Spring 2025 CS4641/CS7641 Homework 3

Instructor: Dr. Mahdi Roozbahani

Deadline: Friday, March 28th, 11:59 pm EST

- No unapproved extension of the deadline is allowed. Submission past our 48-hour penalized acceptance period will lead to 0 credit.
- Discussion is encouraged on Ed as part of the Q/A. However, all assignments should be done individually.
- Plagiarism is a **serious offense**. You are responsible for completing your own work. You are not allowed to copy and paste, or paraphrase, or submit materials created or published by others, as if you created the materials. All materials submitted must be your own.
- All incidents of suspected dishonesty, plagiarism, or violations of the Georgia Tech Honor Code will be subject to the institute's Academic Integrity procedures. If we observe any (even small) similarities/plagiarisms detected by Gradescope or our TAs, **WE WILL DIRECTLY REPORT ALL CASES TO OSI**, which may, unfortunately, lead to a very harsh outcome. **Consequences can be severe, e.g., academic probation or dismissal, grade penalties, a 0 grade for assignments concerned, and prohibition from withdrawing from the class.**

Instructions for the assignment

- This assignment consists of both programming and theory questions.
- Unless a theory question explicitly states that no work is required to be shown, you must provide an explanation, justification, or calculation for your answer.
- To switch between cell for code and for markdown, see the menu -> Cell -> Cell Type
- You can directly type LaTeX equations into markdown cells using $ ... $ for inline and $$ ... $$ for equations. Using LaTeX to format your answers is highly encouraged.
- If a question requires a picture, you could use this syntax `<img src="" style="width: 300px;"/>` to include them within your ipython notebook.
- Your write-up must be submitted in PDF format. Please ensure all questions are answered within the Jupyter Notebook using either Markdown or LaTeX. We will **NOT** accept handwritten work. Make sure that your work is formatted correctly, for example, submit $\sum_{i=0} x_i$ instead of \text{sum_{i=0} x_i}
- When submitting the non-programming part of your assignment, you must correctly map pages of your PDF to each question/subquestion to reflect where they appear. **Improperly mapped questions may not be graded correctly and/or will result in point deductions for the error.**
- All assignments should be done individually, and each student must write up and submit their own answers.
- **Graduate Students**: You are required to complete any sections marked as Bonus for Undergrads

Using the autograder

- Grads will find three assignments on Gradescope that correspond to HW3: "Assignment 3 Programming", "Assignment 3 - Non-programming" and "Assignment 3 Programming - Bonus for all". Undergrads will find an additional assignment called "Assignment 3 Programming - Bonus for Undergrads".
- You will submit your code for the autograder in the Assignment 3 Programming sections. Please refer to the Deliverables and Point Distribution section for what parts

are considered required, bonus for undergrads, and bonus for all.

- We provided you different .py files and we added libraries in those files please DO NOT remove those lines and add your code after those lines. Note that these are the only allowed libraries that you can use for the homework.
- You are allowed to make as many submissions until the deadline as you like. Additionally, note that the autograder tests each function separately, therefore it can serve as a useful tool to help you debug your code if you are not sure of what part of your implementation might have an issue.
- **For the "Assignment 3 - Non-programming" part, you will need to submit to Gradescope a PDF copy of your Jupyter Notebook with the cell outputs.** Please refer to the **Deliverables and Point Distribution** section for an outline of the non-programming questions.
- **When submitting to Gradescope, please make sure to mark the page(s) corresponding to each problem/sub-problem. The pages in the PDF should be of size 8.5" x 11", otherwise there may be a deduction in points for extra long sheets.**

Using the local tests

- For some of the programming questions we have included a local test using a small toy dataset to aid in debugging. The local test sample data and outputs are stored in .py files in the **local_tests_folder**. The actual local tests are stored in **localtests.py**. Both can be found under the **utilities** folder.
- There are no points associated with passing or failing the local tests, you must still pass the autograder to get points.
- **It is possible to fail the local test and pass the autograder** since the autograder has a certain allowed error tolerance while the local test allowed error may be smaller. Likewise, passing the local tests does not guarantee passing the autograder.
- **You do not need to pass both local and autograder tests to get points, passing the Gradescope autograder is sufficient for credit.**
- It might be helpful to comment out the tests for functions that have not been completed yet.
- It is recommended to test the functions as it gets completed instead of completing the whole class and then testing. This may help in isolating errors. Do not solely rely on the local tests, continue to test on the autograder regularly as well.

Deliverables and Points Distribution

Q1: Eigenfaces [30pts]

Deliverables: eigenfaces.py

- **1.1 Eigenfaces** [30pts] - **[Programming]**
  - svd [5pts]
  - compress [5pts]
  - rebuild_svd [5pts]
  - compression_ratio [5pts]
  - recovered_variance_proportion [5pts]
  - compute_eigenfaces [5pts]
- **1.2 Eigenface Demonstration** [0pts]

Q2: Understanding PCA [20pts + 2.1% Bonus for All]

Deliverables: pca.py, svd_recommender.py, and Written portion

- **2.1 PCA Implementation** [10pts] - **[Programming]**
  - fit [5pts]
  - transform [2pts]
  - transform_rv [3pts]
- **2.2 Visualize** [5pts] - **[Written]**
- **2.3 PCA Reduced Facemask Dataset Analysis** [5pts] - **[Written]**
- **2.4 Movie Recommendation with SVD** [2.1% Bonus for All] - **[Programming]** | **[Written]**
  - recommender_svd [0.70%]
  - predict [0.70%]
  - Comparison: SVD vs. PCA [0.70%]

Q3: Regression and Regularization [72pts: 52pts + 2.3% Bonus for All + 20pts Grad / 6% Bonus for Undergrads]

Deliverables: regression.py and Written portion

- **3.1 About RMSE** [3pts] - **[Written]**
- **3.2 Regression and Regularization Implementations** [50pts: 30pts + 20pts Grad / 6% Bonus for Undergrad] - **[Programming]**
  - RMSE [5pts]
  - Construct Poly Features 1D [2pts]
  - Construct Poly Features 2D [3pts]
  - Prediction [5pts]
  - Linear Fit Closed Form [5pts]
  - Ridge Fit Closed Form [5pts]
  - Cross Validation [5pts]
  - Linear Gradient Descent [5pts Grad/1.5% Bonus for Undergrad]
  - Linear Stochastic Gradient Descent [5pts Grad/1.5% Bonus for Undergrad]
  - Ridge Gradient Descent [5pts Grad/1.5% Bonus for Undergrad]
  - Ridge Stochastic Gradient Descent [5pts Grad/1.5% Bonus for Undergrad]
- **3.3 Testing: General Functions and Linear Regression** [5pts] - **[Programming]**
- **3.4 Testing: Ridge Regression** [5pts] - **[Programming]**

- **3.5 Linear vs. Ridge Regression Analysis** [4pts] - **[Written]**
- **3.6 Cross Validation Hyperparameter Search** [5pts] - **[Programming]**
- **3.7 Noisy Input Samples in Linear Regression** [2.3% Bonus for All] - **[Written]**

Q4: Naive Bayes and Logistic Regression [40pts]

Deliverables: logistic_regression.py and Written portion

- **4.1 Profile Screening Problem** [7pts] - **[Written]**
  - Profile Screening Problem using Naive Bayes [5pts]
  - AI-Driven Profile Screening [2pts]
- **4.2 News Data Sentiment Classification Using Logistic Regression** [30 pts] - **[Programming]**
  - sigmoid [2 pts]
  - bias_augment [3 pts]
  - predict_probs [5 pts]
  - predict_labels [2 pts]
  - loss [3 pts]
  - gradient [3 pts]
  - accuracy [2 pts]
  - evaluate [5 pts]
  - fit [5 pts]
- **4.3 Logistic Regression Threshold Experiments** [3 pts] - **[Programming]**

Q5: Feature Selection [30pts]

Deliverables: feature_reduction.py

- **5.1 Feature Reduction** [30pts] - **[Programming]**
  - forward_selection [15pts]
  - backward_elimination [15pts]

Q6: Imbalanced Classes in Classification Tasks [5.6% Bonus for All]

Deliverables: smote.py

- **6.1 A More Comprehensive Measure** [1.75% Bonus for All] - **[Programming]**
  - generate_confusion_matrix [0.75%]

- roc_auc (Receiver Operating Characteristics Area under the Curve) [1%]

- **6.2 SMOTE** [3.85% Bonus for All] - **[Programming]**
  - interpolate [1%]
  - k_nearest_neighbors [0.75%]
  - smote [2.1%]

Points Totals:

- Total Base Undergrads: 172 pts
- Total Base Grads: 192 pts

Submission Instructions

Use the zip utility `python zip_submission.py` to zip up your assignment. This utility will generate a zip file `HW3_programming.zip`, which you can submit to every section (base, bonus for undergrad, and bonus for all). Additionally, this utility will convert this notebook into a pdf for submission to the non-programming section.

