1$", rotation=0)
plt.axis([2.6, 4.6, 2.3, 3.4])
plt.grid()
save_fig("gradient_descent_paths_plot")
plt.show()

```



Provide a comparison of the 3 gradient descent algorithms included in this section. The comparison must provide an analysis based on different factors such as execution time (among others).

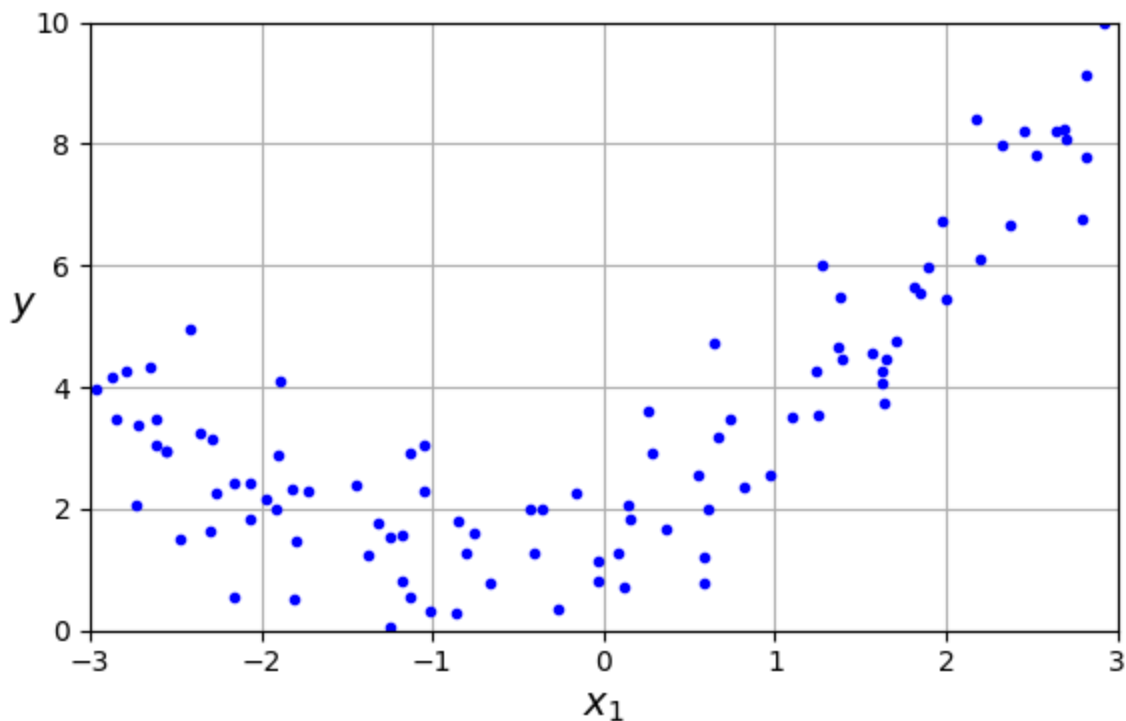
3.3 Polynomial Regression

What if your data is actually more complex than a simple straight line? Surprisingly, you can actually use a linear model to fit nonlinear data. A simple way to do this is to add powers of

each feature as new features, then train a linear model on this extended set of features. This technique is called Polynomial Regression.

```
In [24]: np.random.seed(42)
m = 100
X = 6 * np.random.rand(m, 1) - 3
y = 0.5 * X ** 2 + X + 2 + np.random.randn(m, 1)
```

```
In [25]: # extra code - this cell generates and saves Figure 4-12
plt.figure(figsize=(6, 4))
plt.plot(X, y, "b.")
plt.xlabel("$x_1$")
plt.ylabel("$y$", rotation=0)
plt.axis([-3, 3, 0, 10])
plt.grid()
save_fig("quadratic_data_plot")
plt.show()
```



```
In [26]: from sklearn.preprocessing import PolynomialFeatures

poly_features = PolynomialFeatures(degree=2, include_bias=False)
X_poly = poly_features.fit_transform(X)
X[0]
```

```
Out[26]: array([-0.75275929])
```

```
In [27]: X_poly[0]
```

```
Out[27]: array([-0.75275929,  0.56664654])
```

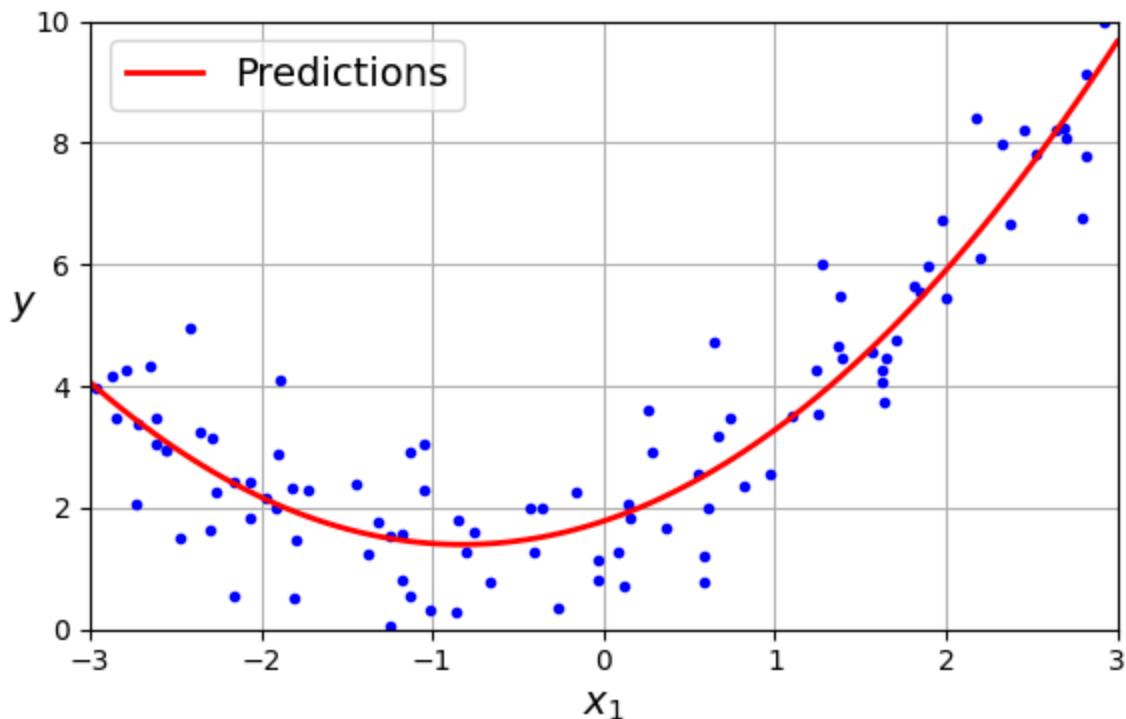
```
In [28]: lin_reg = LinearRegression()
lin_reg.fit(X_poly, y)
lin_reg.intercept_, lin_reg.coef_
```

```
Out[28]: (array([1.78134581]), array([[0.93366893, 0.56456263]]))
```

```
In [29]: # extra code - this cell generates and saves Figure 4-13
```

```
X_new = np.linspace(-3, 3, 100).reshape(100, 1)
X_new_poly = poly_features.transform(X_new)
y_new = lin_reg.predict(X_new_poly)

plt.figure(figsize=(6, 4))
plt.plot(X, y, "b.")
plt.plot(X_new, y_new, "r-", linewidth=2, label="Predictions")
plt.xlabel("$x_1$")
plt.ylabel("$y$", rotation=0)
plt.legend(loc="upper left")
plt.axis([-3, 3, 0, 10])
plt.grid()
save_fig("quadratic_predictions_plot")
plt.show()
```



```
In [30]: # extra code - this cell generates and saves Figure 4-14
```

```
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import make_pipeline

plt.figure(figsize=(6, 4))

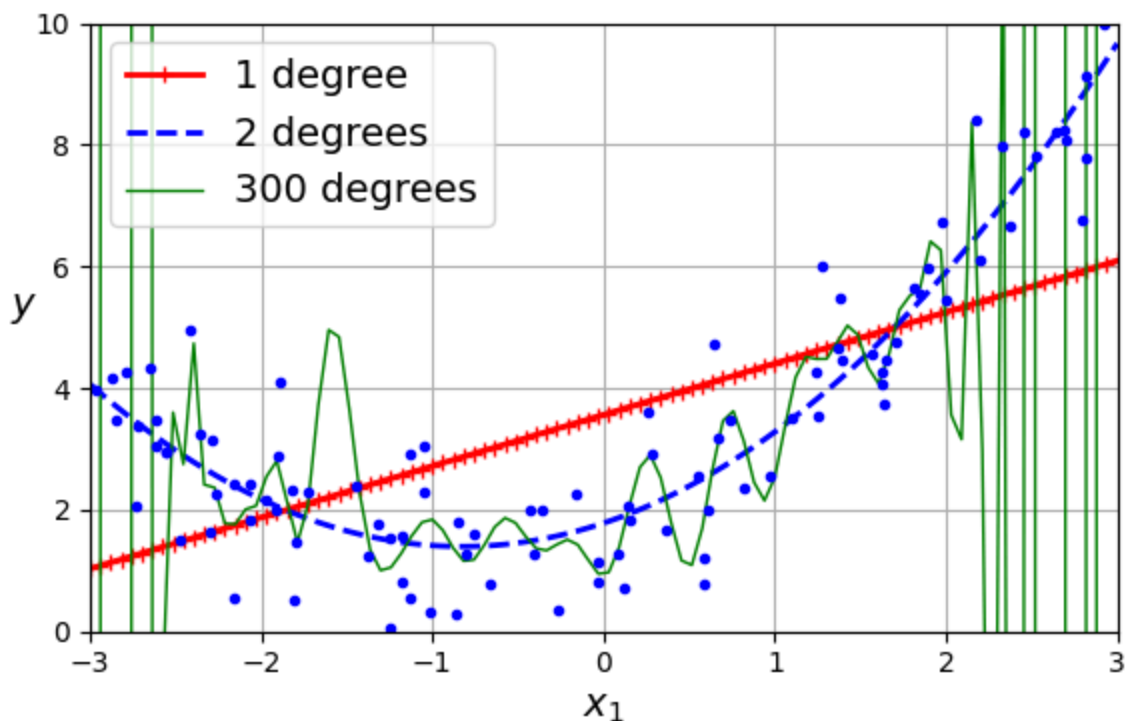
for style, width, degree in (("r-+", 2, 1), ("b--", 2, 2), ("g-", 1, 300)):
    polybig_features = PolynomialFeatures(degree=degree, include_bias=False)
```

```

std_scaler = StandardScaler()
lin_reg = LinearRegression()
polynomial_regression = make_pipeline(polybig_features, std_scaler, lin_reg)
polynomial_regression.fit(X, y)
y_newbig = polynomial_regression.predict(X_new)
label = f"{degree} degree{'s' if degree > 1 else ''}"
plt.plot(X_new, y_newbig, style, label=label, linewidth=width)

plt.plot(X, y, "b.", linewidth=3)
plt.legend(loc="upper left")
plt.xlabel("$x_1$")
plt.ylabel("$y$", rotation=0)
plt.axis([-3, 3, 0, 10])
plt.grid()
save_fig("high_degree_polynomials_plot")
plt.show()

```



Based on this example, how is Polynomial Regression able to find relationships between features? Is this something simple Linear Regression is unable to do?

- making Logistic regression models can capture more complex datasets compared to linear regression

What do we achieve in the figure above? Explain the result of the comparison.

3.4 Learning Curves

If you perform high-degree Polynomial Regression, you will likely fit the training data much better than with plain Linear Regression.

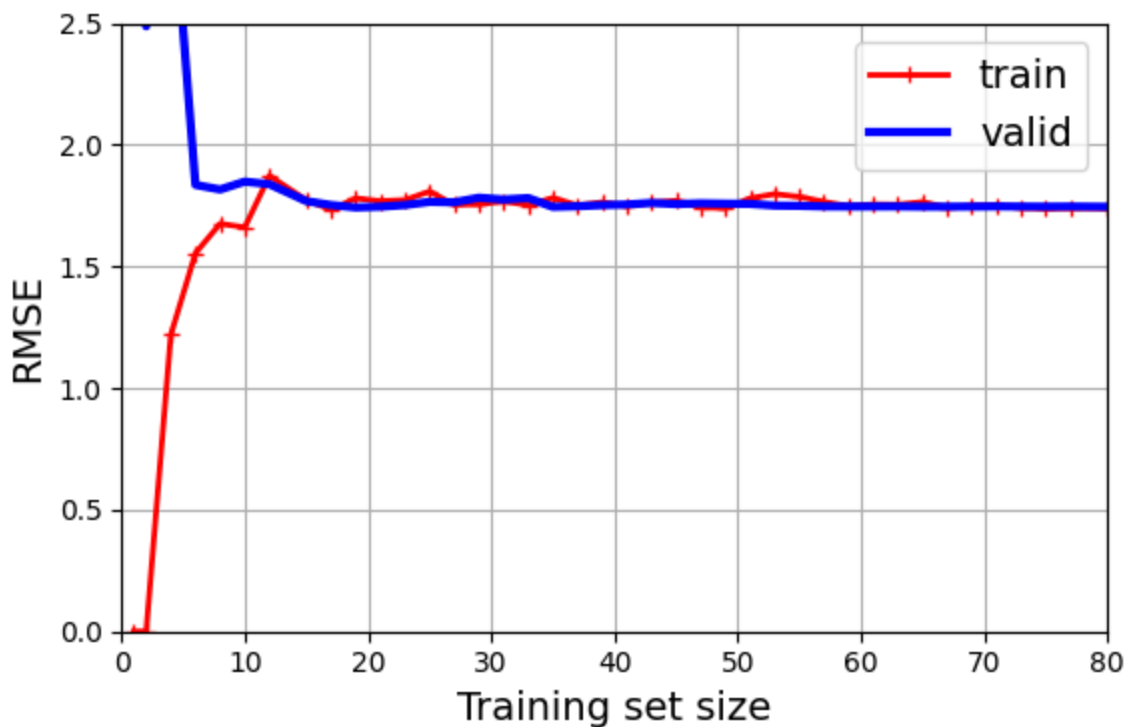
```
In [31]: from sklearn.model_selection import learning_curve

train_sizes, train_scores, valid_scores = learning_curve(
    LinearRegression(), X, y, train_sizes=np.linspace(0.01, 1.0, 40), cv=5,
    scoring="neg_root_mean_squared_error")
train_errors = -train_scores.mean(axis=1)
valid_errors = -valid_scores.mean(axis=1)

plt.figure(figsize=(6, 4)) # extra code - not needed, just formatting
plt.plot(train_sizes, train_errors, "r-+", linewidth=2, label="train")
plt.plot(train_sizes, valid_errors, "b-", linewidth=3, label="valid")