Alternatively, you can manually submit the following files to their respective assignments on Gradescope for the programming portions:

- **Assignment 3 Programming**
  - eigenfaces.py
  - pca.py
  - regression.py
  - logistic_regression.py
  - feature_reduction.py
- **Assignment 3 Bonus for Undergrad - Programming**
  - regression.py
- **Assignment 3 Bonus for All - Programming**
  - smote.py
  - eigenfaces.py (used as an import in svd_recommender.py)
  - svd_recommender.py
- **Assignment 3 Non-Programming**
  - Use some tool to convert this notebook into a pdf.
  - It is your responsibility to make sure that all LaTeX renders, all generated matplotlib figures render, none of your answers are clipped, the font is a reasonable size, the pdf is a reasonable resolution, the pdf is the correct size (8.5"x11"). Failure to do so may result in heavy penalties.

0 Set up

This notebook is tested under python 3.11._, and the corresponding packages can be downloaded from miniconda. You may also want to get yourself familiar with several packages:

- jupyter notebook
- numpy

- matplotlib
- sklearn

There is a VS Code and Anaconda Setup Tutorial on Ed under the "Links" category. You can also find brief setup instructions in the environment folder.

Please implement the functions that have `raise NotImplementedError`, and after you finish the coding, please delete or comment out `raise NotImplementedError`.

Library imports

In [1]:
```
###############################
### DO NOT CHANGE THIS CELL ###
###############################
# This is cell which sets up some of the modules you might need
# Please do not change the cell or import any additional packages.

import os
import sys
import warnings

import matplotlib
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt
from sklearn.datasets import load_breast_cancer, load_iris, make_classification
from sklearn.feature_extraction import text
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, mean_squared_error
from sklearn.model_selection import train_test_split
from sklearn.svm import SVC

print("Version information")

print("python: {}".format(sys.version))
print("matplotlib: {}".format(matplotlib.__version__))
print("numpy: {}".format(np.__version__))

warnings.filterwarnings("ignore")

%matplotlib inline
%load_ext autoreload
%autoreload 2

STUDENT_VERSION = 0
EO_TEXT, EO_FONT, EO_COLOR = (
    "TA VERSION",
    "Chalkduster",
    "gray",
)
EO_ALPHA, EO_SIZE, EO_ROT = 0.7, 90, 40
```

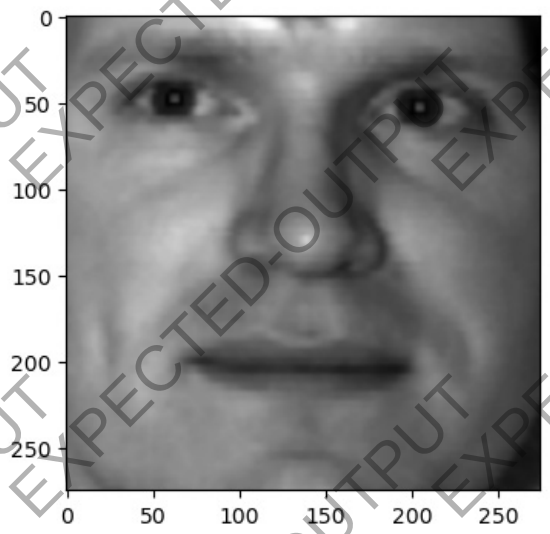
```
Version information
python: 3.11.11 (main, Dec 11 2024, 16:28:39) [GCC 11.2.0]
matplotlib: 3.10.0
numpy: 1.26.3
```

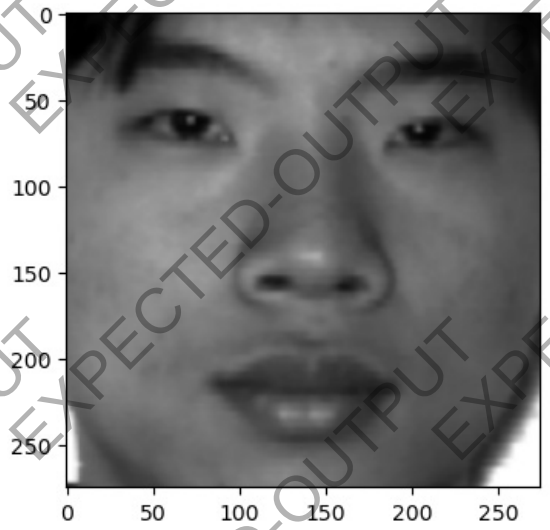
Q1: Eigenfaces [30 pts] **[P]**

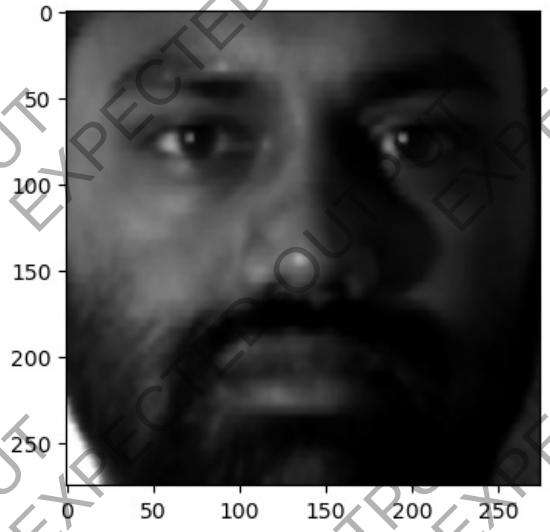
Load images data and plot

The Yale Face Database is a set of pose-normalized images of subjects with varying appearance and lighting. For the purposes of this question, you'll be passed the data already preprocessed. The data is reduced to (64,64), grayscaled, then flattened to (4096,). Below is a sample of some of the images in the dataset.

In [2]:
```
##############################
### DO NOT CHANGE THIS CELL ###
##############################
imgNum = 0
images = []
for filename in sorted(os.listdir("./data/faces")):
    img_path = os.path.join("./data/faces", filename)
    image = plt.imread(img_path)
    image = np.dot(
        image[..., :3], [0.299, 0.587, 0.114]
    ) # convert 3D RGB array into 2D grayscale array
    if imgNum == 0 or imgNum == 9 or imgNum == 19:
        fig = plt.figure(figsize=(4, 4))
        plt.imshow(image, cmap="gray")
        plt.show()
    images.append(image.flatten())  # Flatten into a 1D vector
    imgNum += 1
faces = np.array(images)
```







1.1 Eigenfaces [30pts] **[P]**

Singular Value Decomposition (SVD) is a matrix decomposition technique. For a real-valued matrix M,

$$M_{m\times n} = U_{m\times m}\Sigma_{m\times n}V^T_{n\times n} = \begin{bmatrix} | & | & & | \\ \to u_1 & \to u_2 & \cdots & \to u_m \\ | & | & & | \end{bmatrix} \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_{\min(m,n)} \end{bmatrix} \begin{bmatrix} - & \to v_1^T & - \\ - & \to v_2^T & - \\ & \vdots & \\ - & \to v_n^T & - \end{bmatrix}$$

This product $U\Sigma V^T$ can be simplified to a sum of outer vector products between the left singular vectors u and the right singular vectors v.

$$M = \sum_{i=1}\sigma_i \cdot \to u_i \to v_i^T$$

This has a nice interpretation in ML as a dimensionality reduction technique. Note that the outer product $\to u_i \to v_i^T$ creates a facsimile of the image with even the proper shape: m × n. This is then weighted by the singular value σ_i. Higher singular values therefore denote that the corresponding singular vectors have captured more useful information. Thus, we can discard the singular values that are small (and their corresponding singular vectors), while still retaining much of the variance present in the original image.

This is the standard usage of SVD. For any matrix, you can use that formula $M = \sum_{i=1}\sigma_i \cdot \to u_i \to v_i^T$, but you can terminate the sum early, allowing you to control how much information you want to retain.

The proportion of variance captured by the i^{th} product is $\sigma_{2i}\sum_{j=1}\sigma_{2j}$ and we can sum this over the singular values that we choose to retain to calculate how much variance we capture.

Eigenfaces are used in facial recognition and refer to a set of basis faces derived from a set of faces. To calculate the eigenfaces, apply SVD on the centered matrix of all

images and isolate the right singular vectors (the rows of V^T). This is a technique called principal component analysis, which uses SVD to extract principal components, which, in this case, we call eigenfaces.

In the **eigenfaces.py** file, complete the following functions:

- **svd**: Hint 2 may be useful.
- **compute_eigenfaces**
- **compress**
- **rebuild_svd**
- **compression_ratio**: Hint 1 may be useful
- **recovered_variance_proportion**: Hint 1 may be useful

HINT 1: http://timbaumann.info/svd-image-compression-demo/ is a useful article on image compression and compression ratio.

HINT 2: If you have never used `np.linalg.svd`, it might be helpful to read Numpy's SVD documentation and note the particularities of the V matrix and that it is returned already transposed. Additionally, note how `full_matrices=False` affects output shape.

HINT 3: The shape of S resulting from SVD may change depending on if N > D, N < D, or N = D. Therefore, when checking the shape of S, note that min(N,D) means the value should be equal to whichever is lower between N and D.

1.1.1 Local Tests for Eigenfaces [No Points]

You may test your implementation of the functions contained in **eigenfaces.py** in the cell below. Feel free to comment out tests for functions that have not been completed yet. See Using the Local Tests for more details.

In [3]:
```
##############################
### DO NOT CHANGE THIS CELL ###
##############################
from utilities.localtests import TestEigenfaces

unittest_ef = TestEigenfaces()

unittest_ef.test_svd()
unittest_ef.test_compress()
unittest_ef.test_rebuild_svd()
unittest_ef.test_compression_ratio()
unittest_ef.test_recovered_variance_proportion()
unittest_ef.test_compute_eigenfaces()
```

```
UnitTest passed successfully for "SVD calculation"!
UnitTest passed successfully for "Image compression"!
UnitTest passed successfully for "SVD reconstruction"!
UnitTest passed successfully for "Compression ratio"!
UnitTest passed successfully for "Recovered variance proportion"!
UnitTest passed successfully for "Eigenfaces"!
```

1.2 Eigenface Demonstration [No Points]

This question will utilize your implementation of the functions from Q1.1 to display the eigenfaces calculated for various images as well as the eigenfaces layered upon each other. You can directly execute the following cell without modifying it, provided that you have already implemented the functions in Q1.1

Running this cell is primarily for your own understanding of the compression process.

In [4]:
```
##############################
```

```
### DO NOT CHANGE THIS CELL ###
##############################
from eigenfaces import Eigenfaces

ef = Eigenfaces()
eigenfaces = ef.compute_eigenfaces(faces, 8)  # 8 eigenfaces calculated
eigenfaces = (eigenfaces - eigenfaces.min()) / (
    eigenfaces.max() - eigenfaces.min()
)  # normalize
num_eigenfaces = eigenfaces.shape[0]
fig, axes = plt.subplots(1, num_eigenfaces, figsize=(20, 7))

for i in range(num_eigenfaces):
    ax = axes[i]
    ax.imshow(eigenfaces[i].reshape((275, 275)), cmap="gray")
    ax.axis("off")
    ax.set_title(f"Eigenface {i+1}")

plt.show()
```

In [5]:

```
##############################
### DO NOT CHANGE THIS CELL ###
##############################
from eigenfaces import Eigenfaces

ef = Eigenfaces()
merged_eigenface = np.sum(eigenfaces, axis=0)
merged_eigenface = (merged_eigenface - merged_eigenface.min()) / (
    merged_eigenface.max() - merged_eigenface.min()
)

plt.figure(figsize=(5, 5))
plt.imshow(merged_eigenface.reshape((275, 275)), cmap="gray")
plt.axis("off")
plt.title("Superimposed Eigenfaces")
plt.show()
```

Superimposed Eigenfaces

Q2: Understanding PCA [20 pts + 2.1% Bonus for All] **[P]** | **[W]**

Principal Component Analysis (PCA) is a dimensionality reduction technique that reduces dimensions or features while still preserving the maximum (or close-to) amount of variance. This is useful when analyzing large datasets that contain a high number of dimensions or features that may be correlated. PCA aims to eliminate features that are highly correlated and only retain the important/uncorrelated ones that can describe most or all the variance in the data. This enables better interpretability and visualization of the multi-dimensional data. In this problem, we will investigate how PCA can be used to improve features for regression and classification tasks and how the data itself affects the behavior of PCA.

As seen briefly in the previous question, PCA uses SVD. In PCA, we first center the data by subtracting the mean of each feature. SVD is well suited for this task since each singular value tells us the amount of variance captured in each component for a given matrix (e.g. image). Hence, we can use SVD to extract data only in directions with high variances using either a threshold of the amount of variance or the number of bases/components. Here, we will reduce the data to a set number of components.

Recall that centering the matrix $X_c = X - \bar{X}$, then calculating the Gram matrix $X_{Tc}X_c$ gives us a result proportional to the covariance matrix of X. Conveniently, SVD helps quite a bit. We can write: $X_{Tc}X_c = (U\Sigma V^T)^T U\Sigma V^T = (V\Sigma^T U^T)U\Sigma V^T = V\Sigma^2 V^T$ If you remember from linear algebra, this is similar to an eigendecomposition. V act as the eigenvectors of the covariance matrix, and Σ^2 are proportional to the eigenvalues of the covariance matrix. This means two important things for PCA:

- The matrix V^T, often referred to as the *right singular vectors* of X, is equivalent to the *eigenvectors* of X^TX.
- S^2 is equivalent to the *eigenvalues* of X^TX.

The first n "principal components" of X are obtained by projecting X onto the first n vectors from V^T. Since these are eigenvectors of the covariance matrix, this projection will constitute some linear combination of the features that maximizes variance. Additionally, Σ^2 gives a measure of the variance retained.

2.1 Implementation [10 pts] **[P]**

Implement PCA. In the **pca.py** file, complete the following functions:

- **fit**: You may use `np.linalg.svd`. Set `full_matrices=False`. Hint 1 may be useful.
- **transform**
- **transform_rv**: You may find `np.cumsum` helpful for this function.

Assume a dataset is composed of N datapoints, each of which has D features with D < N. The dimension of our data would be D. However, it is possible that many of these dimensions contain redundant information. Each feature explains part of the variance in our dataset, and some features may explain more variance than others.

HINT 1: Make sure you remember to first center your data by subtracting the mean of each feature.

2.1.1 Local Tests for PCA [No Points]

You may test your implementation of the functions contained in **pca.py** in the cell below. Feel free to comment out tests for functions that have not been completed yet. See Using the Local Tests for more details.