# extra code - beautifies and saves Figure 4-15
plt.xlabel("Training set size")
plt.ylabel("RMSE")
plt.grid()
plt.legend(loc="upper right")
plt.axis([0, 80, 0, 2.5])
save_fig("underfitting_learning_curves_plot")

plt.show()
```



This deserves a bit of explanation. First, let's look at the performance on the training data: when there are just one or two instances in the training set, the model can fit them perfectly, which is why the curve starts at zero. But as new instances are added to the training set, it becomes impossible for the model to fit the training data perfectly, both because the data is noisy and because it is not linear at all. So the error on the training data goes up until it reaches a plateau, at which point adding new instances to the training set doesn't make the average error much better or worse.

Now let's look at the performance of the model on the validation data. When the model is trained on very few training instances, it is incapable of generalizing properly, which is why the validation error is initially quite big. Then as the model is shown more training examples, it learns and thus the validation error slowly goes down. However, once again a straight line cannot do a good job modeling the data, so the error ends up at a plateau, very close to the other curve.

What are these learning curves indicating? Is the model overfitting or underfitting? Explain why.

```
In [32]: from sklearn.pipeline import make_pipeline

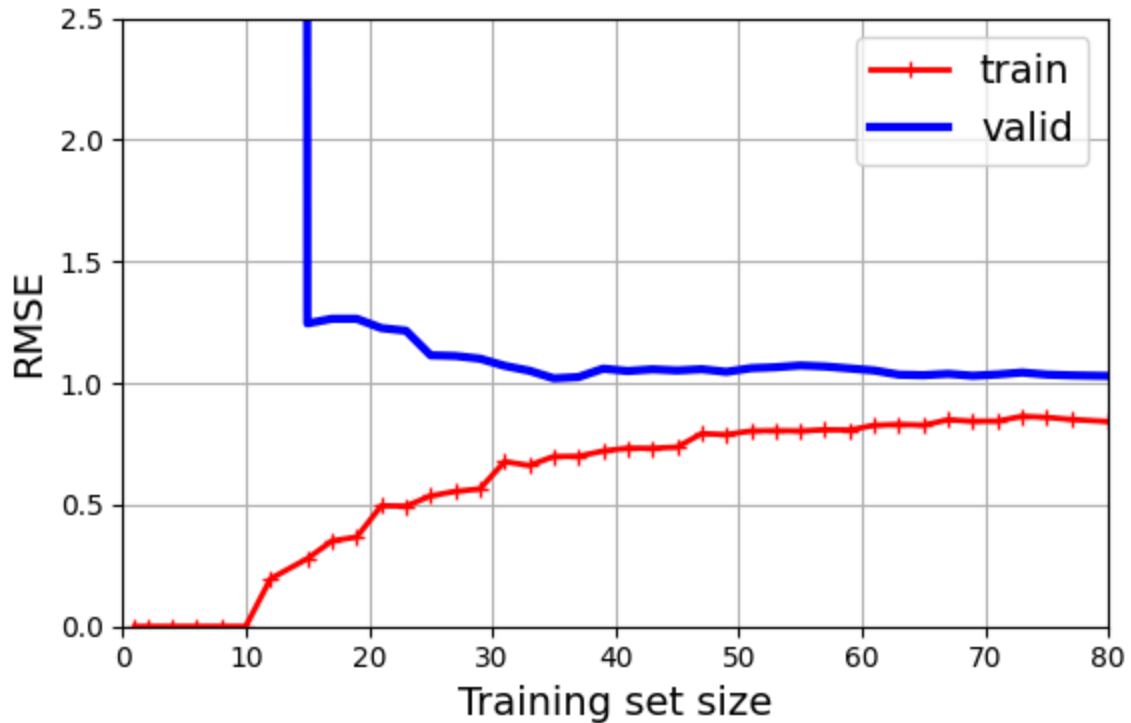
polynomial_regression = make_pipeline(
    PolynomialFeatures(degree=10, include_bias=False),
    LinearRegression())

train_sizes, train_scores, valid_scores = learning_curve(
    polynomial_regression, X, y, train_sizes=np.linspace(0.01, 1.0, 40), cv=5,
    scoring="neg_root_mean_squared_error")
```

```
In [33]: # extra code - generates and saves Figure 4-16

train_errors = -train_scores.mean(axis=1)
valid_errors = -valid_scores.mean(axis=1)

plt.figure(figsize=(6, 4))
plt.plot(train_sizes, train_errors, "r-+", linewidth=2, label="train")
plt.plot(train_sizes, valid_errors, "b-", linewidth=3, label="valid")
plt.legend(loc="upper right")
plt.xlabel("Training set size")
plt.ylabel("RMSE")
plt.grid()
plt.axis([0, 80, 0, 2.5])
save_fig("learning_curves_plot")
plt.show()
```



Compare this learning curve with the one previously shown above. The major differences you must explain here are as follows:

- The learning rate
- The gap between the curves

3.5 Regularized Linear Models

a good way to reduce overfitting is to regularize the model (i.e., to constrain it): the fewer degrees of freedom it has, the harder it will be for it to overfit the data. For example, a simple way to regularize a polynomial model is to reduce the number of polynomial degrees.

3.5.1 Ridge Regression

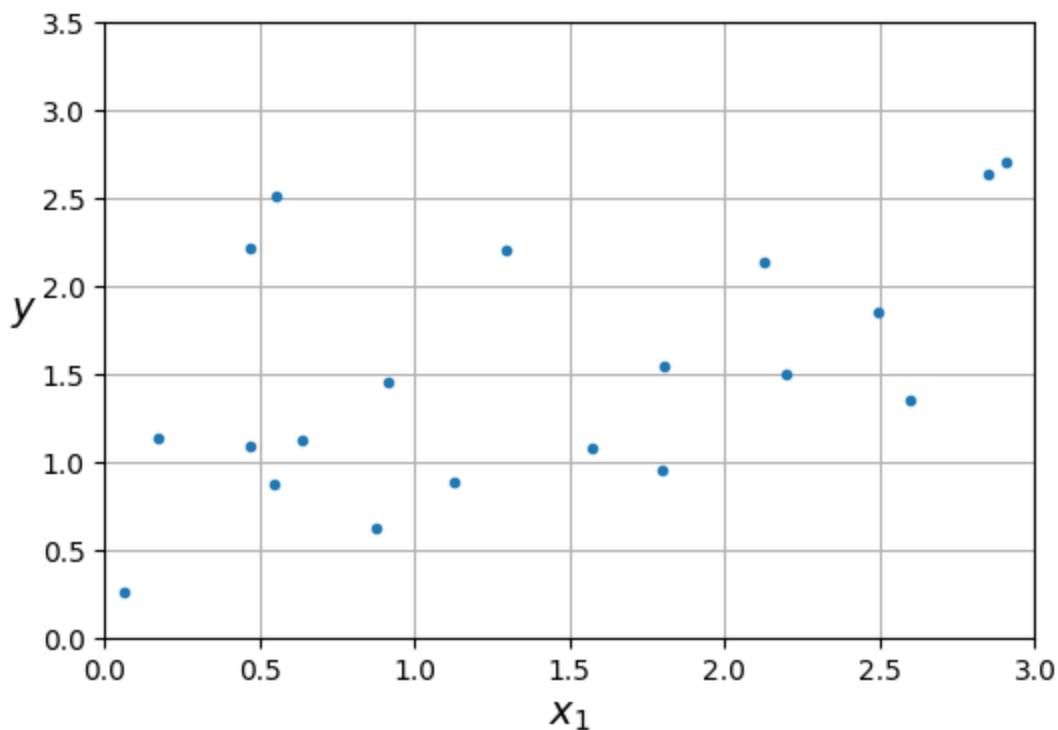
For a linear model, regularization is typically achieved by constraining the weights of the model. We will now look at Ridge Regression, Lasso Regression, and Elastic Net, which implement three different ways to constrain the weights.

This forces the learning algorithm to not only fit the data but also keep the model weights as small as possible. Note that the regularization term should only be added to the cost function during training. Once the model is trained, you want to evaluate the model's performance using the unregularized performance measure.

Let's generate a very small and noisy linear dataset:


```
In [34]: # extra code - we've done this type of generation several times before
np.random.seed(42)
m = 20
X = 3 * np.random.rand(m, 1)
y = 1 + 0.5 * X + np.random.randn(m, 1) / 1.5
X_new = np.linspace(0, 3, 100).reshape(100, 1)
```

```
In [35]: # extra code - a quick peek at the dataset we just generated
plt.figure(figsize=(6, 4))
plt.plot(X, y, ".")
plt.xlabel("$x_1$")
plt.ylabel("$y$", rotation=0)
plt.axis([0, 3, 0, 3.5])
plt.grid()
plt.show()
```



```
In [36]: from sklearn.linear_model import Ridge

ridge_reg = Ridge(alpha=0.1, solver="cholesky")
ridge_reg.fit(X, y)
ridge_reg.predict([[1.5]])
```

```
Out[36]: array([[1.55325833]])
```

As with Linear Regression, we can perform Ridge Regression either by computing a closed-form equation or by performing Gradient Descent. The pros and cons are the same.

```
In [37]: # extra code - this cell generates and saves Figure 4-17

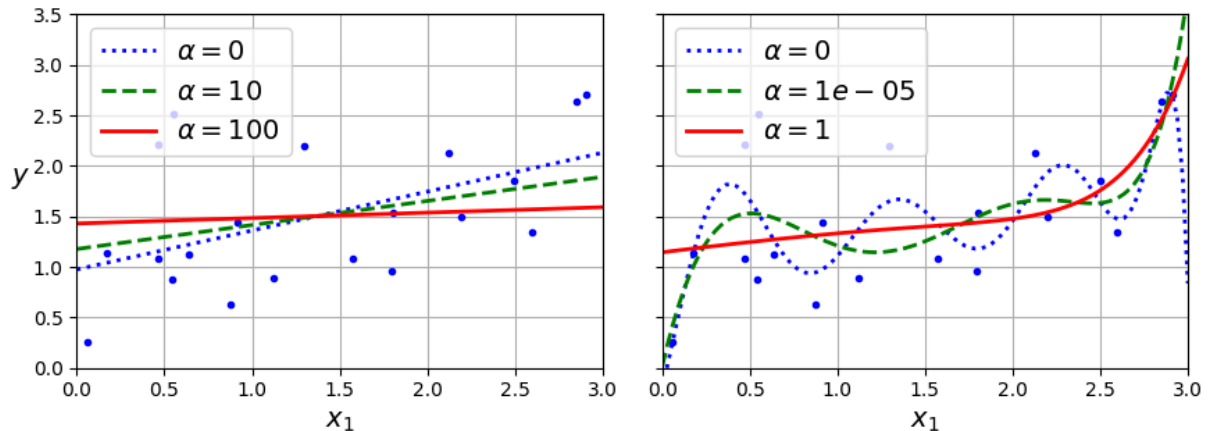
def plot_model(model_class, polynomial, alphas, **model_kwargs):
    plt.plot(X, y, "b.", linewidth=3)
```

```

for alpha, style in zip(alphas, ("b:", "g--", "r-")):
    if alpha > 0:
        model = model_class(alpha, **model_kwargs)
    else:
        model = LinearRegression()
    if polynomial:
        model = make_pipeline(
            PolynomialFeatures(degree=10, include_bias=False),
            StandardScaler(),
            model)
    model.fit(X, y)
    y_new_regul = model.predict(X_new)
    plt.plot(X_new, y_new_regul, style, linewidth=2,
             label=fr"$\alpha = {alpha}$")
plt.legend(loc="upper left")
plt.xlabel("$x_1$")
plt.axis([0, 3, 0, 3.5])
plt.grid()

plt.figure(figsize=(9, 3.5))
plt.subplot(121)
plot_model(Ridge, polynomial=False, alphas=(0, 10, 100), random_state=42)
plt.ylabel("$y$", rotation=0)
plt.subplot(122)
plot_model(Ridge, polynomial=True, alphas=(0, 10**-5, 1), random_state=42)
plt.gca().axes.yaxis.set_ticklabels([])
save_fig("ridge_regression_plot")
plt.show()

```



```

In [38]: sgd_reg = SGDRegressor(penalty="l2", alpha=0.1 / m, tol=None,
                                max_iter=1000, eta0=0.01, random_state=42)
sgd_reg.fit(X, y.ravel()) # y.ravel() because fit() expects 1D targets
sgd_reg.predict([[1.5]])

```

```

Out[38]: array([1.55302613])

```

```

In [39]: # extra code - show that we get roughly the same solution as earlier when
#         we use Stochastic Average GD (solver="sag")
ridge_reg = Ridge(alpha=0.1, solver="sag", random_state=42)
ridge_reg.fit(X, y)
ridge_reg.predict([[1.5]])

```

Out[39]: array([[1.55326019]])

```
In [40]: # extra code - shows the closed form solution of Ridge regression,
#         compare with the next Ridge model's Learned parameters below
alpha = 0.1
A = np.array([[0., 0.], [0., 1.]])
X_b = np.c_[np.ones(m), X]
np.linalg.inv(X_b.T @ X_b + alpha * A) @ X_b.T @ y
```

Out[40]: array([[0.97898394],
[0.3828496]])

```
In [41]: ridge_reg.intercept_, ridge_reg.coef_ # extra code
```

Out[41]: (array([0.97896386]), array([[0.38286422]]))

Discuss the following:

- What is the purpose of the hyperparameter alpha as seen in the figures above?
- Suppose you are using Ridge Regression and you notice that the training error and the validation error are almost equal and fairly high. Would you say that the model suffers from high bias or high variance? Should you increase the regularization hyperparameter α or reduce it?

3.5.2 Lasso Regression

Least Absolute Shrinkage and Selection Operator Regression (simply called Lasso Regression) is another regularized version of Linear Regression: just like Ridge Regression, it adds a regularization term to the cost function, but it uses the ℓ_1 norm of the weight vector instead of half the square of the ℓ_2 norm.

```
In [42]: from sklearn.linear_model import Lasso

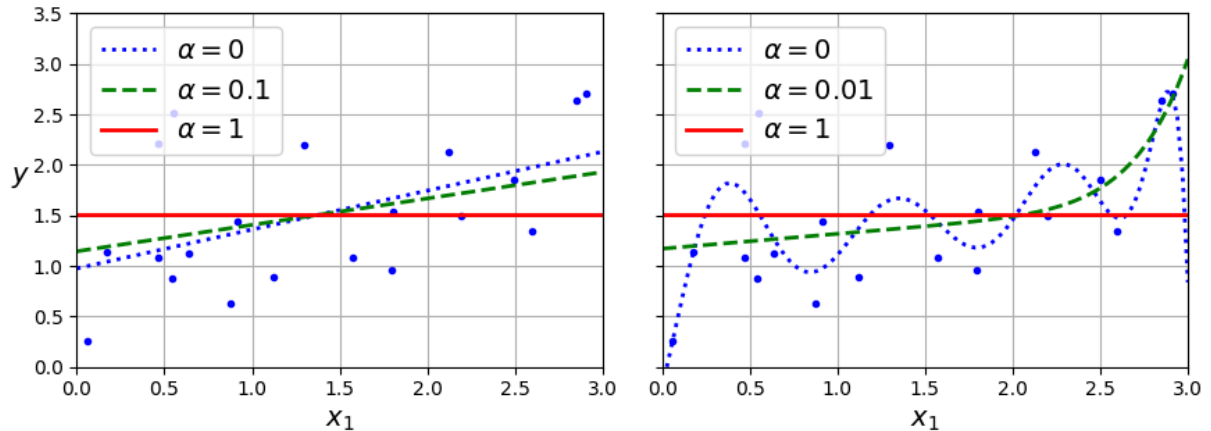
lasso_reg = Lasso(alpha=0.1)
lasso_reg.fit(X, y)
lasso_reg.predict([[1.5]])
```

Out[42]: array([1.53788174])

An important characteristic of Lasso Regression is that it tends to completely eliminate the weights of the least important features (i.e., set them to zero).

```
In [43]: # extra code - this cell generates and saves Figure 4-18
plt.figure(figsize=(9, 3.5))
plt.subplot(121)
plot_model(Lasso, polynomial=False, alphas=(0, 0.1, 1), random_state=42)
plt.ylabel("$y$", rotation=0)
plt.subplot(122)
plot_model(Lasso, polynomial=True, alphas=(0, 1e-2, 1), random_state=42)
plt.gca().axes.yaxis.set_ticklabels([])
```

```
save_fig("lasso_regression_plot")
plt.show()
```



For example, the dashed line in the right plot on Figure 4-18 (with $\alpha = 10^{-7}$) looks quadratic, almost linear: all the weights for the high-degree polynomial features are equal to zero. In other words, Lasso Regression automatically performs feature selection and outputs a sparse model (i.e., with few nonzero feature weights).

```
In [44]: # extra code - this BIG cell generates and saves Figure 4-19

t1a, t1b, t2a, t2b = -1, 3, -1.5, 1.5

t1s = np.linspace(t1a, t1b, 500)
t2s = np.linspace(t2a, t2b, 500)
t1, t2 = np.meshgrid(t1s, t2s)
T = np.c_[t1.ravel(), t2.ravel()]
Xr = np.array([[1, 1], [1, -1], [1, 0.5]])
yr = 2 * Xr[:, :1] + 0.5 * Xr[:, 1:]

J = (1 / len(Xr) * ((T @ Xr.T - yr.T) ** 2).sum(axis=1)).reshape(t1.shape)

N1 = np.linalg.norm(T, ord=1, axis=1).reshape(t1.shape)
N2 = np.linalg.norm(T, ord=2, axis=1).reshape(t1.shape)

t_min_idx = np.unravel_index(J.argmin(), J.shape)
t1_min, t2_min = t1[t_min_idx], t2[t_min_idx]

t_init = np.array([[0.25], [-1]])

def bgd_path(theta, X, y, l1, l2, core=1, eta=0.05, n_iterations=200):
    path = [theta]
    for iteration in range(n_iterations):
        gradients = (core * 2 / len(X) * X.T @ (X @ theta - y)
                     + l1 * np.sign(theta) + l2 * theta)
        theta = theta - eta * gradients
        path.append(theta)
    return np.array(path)

fig, axes = plt.subplots(2, 2, sharex=True, sharey=True, figsize=(10.1, 8))
```

```

for i, N, l1, l2, title in ((0, N1, 2.0, 0, "Lasso"), (1, N2, 0, 2.0, "Ridge")):
    JR = J + l1 * N1 + l2 * 0.5 * N2 ** 2

    tr_min_idx = np.unravel_index(JR.argmin(), JR.shape)
    t1r_min, t2r_min = t1[tr_min_idx], t2[tr_min_idx]

    levels = np.exp(np.linspace(0, 1, 20)) - 1
    levelsJ = levels * (J.max() - J.min()) + J.min()
    levelsJR = levels * (JR.max() - JR.min()) + JR.min()
    levelsN = np.linspace(0, N.max(), 10)

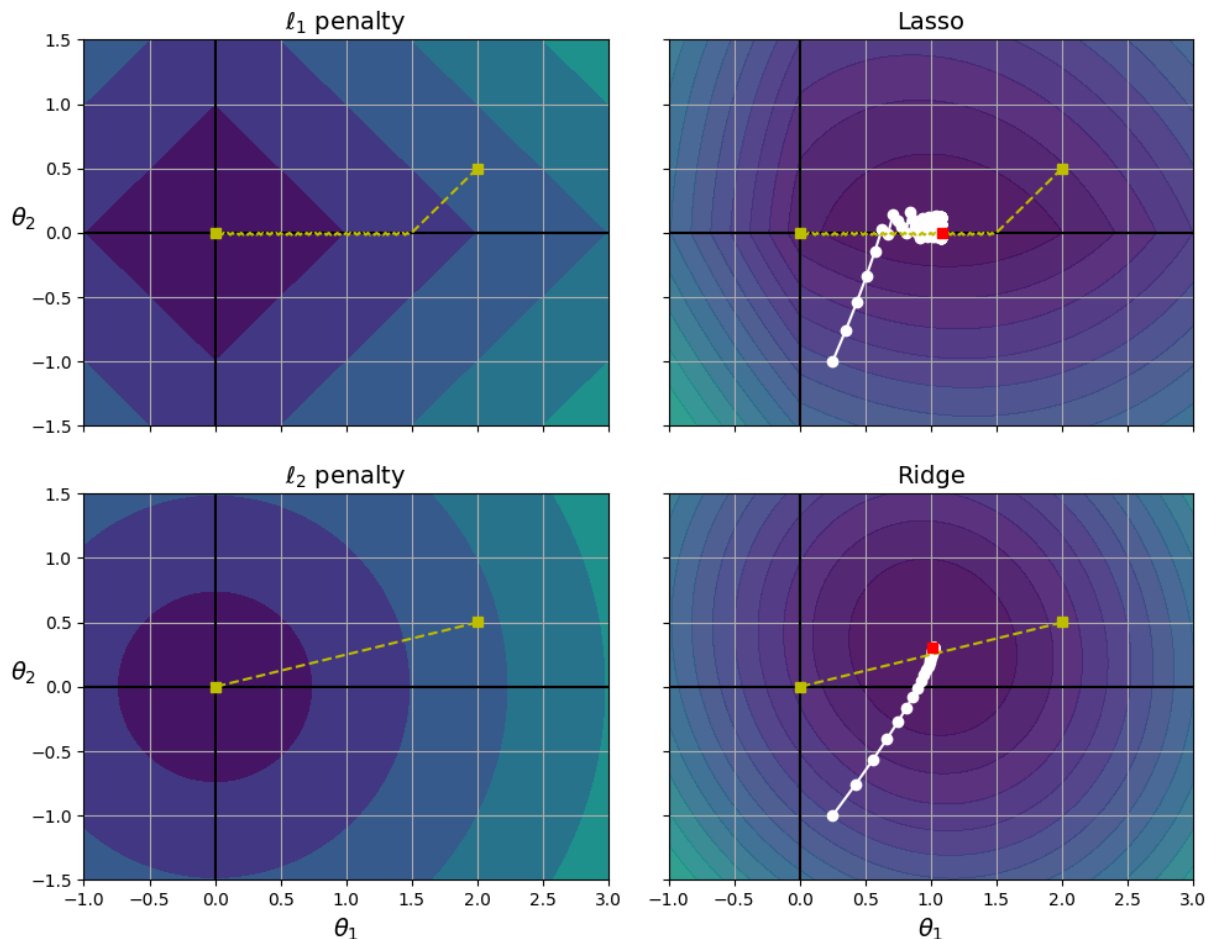
    path_J = bgd_path(t_init, Xr, yr, l1=0, l2=0)
    path_JR = bgd_path(t_init, Xr, yr, l1, l2)
    path_N = bgd_path(theta=np.array([[2.0], [0.5]]), X=Xr, y=yr,
                        l1=np.sign(l1) / 3, l2=np.sign(l2), core=0)

    ax = axes[i, 0]
    ax.grid()
    ax.axhline(y=0, color="k")
    ax.axvline(x=0, color="k")
    ax.contourf(t1, t2, N / 2.0, levels=levelsN)
    ax.plot(path_N[:, 0], path_N[:, 1], "y--")
    ax.plot(0, 0, "ys")
    ax.plot(t1_min, t2_min, "ys")
    ax.set_title(fr"$\ell_{i + 1}$ penalty")
    ax.axis([t1a, t1b, t2a, t2b])
    if i == 1:
        ax.set_xlabel(r"$\theta_1$")
    ax.set_ylabel(r"$\theta_2$", rotation=0)

    ax = axes[i, 1]
    ax.grid()
    ax.axhline(y=0, color="k")
    ax.axvline(x=0, color="k")
    ax.contourf(t1, t2, JR, levels=levelsJR, alpha=0.9)
    ax.plot(path_JR[:, 0], path_JR[:, 1], "w-o")
    ax.plot(path_N[:, 0], path_N[:, 1], "y--")
    ax.plot(0, 0, "ys")
    ax.plot(t1_min, t2_min, "ys")
    ax.plot(t1r_min, t2r_min, "rs")
    ax.set_title(title)
    ax.axis([t1a, t1b, t2a, t2b])
    if i == 1:
        ax.set_xlabel(r"$\theta_1$")

save_fig("lasso_vs_ridge_plot")
plt.show()

```



Using the figure above, compare the Lasso Regression and the Ridge Regression. What kind of performance is observable from the graph?

3.6 Early Stopping

A very different way to regularize iterative learning algorithms such as Gradient Descent is to stop training as soon as the validation error reaches a minimum. This is called early stopping.

Let's go back to the quadratic dataset we used earlier:

```
In [45]: from copy import deepcopy
from sklearn.metrics import mean_squared_error
from sklearn.preprocessing import StandardScaler

# extra code - creates the same quadratic dataset as earlier and splits it
np.random.seed(42)
m = 100
X = 6 * np.random.rand(m, 1) - 3
y = 0.5 * X ** 2 + X + 2 + np.random.randn(m, 1)
X_train, y_train = X[: m // 2], y[: m // 2, 0]
X_valid, y_valid = X[m // 2 :, 0], y[m // 2 :, 0]

preprocessing = make_pipeline(PolynomialFeatures(degree=90, include_bias=False),
                              StandardScaler())
```