In [6]:
```
##############################
### DO NOT CHANGE THIS CELL ###
##############################

from utilities.localtests import TestPCA

unittest_pca = TestPCA()
unittest_pca.test_pca()
unittest_pca.test_transform()
unittest_pca.test_transform_rv()
```

```
UnitTest passed successfully for "PCA fit"!
UnitTest passed successfully for "PCA transform"!
UnitTest passed successfully for "PCA transform with recovered variance"!
```

2.2 Visualize [5 pts] **[W]**

PCA is used to transform multivariate data tables into smaller sets so as to observe the hidden trends and variations in the data. It can also be used as a feature extractor for images. Here you will visualize two datasets using PCA, including Iris Dataset and Facemask Dataset.

In the **pca.py**, complete the following function:

- **visualize**: Use your implementation of PCA and reduce the dataset such that it contains only two features. Using Matplotlib, make a 2D scatterplot of the data points using these features. Make sure to differentiate the data points according to their true labels using color.
  Then, use your implementation of PCA to reduce the dataset such that it contains only three features, and visualize that data with a 3D scatterplot.
  Finally, choose 2 random features in the dataset, make a 2D scatterplot, and examine the difference.

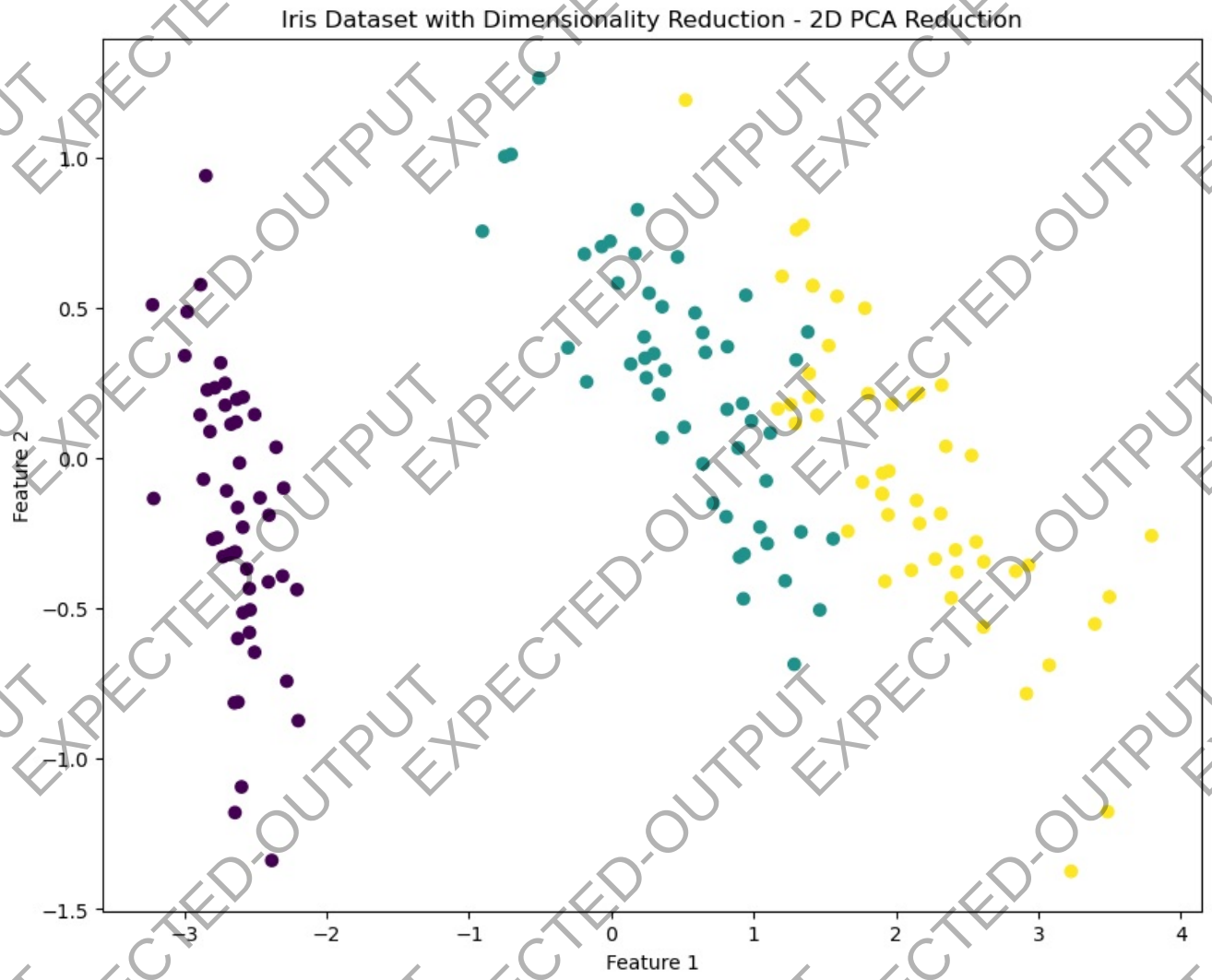
The datasets have already been loaded for you in the subsequent cells.

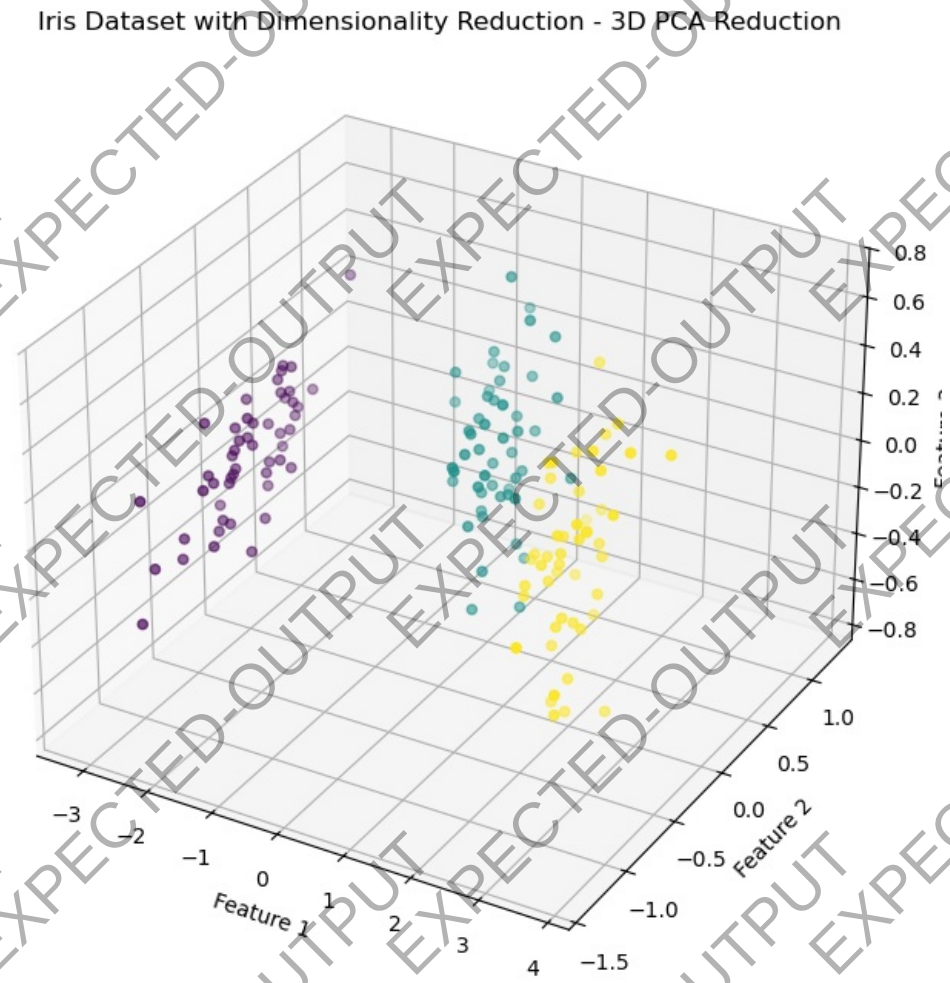
NOTE: Here, we won't be testing for accuracy. Even with correct implementations of PCA, the accuracy can differ from the TA solution. That is fine as long as the visualizations come out similar.