```

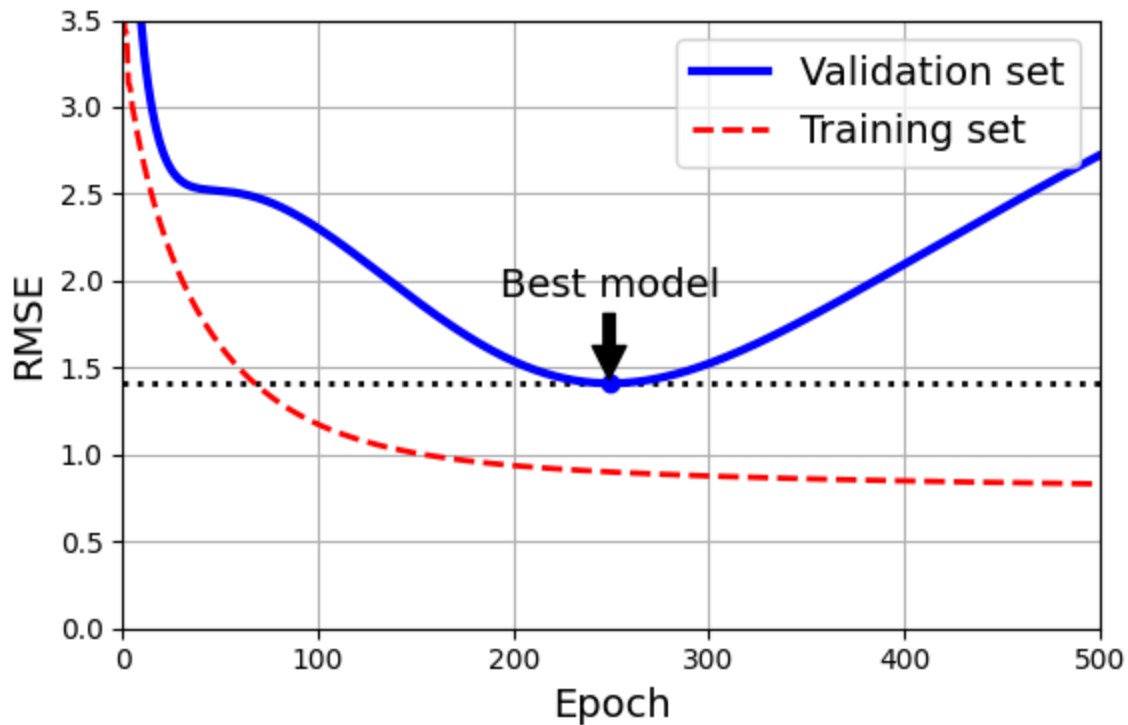
X_train_prep = preprocessing.fit_transform(X_train)
X_valid_prep = preprocessing.transform(X_valid)
sgd_reg = SGDRegressor(penalty=None, eta0=0.002, random_state=42)
n_epochs = 500
best_valid_rmse = float('inf')
train_errors, val_errors = [], [] # extra code - it's for the figure below

for epoch in range(n_epochs):
    sgd_reg.partial_fit(X_train_prep, y_train)
    y_valid_predict = sgd_reg.predict(X_valid_prep)
    val_error = mean_squared_error(y_valid, y_valid_predict, squared=False)
    if val_error < best_valid_rmse:
        best_valid_rmse = val_error
        best_model = deepcopy(sgd_reg)

    # extra code - we evaluate the train error and save it for the figure
    y_train_predict = sgd_reg.predict(X_train_prep)
    train_error = mean_squared_error(y_train, y_train_predict, squared=False)
    val_errors.append(val_error)
    train_errors.append(train_error)

# extra code - this section generates and saves Figure 4-20
best_epoch = np.argmin(val_errors)
plt.figure(figsize=(6, 4))
plt.annotate('Best model',
            xy=(best_epoch, best_valid_rmse),
            xytext=(best_epoch, best_valid_rmse + 0.5),
            ha="center",
            arrowprops=dict(facecolor='black', shrink=0.05))
plt.plot([0, n_epochs], [best_valid_rmse, best_valid_rmse], "k:", linewidth=2)
plt.plot(val_errors, "b-", linewidth=3, label="Validation set")
plt.plot(best_epoch, best_valid_rmse, "bo")
plt.plot(train_errors, "r--", linewidth=2, label="Training set")
plt.legend(loc="upper right")
plt.xlabel("Epoch")
plt.ylabel("RMSE")
plt.axis([0, n_epochs, 0, 3.5])
plt.grid()
save_fig("early_stopping_plot")
plt.show()

```



Do you foresee any useful benefits of implementing early stopping?

3.7 Logistic Regression

Logistic Regression (also called Logit Regression) is commonly used to estimate the probability that an instance belongs to a particular class (e.g., what is the probability that this email is spam?). If the estimated probability is greater than 50%, then the model predicts that the instance belongs to that class (called the positive class, labeled "1"), or else it predicts that it does not (i.e., it belongs to the negative class, labeled "0"). This makes it a binary classifier.

3.7.1 Estimating Probabilities

So how does it work? Just like a Linear Regression model, a Logistic Regression model computes a weighted sum of the input features (plus a bias term), but instead of outputting the result directly like the Linear Regression model does, it outputs the logistic of this result.

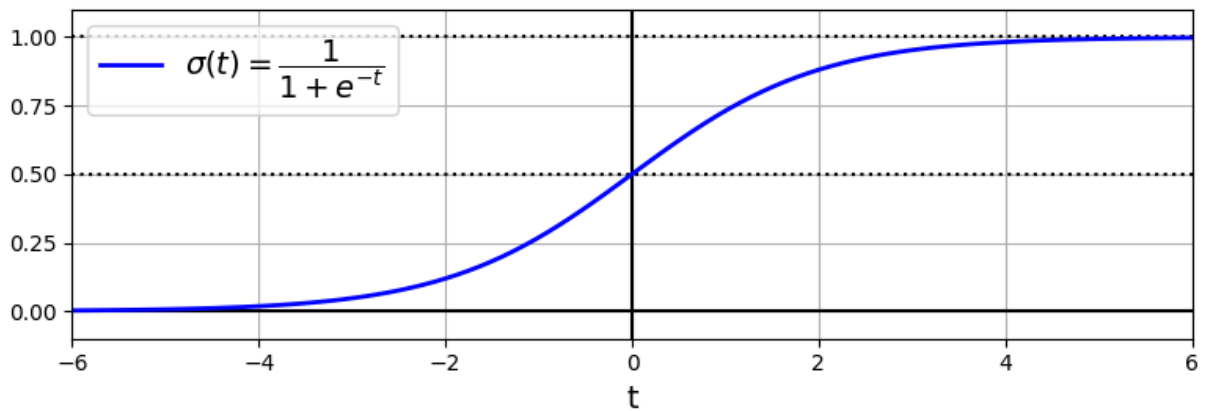
In [46]: *# extra code - generates and saves Figure 4-21*

```
lim = 6
t = np.linspace(-lim, lim, 100)
sig = 1 / (1 + np.exp(-t))

plt.figure(figsize=(8, 3))
plt.plot([-lim, lim], [0, 0], "k-")
plt.plot([-lim, lim], [0.5, 0.5], "k:")
plt.plot([-lim, lim], [1, 1], "k:")
plt.plot([0, 0], [-1.1, 1.1], "k-")
```



```
plt.plot(t, sig, "b-", linewidth=2, label=r"$\sigma(t) = \frac{1}{1 + e^{-t}}$")
plt.xlabel("t")
plt.legend(loc="upper left")
plt.axis([-lim, lim, -0.1, 1.1])
plt.gca().set_yticks([0, 0.25, 0.5, 0.75, 1])
plt.grid()
save_fig("logistic_function_plot")
plt.show()
```



Given the figure above, explain how the equation might be interpreted as the figure and how this provides a logistic regression prediction?

3.7.2 Decision Boundaries

Let's use the iris dataset to illustrate Logistic Regression. This is a famous dataset that contains the sepal and petal length and width of 150 iris flowers of three different species: Iris-Setosa, Iris-Versicolor, and Iris-Virginica

```
In [47]: from sklearn.datasets import load_iris

iris = load_iris(as_frame=True)
list(iris)
```

```
Out[47]: ['data',
          'target',
          'frame',
          'target_names',
          'DESCR',
          'feature_names',
          'filename',
          'data_module']
```

```
In [48]: print(iris.DESCR) # extra code - it's a bit too long
```

```
.. _iris_dataset:
```

```
Iris plants dataset
```

```
-----
```

```
**Data Set Characteristics:**
```

```
:Number of Instances: 150 (50 in each of three classes)
:Number of Attributes: 4 numeric, predictive attributes and the class
:Attribute Information:
  - sepal length in cm
  - sepal width in cm
  - petal length in cm
  - petal width in cm
  - class:
    - Iris-Setosa
    - Iris-Versicolour
    - Iris-Virginica
```

```
:Summary Statistics:
```

```
=====
      Min  Max   Mean   SD   Class Correlation
=====
sepal length:  4.3  7.9   5.84   0.83    0.7826
sepal width:   2.0  4.4   3.05   0.43   -0.4194
petal length:  1.0  6.9   3.76   1.76    0.9490 (high!)
petal width:   0.1  2.5   1.20   0.76    0.9565 (high!)
=====
```

```
:Missing Attribute Values: None
:Class Distribution: 33.3% for each of 3 classes.
:Creator: R.A. Fisher
:Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
:Date: July, 1988
```

The famous Iris database, first used by Sir R.A. Fisher. The dataset is taken from Fisher's paper. Note that it's the same as in R, but not as in the UCI Machine Learning Repository, which has two wrong data points.

This is perhaps the best known database to be found in the pattern recognition literature. Fisher's paper is a classic in the field and is referenced frequently to this day. (See Duda & Hart, for example.) The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.

```
|details-start|
```

```
**References**
```

```
|details-split|
```

- Fisher, R.A. "The use of multiple measurements in taxonomic problems" Annual Eugenics, 7, Part II, 179-188 (1936); also in "Contributions to Mathematical Statistics" (John Wiley, NY, 1950).
- Duda, R.O., & Hart, P.E. (1973) Pattern Classification and Scene Analysis. (Q327.D83) John Wiley & Sons. ISBN 0-471-22361-1. See page 218.

- Dasarathy, B.V. (1980) "Nosing Around the Neighborhood: A New System Structure and Classification Rule for Recognition in Partially Exposed Environments". IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-2, No. 1, 67-71.
- Gates, G.W. (1972) "The Reduced Nearest Neighbor Rule". IEEE Transactions on Information Theory, May 1972, 431-433.
- See also: 1988 MLC Proceedings, 54-64. Cheeseman et al's AUTOCLASS II conceptual clustering system finds 3 classes in the data.
- Many, many more ...

|details-end|

In [49]: `iris.data.head(3)`

Out[49]:

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2

In [50]: `iris.target.head(3)` *# note that the instances are not shuffled*

Out[50]:

	target
0	0
1	0
2	0

dtype: int64

In [51]: `iris.target_names`

Out[51]: `array(['setosa', 'versicolor', 'virginica'], dtype='<U10')`

In [52]:

```

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

X = iris.data[["petal width (cm)"]].values
y = iris.target_names[iris.target] == 'virginica'
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=42)

log_reg = LogisticRegression(random_state=42)
log_reg.fit(X_train, y_train)

```

Out[52]:

▼ LogisticRegression

LogisticRegression(random_state=42)

```

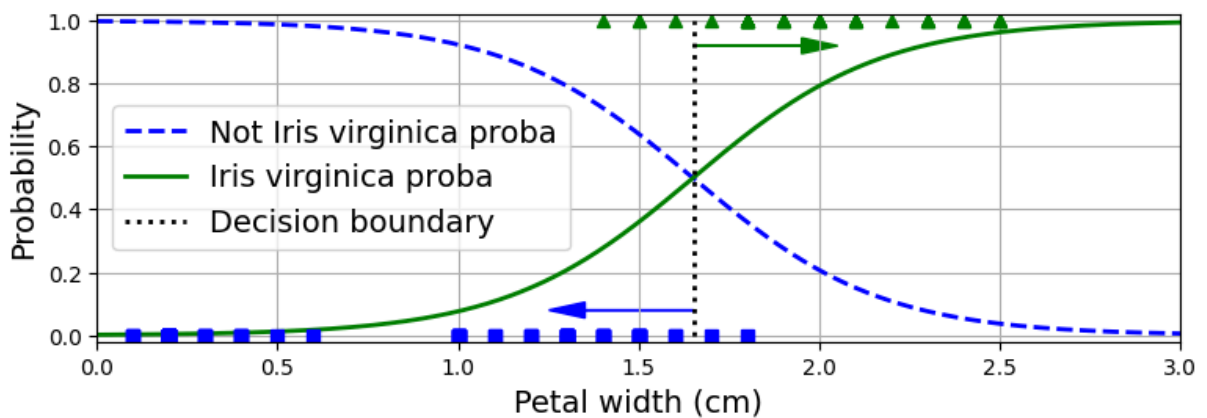
In [53]: X_new = np.linspace(0, 3, 1000).reshape(-1, 1) # reshape to get a column vector
y_proba = log_reg.predict_proba(X_new)
decision_boundary = X_new[y_proba[:, 1] >= 0.5][0, 0]

plt.figure(figsize=(8, 3)) # extra code - not needed, just formatting
plt.plot(X_new, y_proba[:, 0], "b--", linewidth=2,
         label="Not Iris virginica proba")
plt.plot(X_new, y_proba[:, 1], "g-", linewidth=2, label="Iris virginica proba")
plt.plot([decision_boundary, decision_boundary], [0, 1], "k:", linewidth=2,
         label="Decision boundary")

# extra code - this section beautifies and saves Figure 4-23
plt.arrow(x=decision_boundary, y=0.08, dx=-0.3, dy=0,
          head_width=0.05, head_length=0.1, fc="b", ec="b")
plt.arrow(x=decision_boundary, y=0.92, dx=0.3, dy=0,
          head_width=0.05, head_length=0.1, fc="g", ec="g")
plt.plot(X_train[y_train == 0], y_train[y_train == 0], "bs")
plt.plot(X_train[y_train == 1], y_train[y_train == 1], "g^")
plt.xlabel("Petal width (cm)")
plt.ylabel("Probability")
plt.legend(loc="center left")
plt.axis([0, 3, -0.02, 1.02])
plt.grid()
save_fig("logistic_regression_plot")

plt.show()

```



```
In [54]: decision_boundary
```

```
Out[54]: 1.6516516516516517
```

```
In [55]: log_reg.predict([[1.7], [1.5]])
```

```
Out[55]: array([ True, False])
```

```
In [56]: # extra code - this cell generates and saves Figure 4-24
```

```

X = iris.data[["petal length (cm)", "petal width (cm)"].values
y = iris.target_names[iris.target] == 'virginica'
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=42)

```

```

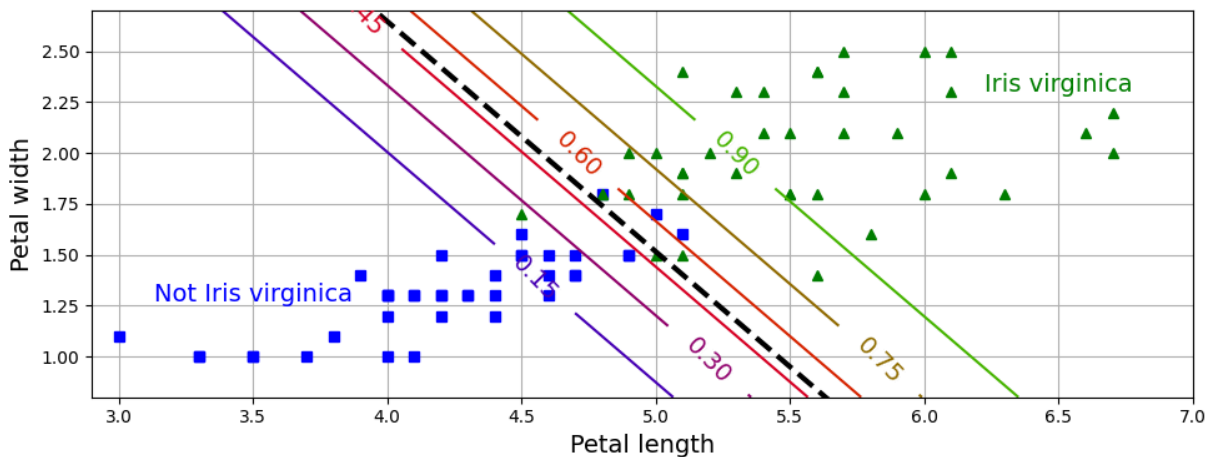
log_reg = LogisticRegression(C=2, random_state=42)
log_reg.fit(X_train, y_train)

# for the contour plot
x0, x1 = np.meshgrid(np.linspace(2.9, 7, 500).reshape(-1, 1),
                     np.linspace(0.8, 2.7, 200).reshape(-1, 1))
X_new = np.c_[x0.ravel(), x1.ravel()] # one instance per point on the figure
y_proba = log_reg.predict_proba(X_new)
zz = y_proba[:, 1].reshape(x0.shape)

# for the decision boundary
left_right = np.array([2.9, 7])
boundary = -((log_reg.coef_[0, 0] * left_right + log_reg.intercept_[0])
             / log_reg.coef_[0, 1])

plt.figure(figsize=(10, 4))
plt.plot(X_train[y_train == 0, 0], X_train[y_train == 0, 1], "bs")
plt.plot(X_train[y_train == 1, 0], X_train[y_train == 1, 1], "g^")
contour = plt.contour(x0, x1, zz, cmap=plt.cm.brg)
plt.clabel(contour, inline=1)
plt.plot(left_right, boundary, "k--", linewidth=3)
plt.text(3.5, 1.27, "Not Iris virginica", color="b", ha="center")
plt.text(6.5, 2.3, "Iris virginica", color="g", ha="center")
plt.xlabel("Petal length")
plt.ylabel("Petal width")
plt.axis([2.9, 7, 0.8, 2.7])
plt.grid()
save_fig("logistic_regression_contour_plot")
plt.show()

```



Given the above code and the figure provided, how do l1 and l2 penalties apply to the Logistic regression model?

3.7.3 Softmax Regression

The Logistic Regression model can be generalized to support multiple classes directly, without having to train and combine multiple binary classifiers. This is called Softmax Regression, or Multinomial Logistic Regression.

The idea is quite simple: when given an instance x , the Softmax Regression model first computes a score $s_k(x)$ for each class k , then estimates the probability of each class by applying the softmax function (also called the normalized exponential) to the scores. The equation to compute $s_k(x)$ should look familiar, as it is just like the equation for Linear Regression prediction

```
In [57]: X = iris.data[["petal length (cm)", "petal width (cm)"]].values
y = iris["target"]
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=42)

softmax_reg = LogisticRegression(C=30, random_state=42)
softmax_reg.fit(X_train, y_train)
```

```
Out[57]: LogisticRegression
LogisticRegression(C=30, random_state=42)
```

```
In [58]: softmax_reg.predict([[5, 2]])
```

```
Out[58]: array([2])
```

```
In [59]: softmax_reg.predict_proba([[5, 2]]).round(2)
```

```
Out[59]: array([[0. , 0.04, 0.96]])
```

```
In [60]: # extra code - this cell generates and saves Figure 4-25

from matplotlib.colors import ListedColormap

custom_cmap = ListedColormap(["#fafab0", "#9898ff", "#a0faa0"])

x0, x1 = np.meshgrid(np.linspace(0, 8, 500).reshape(-1, 1),
                     np.linspace(0, 3.5, 200).reshape(-1, 1))
X_new = np.c_[x0.ravel(), x1.ravel()]

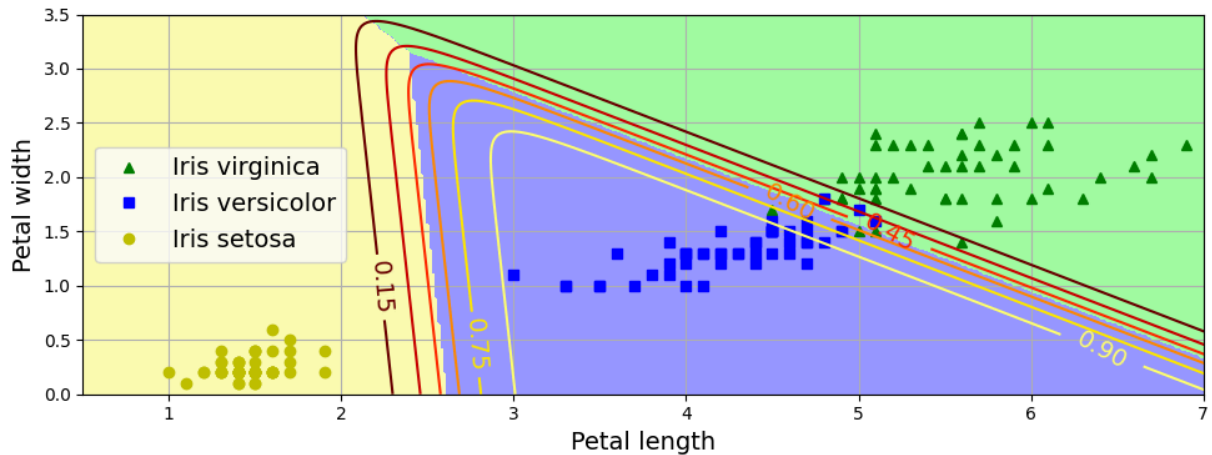
y_proba = softmax_reg.predict_proba(X_new)
y_predict = softmax_reg.predict(X_new)

zz1 = y_proba[:, 1].reshape(x0.shape)
zz = y_predict.reshape(x0.shape)

plt.figure(figsize=(10, 4))
plt.plot(X[y == 2, 0], X[y == 2, 1], "g^", label="Iris virginica")
plt.plot(X[y == 1, 0], X[y == 1, 1], "bs", label="Iris versicolor")
plt.plot(X[y == 0, 0], X[y == 0, 1], "yo", label="Iris setosa")

plt.contourf(x0, x1, zz, cmap=custom_cmap)
contour = plt.contour(x0, x1, zz1, cmap="hot")
plt.clabel(contour, inline=1)
plt.xlabel("Petal length")
plt.ylabel("Petal width")
plt.legend(loc="center left")
```

```
plt.axis([0.5, 7, 0, 3.5])
plt.grid()
save_fig("softmax_regression_contour_plot")
plt.show()
```



Given the code above as an example, how many classes can the Softmax Regression predict? Can it be used to on non-mutually exclusive classes (such as to recognize different faces in a picture)?

4. Supplementary Activity

4.1 Linear Regression

- Choose your own dataset
- Import the dataset
- Determine the number of datapoints, columns and data types
- Remove unnecessary columns
- Do data cleaning such as removing empty values.
- Perform descriptive statistics such as mean, median and mode
- Compute the correlation
- Use One-hot encoding in categorical features
- Apply feature scaling
- Perform linear regression using one independent variable only using numpy
- Perform linear regression using multiple independent variable using statsmodel
- Perform linear regression using sklearn

```
In [119... import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

```
In [62]: ictemp = pd.read_csv('/content/Ice Cream Sales - temperatures.csv')
ictemp.head()
```

Out[62]:

	Temperature	Ice Cream Profits
0	39	13.17
1	40	11.88
2	41	18.82
3	42	18.65
4	43	17.02

```
In [63]: ictemp.rename(columns={'Temperature':'Temperature(F)',
                                'Ice Cream Profits':'IceCreamProfits(USD)'}, inplace=True)
ictemp.head()
```

Out[63]:

	Temperature(F)	IceCreamProfits(USD)
0	39	13.17
1	40	11.88
2	41	18.82
3	42	18.65
4	43	17.02

```
In [64]: ictemp.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 365 entries, 0 to 364
Data columns (total 2 columns):
#   Column                Non-Null Count  Dtype  
---  -
0   Temperature(F)        365 non-null   int64  
1   IceCreamProfits(USD)  365 non-null   float64
dtypes: float64(1), int64(1)
memory usage: 5.8 KB
```

```
In [71]: ictempstat = ictemp.describe()
```

```
In [72]: ictempstat
```


Out[72]:

	Temperature(F)	IceCreamProfits(USD)
count	365.000000	365.000000
mean	71.980822	52.103616
std	13.258510	15.989004
min	39.000000	11.880000
25%	63.000000	40.650000
50%	73.000000	53.620000
75%	82.000000	63.630000
max	101.000000	89.290000

```
In [91]: mode = pd.DataFrame(ictemp.mode().loc[0])
mode.rename(columns={0: 'Mode'}, inplace=True)
mode = mode.transpose()
mode
```

Out[91]:

	Temperature(F)	IceCreamProfits(USD)
Mode	77.0	44.85

```
In [104... ictempstat = pd.concat([ictempstat, mode],axis=0)
```

```
In [103... ictempstat = pd.concat([ictempstat, mode],axis=0)
ictempstat.drop(['min', '25%', '75%', 'max'], inplace=True)
```

Out[103...]

	Temperature(F)	IceCreamProfits(USD)
mean	71.980822	52.103616
std	13.258510	15.989004
50%	73.000000	53.620000

```
In [105... ictempstat
```

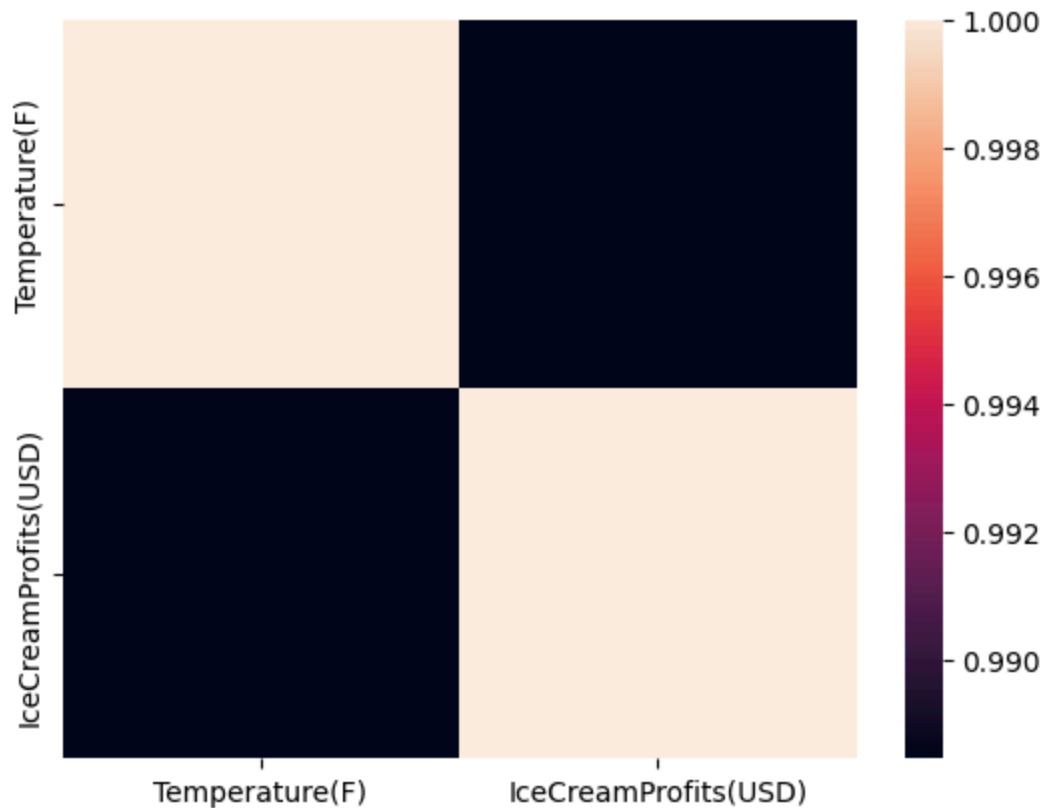
Out[105...]