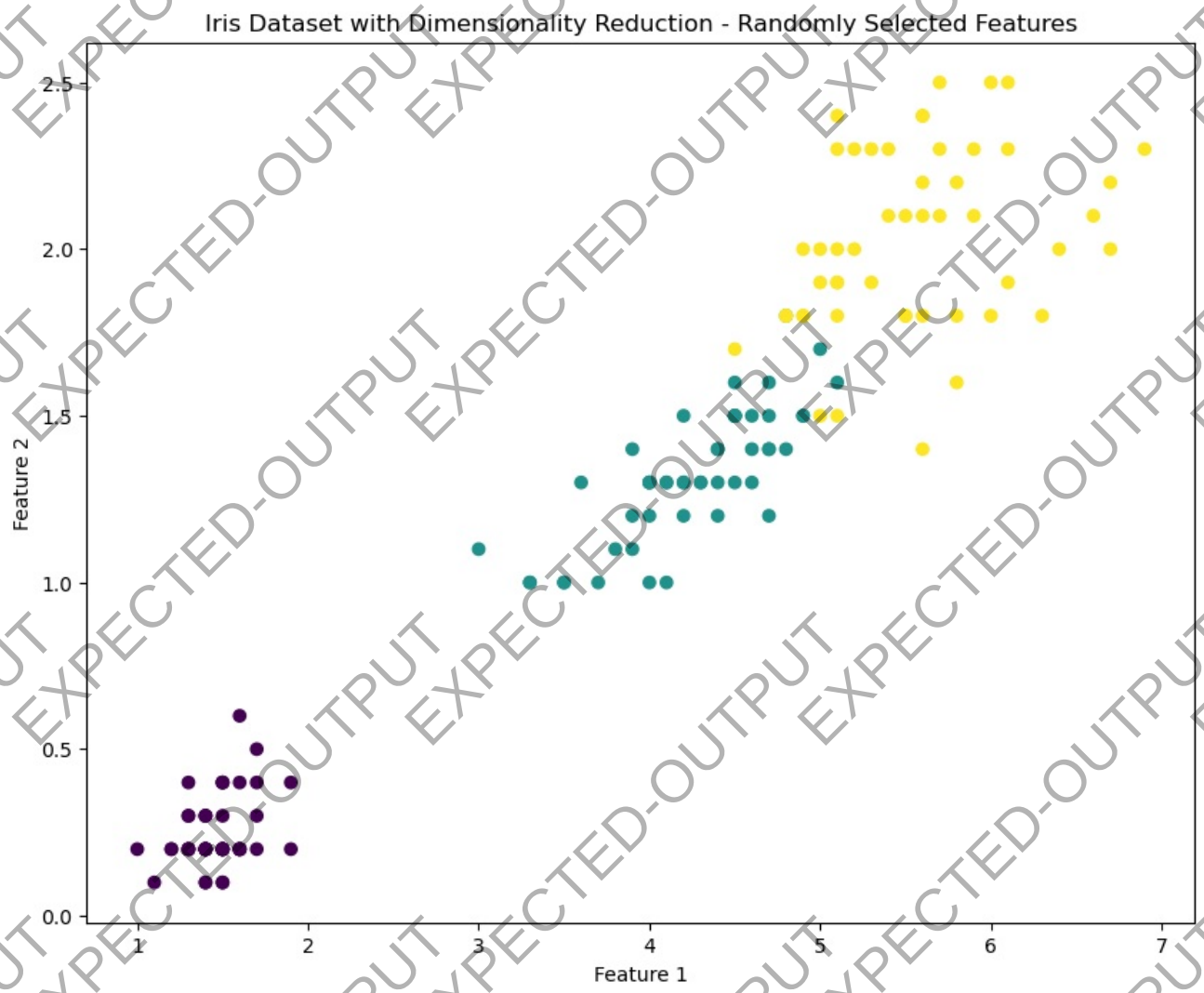
Iris Dataset

In [7]:
```
##############################
### DO NOT CHANGE THIS CELL ###
##############################
```

```python
# Use PCA for visualization of iris dataset

from pca import PCA

iris_data = load_iris(return_X_y=True)

X = iris_data[0]
y = iris_data[1]

fig_title = "Iris Dataset with Dimensionality Reduction"
PCA().visualize(X, y, fig_title)
```