	Temperature(F)	IceCreamProfits(USD)
mean	71.980822	52.103616
std	13.258510	15.989004
50%	73.000000	53.620000
Mode	77.000000	44.850000

```
In [109... ictempcorr = ictemp.corr()
```

```
In [111... sns.heatmap(ictempcorr)
```

```
Out[111... <Axes: >
```



```
In [113... ictemp.head()
```

```
Out[113... 
```

	Temperature(F)	IceCreamProfits(USD)
0	39	13.17
1	40	11.88
2	41	18.82
3	42	18.65
4	43	17.02

Linear Regression using numPy

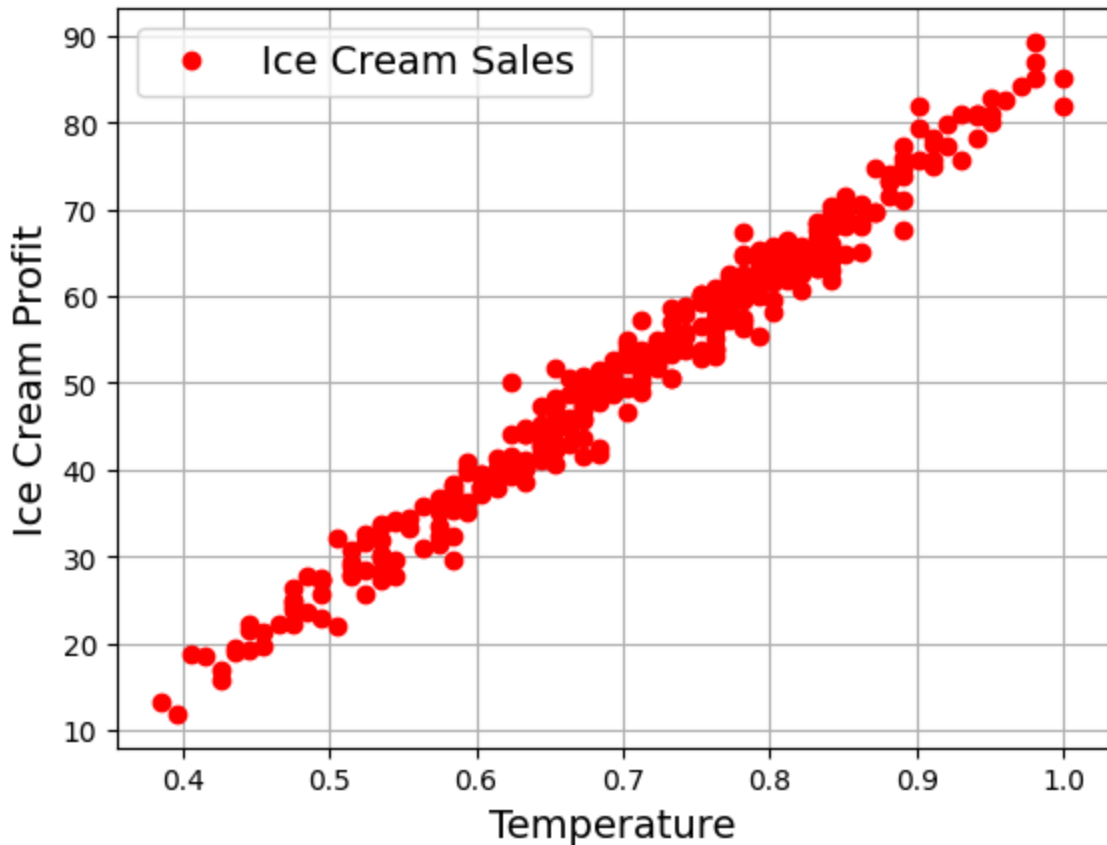
```
In [112... ictemparr = np.array(ictemp.values, 'float') #converting dataframe to numpy array
ictemparr[0:5, :] # equivalent to df.head()
```

```
Out[112...] array([[39. , 13.17],
      [40. , 11.88],
      [41. , 18.82],
      [42. , 18.65],
      [43. , 17.02]])
```

```
In [116...] X = ictemparr[:,0] # getting the temperature column
            y = ictemparr[:,1] # getting the ice cream profit
```

```
In [117...] X = X/np.max(X) # normalising the feature
```

```
In [121...] plt.plot(X,y,'ro')
            plt.plot(X,x@theta,'-')
            plt.xlabel('Temperature')
            plt.ylabel('Ice Cream Profit')
            plt.legend(['Ice Cream Sales'])
            plt.grid()
```



```
In [139...] def compute_cost(x,y,theta):
            """ computes the difference between our hypothesis and actual data points"""
            m = len(y)
            a = (1/(2*m))*np.sum(x@theta-y**2)
            return a
```

```
In [172...] def gradient(x,y,theta):

            alpha = 0.00001
            iteration = 2000
```

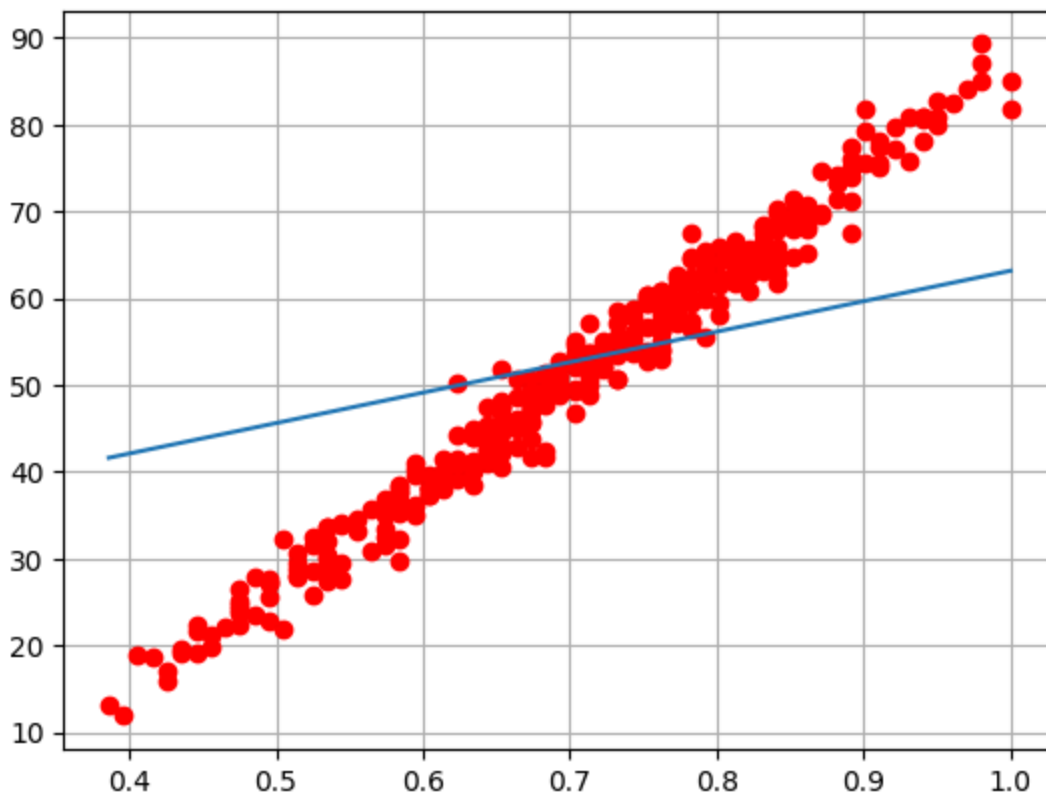
```
#gradient descend algorithm
J_history = np.zeros([iteration, 1]);
for iter in range(0,2000):
    error = (x @ theta) - y
    temp0 = theta[0] - ((alpha/m) * np.sum(error*x[:,0]))
    temp1 = theta[1] - ((alpha/m) * np.sum(error*x[:,1]))
    theta = np.array([temp0,temp1]).reshape(2,1)
    J_history[iter] = (1 / (2*m)) * (np.sum(((x @ theta)-y)**2)) #compute
return theta, J_history
```

```
In [173... m = np.size(y)
X = X.reshape([m,1])
x = np.hstack([np.ones_like(X),X])
theta = np.zeros([2,1])

theta , J = gradient(x,y,theta)
print(theta)
```

```
[[28.06299243]
 [35.10352046]]
```

```
In [175... plt.plot(X,y,'ro')
plt.plot(X,x@theta) # plotting the model
plt.grid()
```



Linear Regression using statsmodel

```
In [185... import statsmodels.api as sm

X = ictemp['Temperature(F)']
y = ictemp['IceCreamProfits(USD)']
```

```
X = sm.add_constant(X)
model = sm.OLS(y, X).fit()
print(model.summary())
```

OLS Regression Results

```
=====
Dep. Variable:      IceCreamProfits(USD)    R-squared:                0.977
Model:              OLS                    Adj. R-squared:           0.977
Method:             Least Squares          F-statistic:             1.544e+04
Date:               Fri, 06 Sep 2024        Prob (F-statistic):      1.57e-299
Time:               12:19:42                Log-Likelihood:          -840.52
No. Observations:   365                    AIC:                     1685.
Df Residuals:       363                    BIC:                     1693.
Df Model:           1
Covariance Type:    nonrobust
=====
```

	coef	std err	t	P> t	[0.025	0.975]
const	-33.6982	0.702	-47.991	0.000	-35.079	-32.317
Temperature(F)	1.1920	0.010	124.245	0.000	1.173	1.211

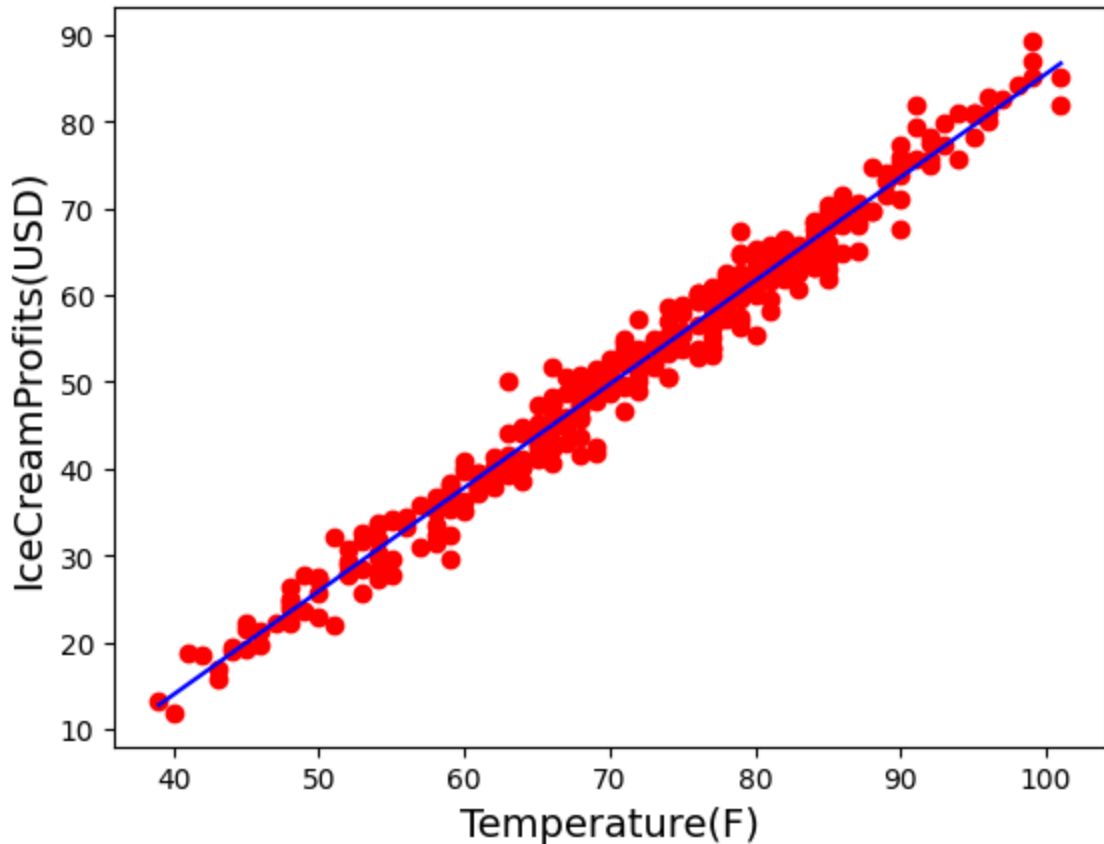
```
=====
Omnibus:            2.964    Durbin-Watson:           2.070
Prob(Omnibus):      0.227    Jarque-Bera (JB):         3.089
Skew:               -0.082    Prob(JB):                 0.213
Kurtosis:           3.420    Cond. No.                 405.
=====
```

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

```
In [189... y_pred = model.predict(X) # gets the model predicted values
plt.plot(ictemp['Temperature(F)'], y, 'ro') # plot the data from our dataset
plt.plot(ictemp['Temperature(F)'], y_pred, color='Blue') #plots the linear regressi
plt.xlabel('Temperature(F)')
plt.ylabel('IceCreamProfits(USD)')
```

```
Out[189... Text(0, 0.5, 'IceCreamProfits(USD)')
```



linear regression model using sklearn

```
In [196... #splitting the feature and target
X = ictemp['Temperature(F)'].values
y = ictemp['IceCreamProfits(USD)'].values
```

```
In [197... # splitting the data set by training and testing
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_sta
```

```
In [199... # making the model
model = LinearRegression()
model.fit(X_train.reshape(-1, 1), y_train)
```

```
Out[199... ▼ LinearRegression
LinearRegression()
```

```
In [200... #checking the coefficient
model.coef_
```

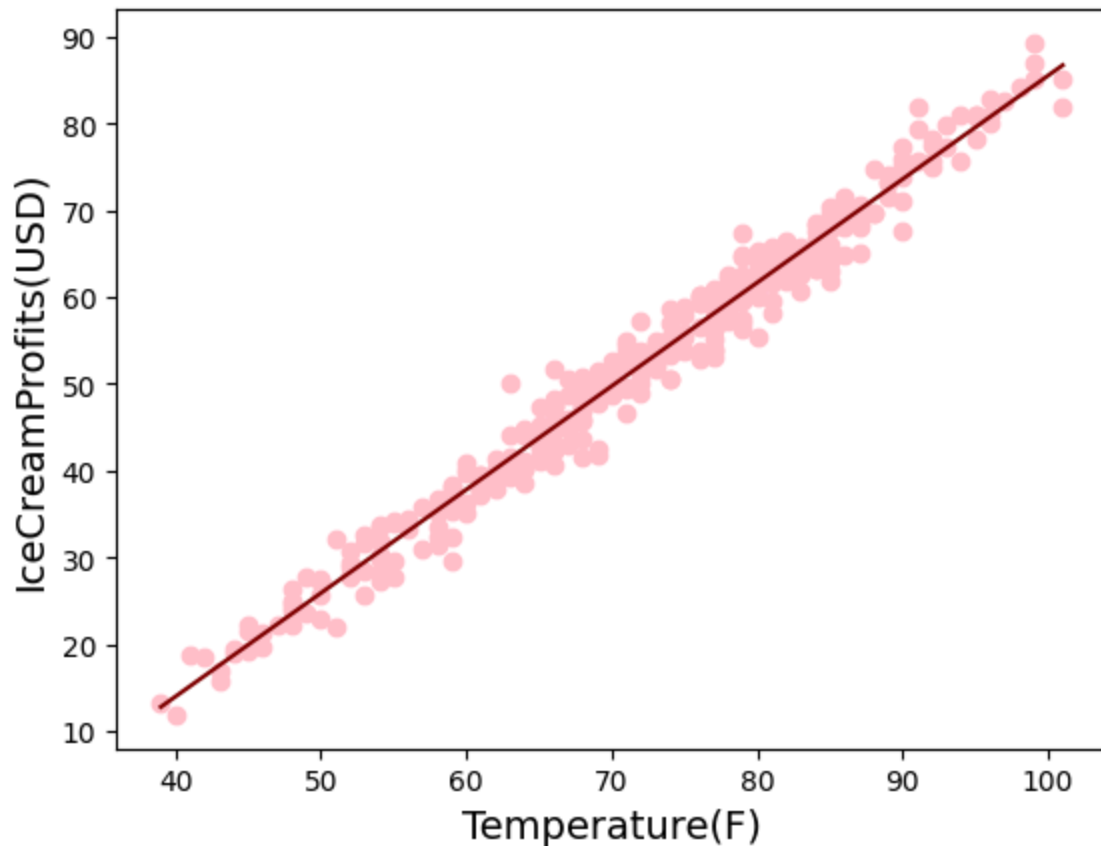
```
Out[200... array([1.19231754])
```

```
In [203... #predict using the test value
Y_pred = model.predict(X.reshape(-1, 1))
```

```
In [206... plt.scatter(ictemp['Temperature(F)'], ictemp['IceCreamProfits(USD)'], color = 'Pink')
plt.plot(ictemp['Temperature(F)'], Y_pred, color='Maroon') #plots the linear regres
```

```
plt.xlabel('Temperature(F)')  
plt.ylabel('IceCreamProfits(USD)')
```

Out[206... Text(0, 0.5, 'IceCreamProfits(USD)')



4.2 Polynomial Regression

- Choose your own dataset
- Import the dataset
- Perform polynomial regression using sklearn and polyfit
- Measure the performance for each polynomial degree.
- Plot the performance of the model for each polynomial degree.

```
In [66]: adsales = pd.read_csv('/content/Advertising Budget and Sales.csv')  
adsales.head()
```

Out[66]:

	Unnamed: 0	TV Ad Budget (\$)	Radio Ad Budget (\$)	Newspaper Ad Budget (\$)	Sales (\$)
0	1	230.1	37.8	69.2	22.1
1	2	44.5	39.3	45.1	10.4
2	3	17.2	45.9	69.3	9.3
3	4	151.5	41.3	58.5	18.5
4	5	180.8	10.8	58.4	12.9

In [67]: `adsales.drop(columns=['Unnamed: 0'], inplace=True)`
`adsales.head()`

Out[67]:

	TV Ad Budget (\$)	Radio Ad Budget (\$)	Newspaper Ad Budget (\$)	Sales (\$)
0	230.1	37.8	69.2	22.1
1	44.5	39.3	45.1	10.4
2	17.2	45.9	69.3	9.3
3	151.5	41.3	58.5	18.5
4	180.8	10.8	58.4	12.9

In [68]: `adsales.rename(columns={'TV Ad Budget ($)': 'TVBudget(USD)',
'Radio Ad Budget ($)': 'RadioBudget(USD)',
'Newspaper Ad Budget ($)': 'NewspaperBudget(USD)',
'Sales ($)': 'Sales(USD)'}, inplace=True)`
`adsales.head()`

Out[68]:

	TVBudget(USD)	RadioBudget(USD)	NewspaperBudget(USD)	Sales(USD)
0	230.1	37.8	69.2	22.1
1	44.5	39.3	45.1	10.4
2	17.2	45.9	69.3	9.3
3	151.5	41.3	58.5	18.5
4	180.8	10.8	58.4	12.9

In [69]: `adsales.info()`


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 4 columns):
#   Column                Non-Null Count  Dtype
---  -
0   TVBudget(USD)         200 non-null   float64
1   RadioBudget(USD)      200 non-null   float64
2   NewspaperBudget(USD)  200 non-null   float64
3   Sales(USD)            200 non-null   float64
dtypes: float64(4)
memory usage: 6.4 KB
```

```
In [ ]: adsales.drop(columns=['RadioBudget(USD)', 'NewspaperBudget(USD)'], inplace=True)
```

```
In [212... adsales.describe()
```

```
Out[212...      TVBudget(USD)  Sales(USD)
count      200.000000  200.000000
mean       147.042500   14.022500
std         85.854236    5.217457
min          0.700000    1.600000
25%         74.375000   10.375000
50%        149.750000   12.900000
75%        218.825000   17.400000
max        296.400000   27.000000
```

Polynomial regression using sklearn

```
In [218... X = adsales['TVBudget(USD)'].values/np.max(adsales['TVBudget(USD)']) # normalizing
y = adsales['Sales(USD)'].values
```

```
In [259... from sklearn.preprocessing import PolynomialFeatures

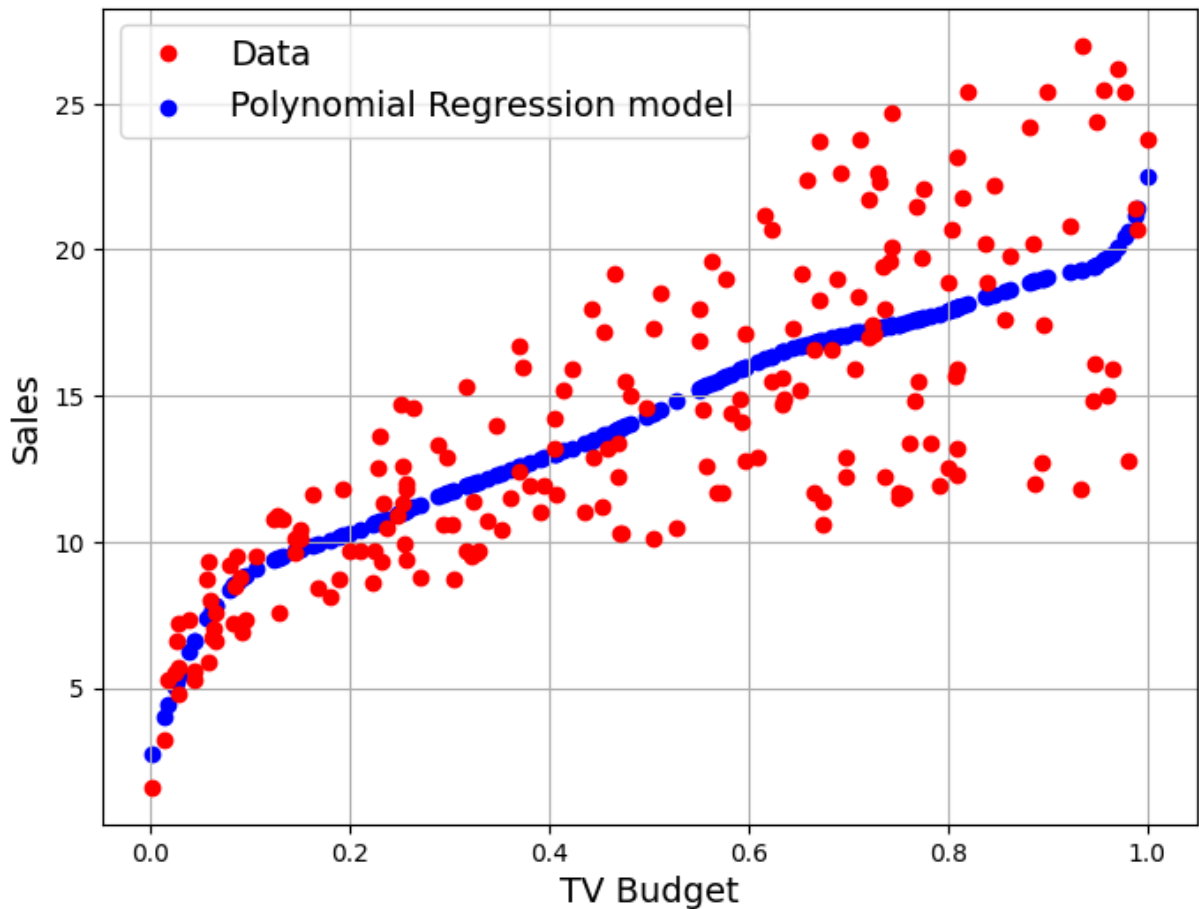
poly = PolynomialFeatures(degree=10)
X_poly = poly.fit_transform(X.reshape(-1, 1))

poly.fit(X_poly, y)
lin2 = LinearRegression()
lin2.fit(X_poly, y)
y_pred = lin2.predict(X_poly)
mse = mean_squared_error(y, y_pred)
print('the measure of the mean squared is: ',mse)
```

the measure of the mean squared is: 10.012092889244299

```
In [256... plt.figure(figsize=(8, 6))
plt.plot(X,y,'ro')
plt.scatter(X, lin2.predict(X_poly), color='Blue')
```

```
plt.xlabel('TV Budget')
plt.ylabel('Sales')
plt.legend(['Data', 'Polynomial Regression model'])
plt.grid()
```



4.3 Logistic Regression

- Choose your own dataset
- Import the dataset
- Determine the number of datapoints, columns and data types
- Remove unnecessary columns
- Do data cleaning such as removing empty values(NaN), replacing missing data .
- Perform descriptive statistics such as mean, median and mode
- Perform data visualization
- Solve classification problem using Logistic Regression
- Evaluate the model using classification report, accuracy and confusion matrix

In [270...

```
matrisk = pd.read_csv('/content/Maternal Health Risk Data Set.csv')
matrisk.head()
```

Out[270...

	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel
0	25	130	80	15.0	98.0	86	high risk
1	35	140	90	13.0	98.0	70	high risk
2	29	90	70	8.0	100.0	80	high risk
3	30	140	85	7.0	98.0	70	high risk
4	35	120	60	6.1	98.0	76	low risk

In [287...

```
#doing one hot coding
matrisk.replace({'low risk':0,'mid risk':1,'high risk':2},inplace = True)
matrisk.head()
```

Out[287...

	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel
0	25	130	80	15.0	98.0	86	2
1	35	140	90	13.0	98.0	70	2
2	29	90	70	8.0	100.0	80	2
3	30	140	85	7.0	98.0	70	2
4	35	120	60	6.1	98.0	76	0

In [288...