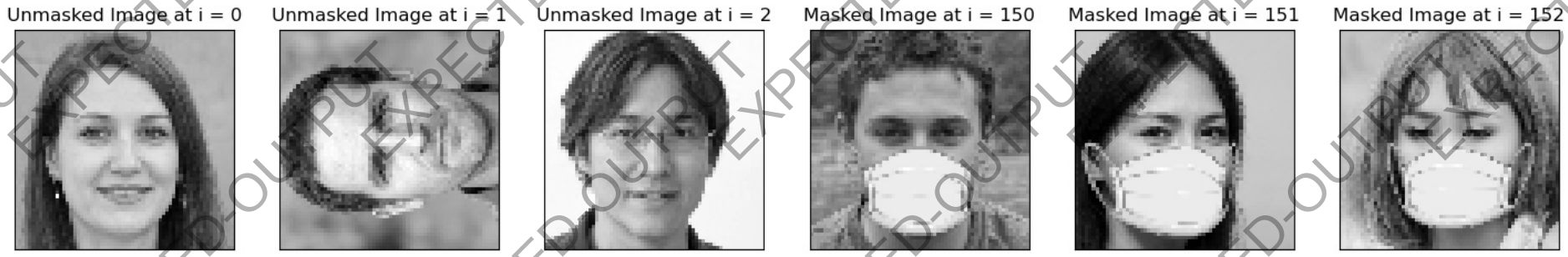


2.3 PCA Reduced Facemask Dataset Analysis [5 pts] **[W]**

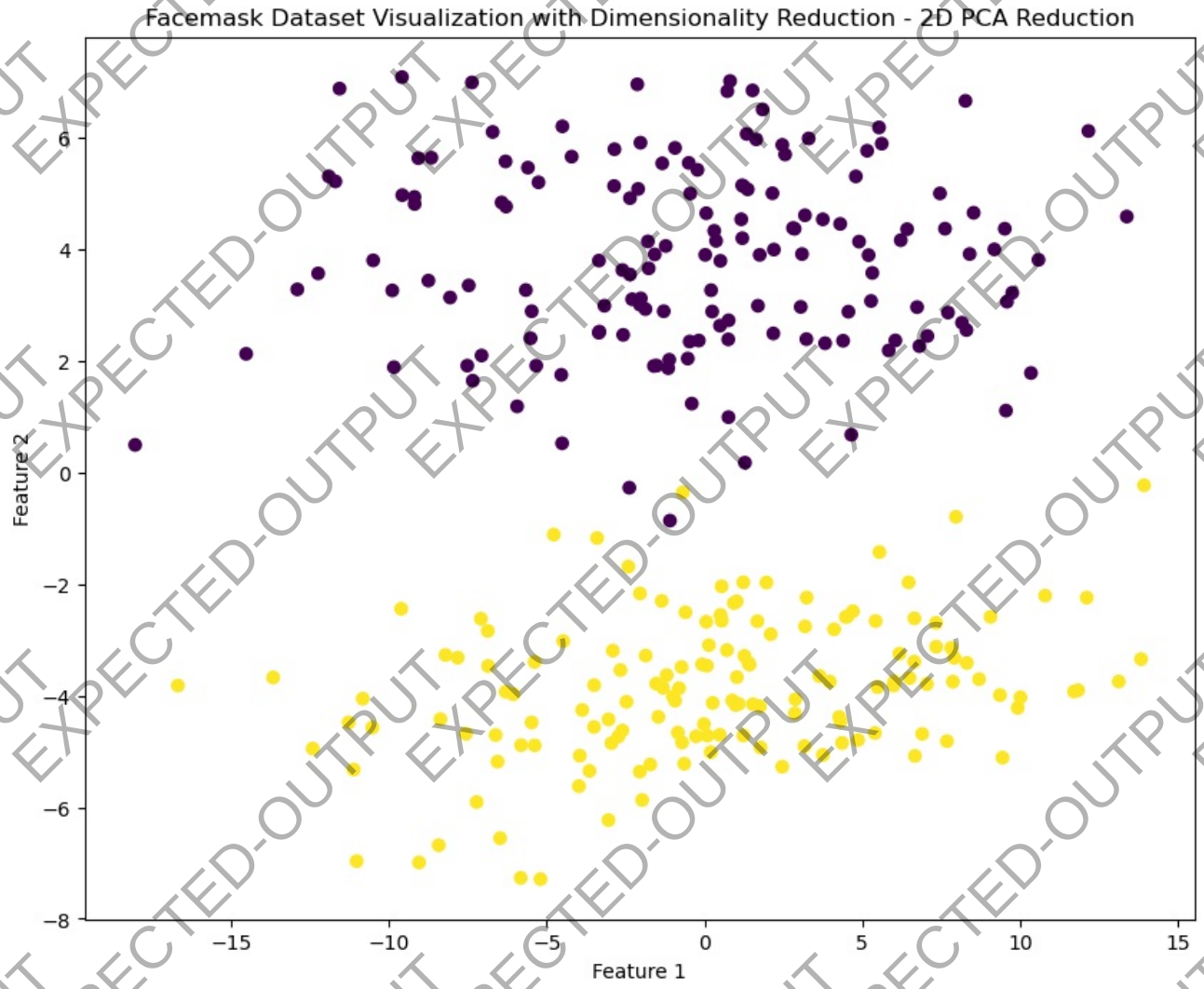
Facemask Dataset

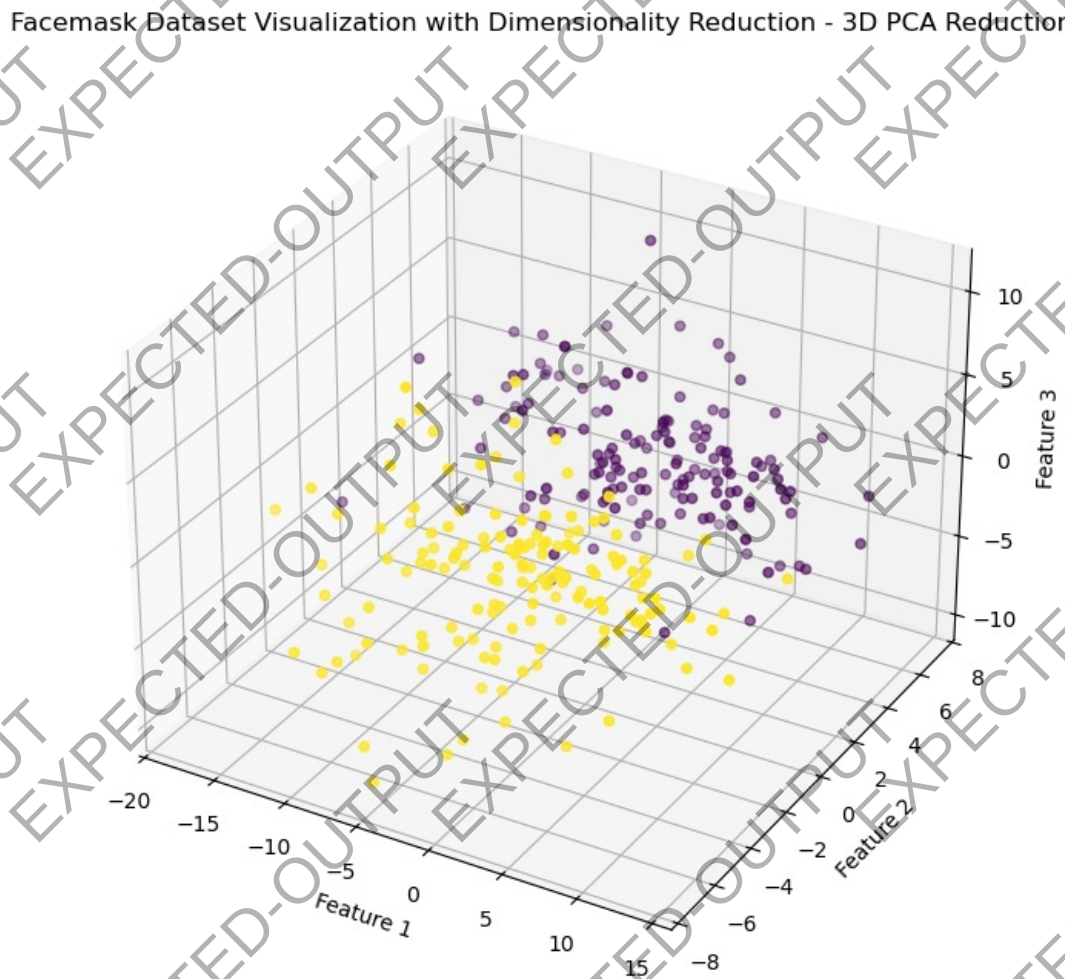
The masked and unmasked dataset is made up of grayscale images of human faces facing forward. Half of these images are faces that are completely unmasked, and the remaining images show half of the face covered with an artificially generated face mask. The images have already been preprocessed, they are also reduced to a small size of 64x64 pixels and then reshaped into a feature vector of 4096 pixels. Below is a sample of some of the images in the dataset.