```
#checking null values
matrisk.isna().sum()
```

Out[288...

	0
Age	0
SystolicBP	0
DiastolicBP	0
BS	0
BodyTemp	0
HeartRate	0
RiskLevel	0

dtype: int64

In [303...

```
matriskstat = matrisk.describe()
matriskstat
```

Out[303...

	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	Ri
count	1014.000000	1014.000000	1014.000000	1014.000000	1014.000000	1014.000000	1014
mean	29.871795	113.198225	76.460552	8.725986	98.665089	74.301775	0
std	13.474386	18.403913	13.885796	3.293532	1.371384	8.088702	0
min	10.000000	70.000000	49.000000	6.000000	98.000000	7.000000	0
25%	19.000000	100.000000	65.000000	6.900000	98.000000	70.000000	0
50%	26.000000	120.000000	80.000000	7.500000	98.000000	76.000000	1
75%	39.000000	120.000000	90.000000	8.000000	98.000000	80.000000	2
max	70.000000	160.000000	100.000000	19.000000	103.000000	90.000000	2

In [304...

```
matriskstat = pd.concat([matriskstat,
                          matrisk.mode().rename(index={0:'Mode'})],axis=0)
matriskstat.drop(['count','min','25%','75%','max'], inplace=True)
matriskstat.rename(index={'50%':'median'},inplace = True)
matriskstat.head()
```

Out[304...

	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel
mean	29.871795	113.198225	76.460552	8.725986	98.665089	74.301775	0.867850
std	13.474386	18.403913	13.885796	3.293532	1.371384	8.088702	0.807353
median	26.000000	120.000000	80.000000	7.500000	98.000000	76.000000	1.000000
Mode	23.000000	120.000000	80.000000	7.500000	98.000000	70.000000	0.000000

In [320...

```
featureCols = ['Age','SystolicBP','DiastolicBP','BS','BodyTemp','HeartRate']
X = matrisk[featureCols]
y = matrisk.drop(columns=featureCols)
```

In [321...

```
y.head()
```

Out[321...

	RiskLevel
0	2
1	2
2	2
3	2
4	0

In [322...

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20, random_st
```

```
In [323... from sklearn.linear_model import LogisticRegression
#training part
model = LogisticRegression()
model.fit(X_train, y_train)
```

/usr/local/lib/python3.10/dist-packages/sklearn/utils/validation.py:1183: DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples,), for example using ravel().

```
y = column_or_1d(y, warn=True)
```

/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:460: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:

<https://scikit-learn.org/stable/modules/preprocessing.html>

Please also refer to the documentation for alternative solver options:

https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression

```
n_iter_i = _check_optimize_result(
```

```
Out[323... ▾ LogisticRegression
LogisticRegression()
```

```
In [328... #testing part
y_pred = model.predict(X_test)
```

```
In [329... y_pred
```

```
Out[329... array([1, 0, 0, 1, 2, 1, 2, 0, 0, 1, 1, 2, 0, 0, 1, 1, 0, 0, 0, 1, 2, 0,
        0, 2, 2, 0, 2, 0, 0, 0, 0, 1, 2, 0, 0, 2, 1, 0, 1, 1, 0, 1, 0, 0,
        0, 2, 0, 0, 0, 0, 0, 1, 2, 2, 0, 1, 1, 0, 1, 2, 0, 0, 1, 1, 1, 1,
        0, 0, 0, 1, 2, 0, 2, 0, 0, 2, 0, 0, 1, 0, 0, 0, 1, 0, 1, 1, 0, 0,
        0, 0, 2, 0, 2, 1, 0, 2, 0, 0, 0, 0, 1, 2, 0, 0, 0, 0, 1, 0, 1, 0,
        0, 0, 2, 2, 1, 0, 1, 1, 0, 0, 2, 0, 2, 1, 2, 0, 2, 1, 0, 2, 0, 0,
        2, 1, 1, 1, 0, 0, 2, 1, 1, 1, 0, 1, 0, 0, 0, 2, 2, 0, 2, 1, 0, 1,
        0, 1, 0, 1, 0, 2, 2, 0, 1, 2, 0, 2, 0, 1, 2, 1, 0, 2, 0, 2, 0, 2,
        1, 1, 1, 0, 0, 1, 0, 1, 0, 0, 2, 1, 0, 0, 1, 0, 0, 0, 0, 0, 2, 2,
        0, 0, 2, 0, 1])
```

```
In [331... sns.regplot(x=X_test['DiastolicBP'], y=y_pred, data=matrisk, logistic=True)
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

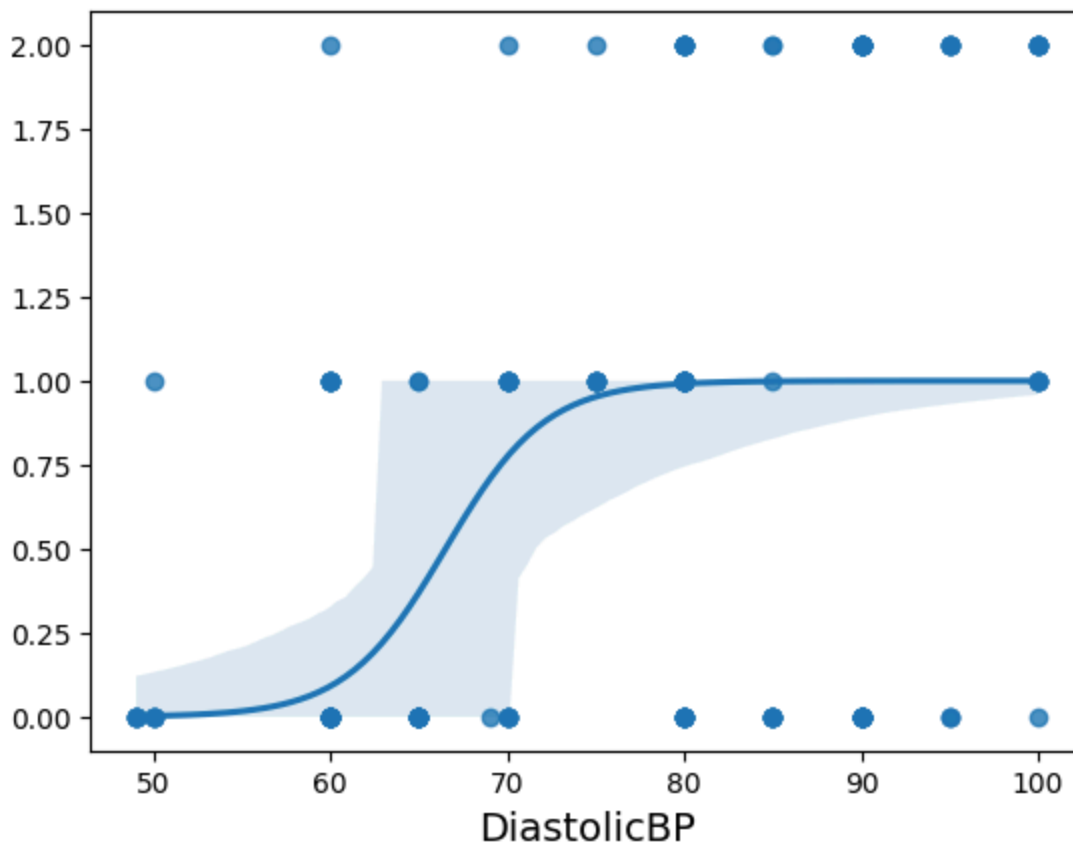
[illegible]

```

RuntimeWarning: overflow encountered in exp
  t = np.exp(-z)
/usr/local/lib/python3.10/dist-packages/statsmodels/genmod/families/links.py:198: RuntimeWarning: overflow encountered in exp
  t = np.exp(-z)
/usr/local/lib/python3.10/dist-packages/statsmodels/genmod/families/links.py:198: RuntimeWarning: overflow encountered in exp
  t = np.exp(-z)
/usr/local/lib/python3.10/dist-packages/statsmodels/genmod/families/links.py:198: RuntimeWarning: overflow encountered in exp
  t = np.exp(-z)
/usr/local/lib/python3.10/dist-packages/statsmodels/genmod/families/links.py:198: RuntimeWarning: overflow encountered in exp
  t = np.exp(-z)
/usr/local/lib/python3.10/dist-packages/statsmodels/genmod/families/links.py:198: RuntimeWarning: overflow encountered in exp
  t = np.exp(-z)
/usr/local/lib/python3.10/dist-packages/statsmodels/genmod/families/links.py:198: RuntimeWarning: overflow encountered in exp
  t = np.exp(-z)
/usr/local/lib/python3.10/dist-packages/statsmodels/genmod/families/links.py:198: RuntimeWarning: overflow encountered in exp
  t = np.exp(-z)

```

Out[331... <Axes: xlabel='DiastolicBP'>



```

In [306... #making confusion matrix
from sklearn.metrics import confusion_matrix
confmat = pd.DataFrame(confusion_matrix(y_test, y_pred))
confmat

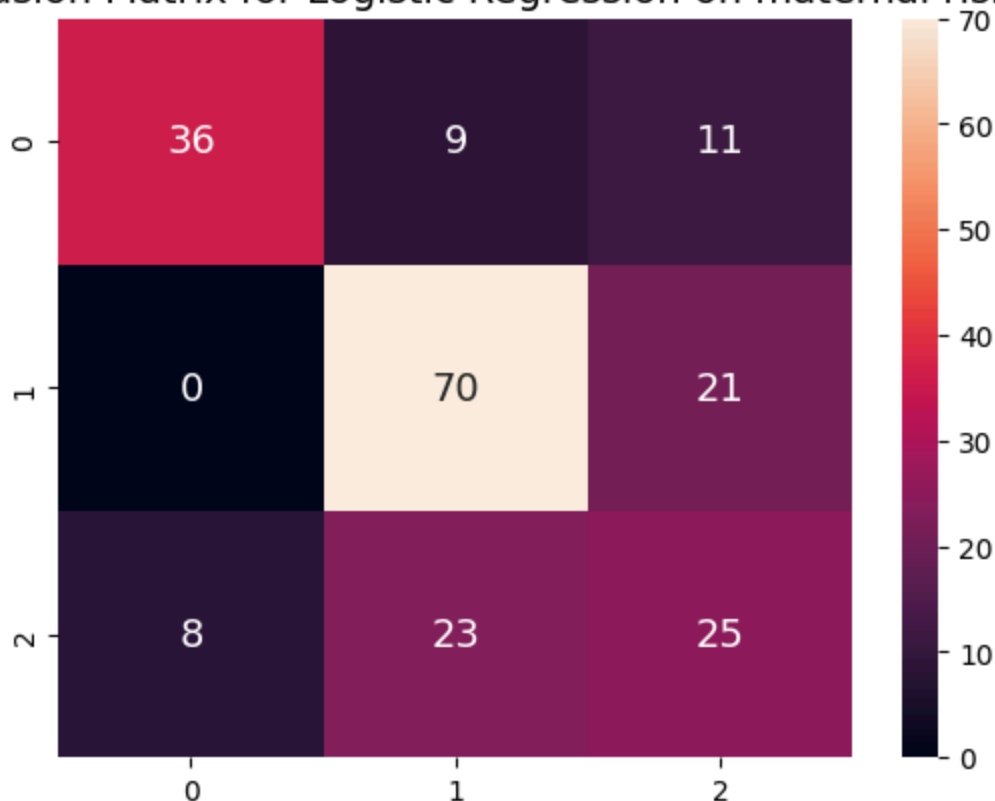
```

```
Out[306...
      0  1  2
0  36  9 11
1   0 70 21
2   8 23 25
```

```
In [310... sns.heatmap(confmat, annot=True, fmt='d')
plt.title('Confusion Matrix for Logistic Regression on maternal risk')
```

```
Out[310... Text(0.5, 1.0, 'Confusion Matrix for Logistic Regression on maternal risk')
```

Confusion Matrix for Logistic Regression on maternal risk



```
In [333... from sklearn.metrics import accuracy_score
accuracy = accuracy_score(y_test, y_pred)
print('Logistic Regression accuracy: ', accuracy)
```

Logistic Regression accuracy: 0.645320197044335

```
In [334... from sklearn.metrics import classification_report

report = classification_report(y_test, y_pred)
print('classification report: ', report)
```

classification report:			precision	recall	f1-score	support
	0	0.69	0.77	0.73		91
	1	0.44	0.45	0.44		56
	2	0.82	0.64	0.72		56
accuracy			0.65			203
macro avg			0.65	0.62	0.63	203
weighted avg			0.65	0.65	0.65	203

5. Summary, Conclusions and Lessons Learned

in this lab activity we have learned 3 types of regression, the linear, polynomial, and logistic regression. I have learned and concluded that linear regressions model or plot is straight line proportional or inversely proportional depending on the relationship of the variables. While the polynomial regression may not have a linear relationship rather on the plot or on the model it is curved. And lastly, the logistic regression creates a sigmoid function relationship on the variables. I also learned that the targets of linear and polynomial are continuous while logistic is discrete.

Proprietary Clause

Property of the Technological Institute of the Philippines (T.I.P.). No part of the materials made and uploaded in this learning management system by T.I.P. may be copied, photographed, printed, reproduced, shared, transmitted, translated, or reduced to any electronic medium or machine-readable form, in whole or in part, without the prior consent of T.I.P.

Disclaimer

Contents of this Notebook are based on Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow 3rd Edition by Aurelien Geron. A copy of the complete text can be obtained here: <https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/>