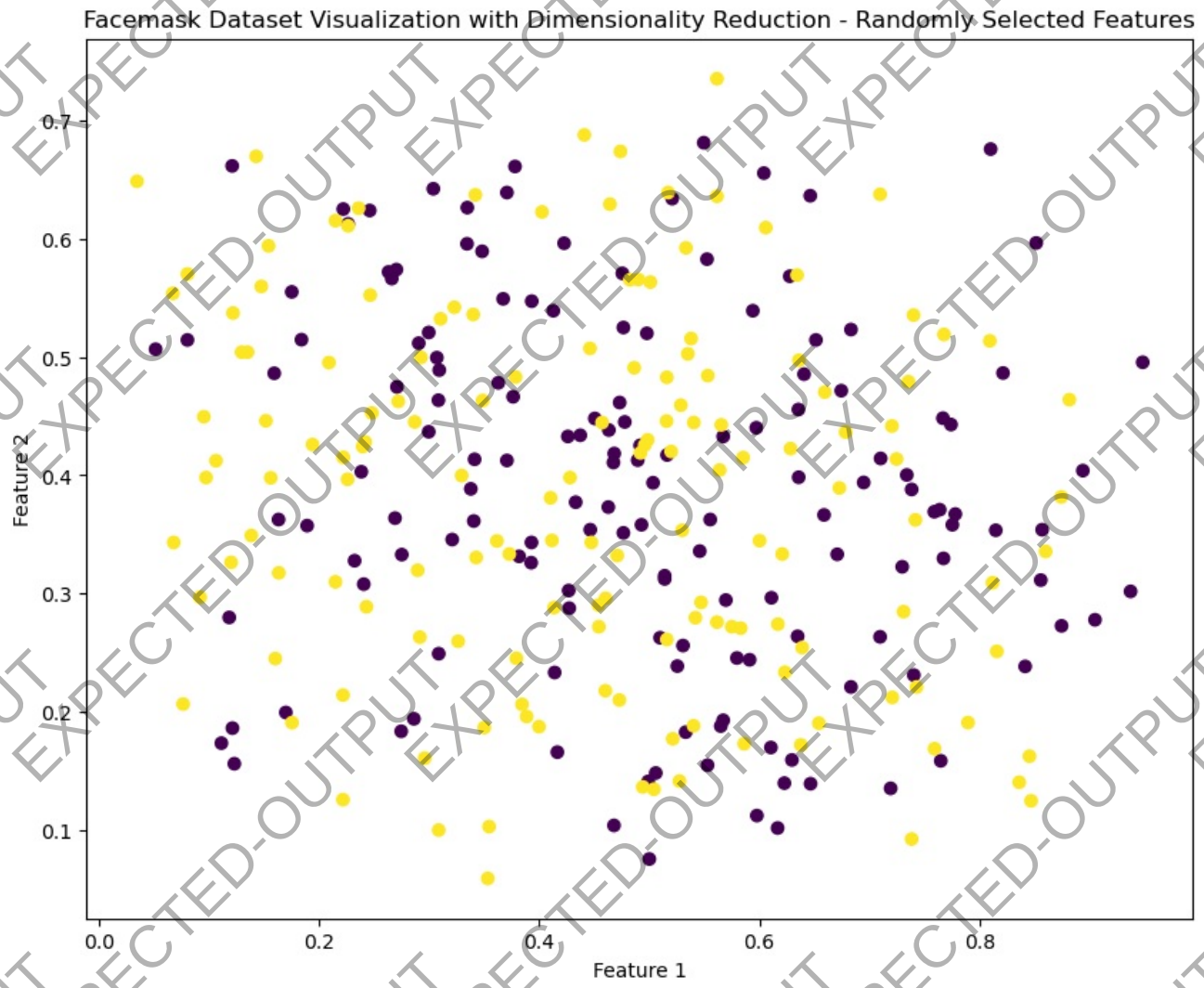
In [8]:
```
##############################
### DO NOT CHANGE THIS CELL ###
##############################

X = np.load("./data/smallflat_64.npy")
y = np.load("./data/masked_labels.npy").astype("int")
i = 0
fig = plt.figure(figsize=(18, 18))
for idx in [0, 1, 2, 150, 151, 152]:
    ax = fig.add_subplot(6, 6, i + 1, xticks=[], yticks=[])
    image = (
        np.rot90(X[idx].reshape(64, 64), k=1)
        if idx % 2 == 1 and idx < 150
        else X[idx].reshape(64, 64)
    )
    m_status = "Unmasked" if idx < 150 else "Masked"
    ax.imshow(image, cmap="gray")
    ax.set_title(f"{m_status} Image at i = {idx}")
    i += 1
```



In [9]:
```
##############################
### DO NOT CHANGE THIS CELL ###
##############################
# Use PCA for visualization of masked and unmasked images

X = np.load("./data/smallflat_64.npy")
y = np.load("./data/masked_labels.npy")

fig_title = "Facemask Dataset Visualization with Dimensionality Reduction"
PCA().visualize(X, y, fig_title)


print(
    "*In this plot, the 0 points are unmasked images and the 1 points are masked images."
)
```


Facemask Dataset Visualization with Dimensionality Reduction - 2D PCA Reduction
Feature 2
Feature 1
6
4
2
0
−2
−4
−6
−8
−15
−10
−5
0
5
10
15





*In this plot, the 0 points are unmasked images and the 1 points are masked images.

What do you think of this 2 dimensional plot, knowing that the original dataset was originally a set of flattened image vectors that had 4096 pixels/features?.

1. Examine the 2-dimensional plot of the *facemask* dataset reduced to 2 principal components. How might PCA help in dealing with noise, overfitting, and feature correlation in a high-dimensional dataset? Discuss scenarios where using PCA-reduced data could outperform a classifier trained on the original high-dimensional data. **(2 pts)**

   Answer ...

2. When applying PCA to reduce a high-dimensional dataset (e.g., 4096 features) to just 2 components, how might this impact the classifier's ability to capture non-linear patterns in the data? Discuss how the trade-off between maximizing variance (PCA's goal) and preserving class separability could affect classification performance. **(2 pts)**

   Answer ...

3. When using PCA to reduce a high-dimensional dataset (e.g., 4096 features) to a lower-dimensional representation, it primarily selects components that maximize

variance. Can PCA also be considered a feature selection method? Discuss how PCA can be used for feature selection. **(1 pts)**

Answer ...

2.4 SVD Recommender [2.1% Bonus for All] **[P]**

Introduction:

Principal Component Analysis (PCA) is a dimensionality reduction technique while SVD is commonly used in recommender systems to extract latent factors from sparse data, while another widely used. PCA is often applied to high-dimensional datasets to reduce complexity, remove noise, and improve model performance.

In this section, we aim to introduce SVD in the context of feature selection. While both PCA and SVD methods reduce dimensionality, their underlying principles and applications differ significantly. PCA finds a new set of uncorrelated features (principal components) by maximizing variance, whereas SVD decomposes a matrix into singular vectors and values, which can be interpreted as latent factors in recommendation problems.

Let us try to tackle the famous problem of movie recommendation using just our SVD functions that we have implemented. We are given a table of reviews that 600+ users have provided for close to 10,000 different movies. Our challenge is to predict how much a user would rate a movie that they have not seen (or rated) yet. Once we have these ratings, we would then be able to predict which movies to recommend to that user.

Understanding How SVD Helps in Movie Recommendation

We are given a dataset of user-movie ratings (R) that looks like the following:

UserID/MovieID	1	2	3	4	...	...	8370
1	nan	nan	2	1	...	...	3
2	nan	nan	nan	nan	...	...	nan
3	nan	3	4	nan	...	...	nan
4	1	nan	nan	nan	...	...	5
....	...	...	...	...	...	...	...
....	...	...	...	...	...	...	...
....	...	...	...	...	...	...	...
671	4	nan	nan	nan	...	...	nan

Ratings in the matrix range from 1-5. In addition, the matrix contains `nan` wherever there is no rating provided by the user for the corresponding movie. One simple way to utilize this matrix to predict movie ratings for a given user-movie pair would be to fill in each row / column with the average rating for that row / column. For example: For each movie, if any rating is missing, we could just fill in the average value of all available ratings and expect this to be around the actual / expected rating.

While this may sound like a good approximation, it turns out that by just using SVD we can improve the accuracy of the predicted rating.

How does SVD fit into this picture?

Recall how we previously used SVD to compress images by throwing out less important information. We could apply the same idea to our above matrix (R) to generate another matrix (R_) which will provide the same information, i.e ratings for any user-movie pairs but by combining only the most important features.

Let's look at this with an example:

Assume that decomposition of matrix R looks like:

$R = U\Sigma V^T$

We can re-write this decomposition as follows:

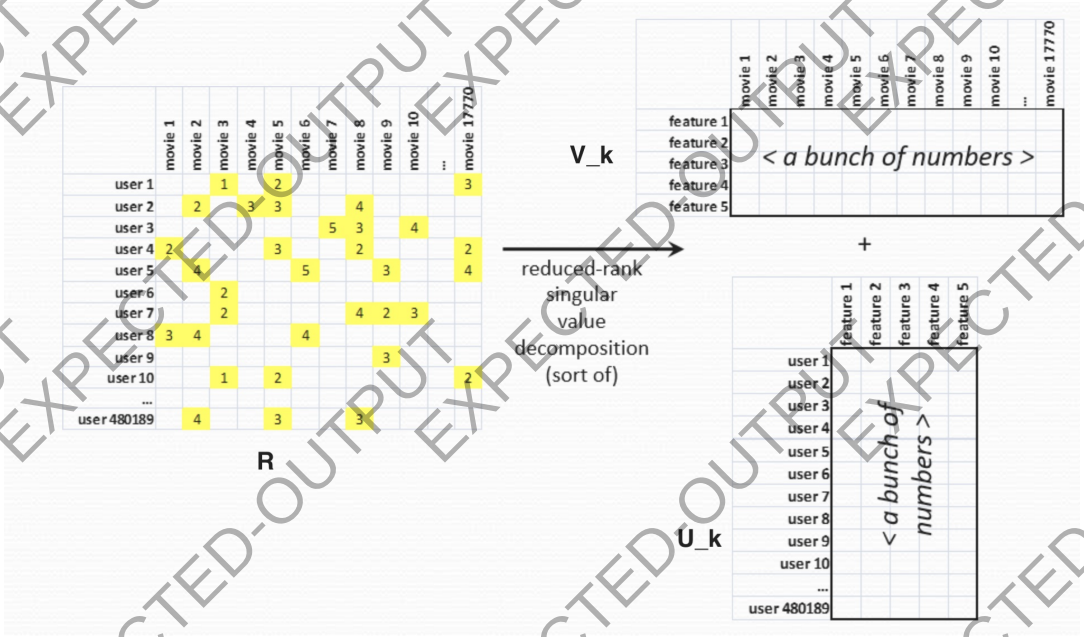
$R = U\sqrt{\Sigma}\sqrt{\Sigma}V^T$

If we were to take only the top K singular values from this matrix, we could again write this as:

$R\_ = U\sqrt{\Sigma_k}\sqrt{\Sigma_k}V^T$

Thus we have now effectively separated our ratings matrix R into two matrices given by: $U_k = U_{[:k]}\sqrt{\Sigma_k}$ and $V_k = \sqrt{\Sigma_k}V_{T[:k]}$

There are many ways to visualize the importance of U and V matrices but with respect to our context of movie ratings, we can visualize these matrices as follows:



We can imagine each row of U_k to be holding some information how much each user likes a particular feature (feature1, feature2, feature 3...feature k). On the contrary, we can imagine each column of V_{Tk} to be holding some information about how much each movie relates to the given features (feature 1, feature 2, feature 3 ... feature k).

Lets denote the row of U_k by u_i and the column of V_{Tk} by m_j. Then the dot-product: $u_i \cdot m_j$ can provide us with information on how much a user *i* likes movie *j*.

What have we achieved by doing this?

Starting with a matrix R containing very few ratings, we have been able to summarize the sparse matrix of ratings into matrices U_k and V_k which each contain feature vectors about the Users and the Movies. Since these feature vectors are summarized from only the most important K features (by our SVD), we can predict any User-Movie rating that is closer to the actual value than just taking any average rating of a row / column (recall our brute force solution discussed above).

Now this method in practice is still not close to the state-of-the-art but for a naive and simple method we have used, we can still build some powerful visualizations as we will see in part 3.

We have divided the task into 3 parts:

1. Implement `recommender_svd` to return matrices U_k and V_k
2. Implement `predict` to predict top 3 movies a given user would watch
3. *(Ungraded)* Feel free to run the final cell labeled to see some visualizations of the feature vectors you have generated

Hint: Movie IDs are IDs assigned to the movies in the dataset and can be greater than the number of movies. This is why we have given movies_index and users_index as well that map between the movie IDs and the indices in the ratings matrix. Please make sure to use this as well.

For a more detailed explanation and practical examples, you can check out this resource:
Singular Value Decomposition (SVD) - GeeksforGeeks

In [10]:
```
##############################
### DO NOT CHANGE THIS CELL ###
##############################

from regression import Regression
from svd_recommender import SVDRecommender
```

In [11]:
```
##############################
### DO NOT CHANGE THIS CELL ###
##############################

recommender = SVDRecommender()
recommender.load_movie_data()
regression = Regression()
# Read the data into the respective train and test dataframes
train, test = recommender.load_ratings_datasets()
print("-----------------------------------------")
print("Train Dataset Stats:")
print("Shape of train dataset: {}".format(train.shape))
print("Number of unique users (train): {}".format(train["userId"].unique().shape[0]))
print("Number of unique users (train): {}".format(train["movieId"].unique().shape[0]))
print("Sample of Train Dataset:")
print("-----------------------------------------")
print(train.head())
print("-----------------------------------------")
print("Test Dataset Stats:")
print("Shape of test dataset: {}".format(test.shape))
print("Number of unique users (test): {}".format(test["userId"].unique().shape[0]))
print("Number of unique users (test): {}".format(test["movieId"].unique().shape[0]))
print("Sample of Test Dataset:")
print("-----------------------------------------")
print(test.head())
print("-----------------------------------------")

# We will first convert our dataframe into a matrix of Ratings: R
# R[i][j] will indicate rating for movie:(j) provided by user:(i)
# users_index, movies_index will store the mapping between array indices and actual userId / movieId
R, users_index, movies_index = recommender.create_ratings_matrix(train)
print("Shape of Ratings Matrix (R): {}".format(R.shape))

# Replacing `nan` with average rating given for the movie by all users
# Additionally, zero-centering the array to perform SVD
mask = np.isnan(R)
masked_array = np.ma.masked_array(R, mask)
r_means = np.array(np.mean(masked_array, axis=0))
R_filled = masked_array.filled(r_means)
```

```
R_filled = R_filled - r_means
```

```
-----------------------------------------
Train Dataset Stats:
Shape of train dataset: (88940, 4)
Number of unique users (train): 671
Number of unique users (train): 8370
Sample of Train Dataset:
-----------------------------------------
   userId  movieId  rating   timestamp
0       1     2294     2.0  1260759108
1       1     2455     2.5  1260759113
2       1     3671     3.0  1260759117
3       1     1339     3.5  1260759125
4       1     1343     2.0  1260759131
-----------------------------------------
Test Dataset Stats:
Shape of test dataset: (10393, 4)
Number of unique users (test): 671
Number of unique users (test): 4368
Sample of Test Dataset:
-----------------------------------------
   userId  movieId  rating   timestamp
0       1     2968     1.0  1260759200
1       1     1405     1.0  1260759203
2       1     1172     4.0  1260759205
3       2       52     3.0   835356031
4       2      314     4.0   835356044
-----------------------------------------
Shape of Ratings Matrix (R): (671, 8370)
```

2.4.1 SVD Recommender for Movies [1.4% Bonus for All] **[P]**

In **svd_recommender.py** file, complete the following function:

- **recommender_svd**: Use the above equations to output U_k and V_k. You can utilize the `svd` method from Numpy and `compress` method from `eigenfaces.py` to retrieve your initial U, Σ and V matrices. Then, calculate U_k and V_k based on the decomposition example above.

- **predict**: Predict the next 3 movies (sorted by high to low rating) that the user would be most interested in watching among the ones above.

Our goal here is to predict movies that a user would be interested in watching next. Since our dataset contains a large list of movies and our model is very naive, filtering among this huge set for top 3 movies can produce results that we may not correlate immediately. Therefore, we'll restrict this prediction to only movies among a subset as given by movies_pool.

Let us consider a user (ID: 660) who has already watched and rated well (>3) on the following movies:

- Iron Man (2008)
- Thor: The Dark World (2013)
- Avengers, The (2012)

The following cell tries to predict which among the movies given by the list below, the user would be most interested in watching next:
`movies_pool`:

- Ant-Man (2015)
- Iron Man 2 (2010)
- Avengers: Age of Ultron (2015)

- Thor (2011)
- Captain America: The First Avenger (2011)
- Man of Steel (2013)
- Star Wars: Episode IV - A New Hope (1977)
- Ladybird Ladybird (1994)
- Man of the House (1995)
- Jungle Book, The (1994)

HINT: You can use the method `get_movie_id_by_name` to convert movie names into movie IDs and vice-versa.

NOTE: The user may have already watched and rated some of the movies in `movies_pool`. Remember to filter these out before returning the output. The original Ratings Matrix, R might come in handy here along with `np.isnan`

2.4.2 Local Test for `recommender_svd` Function [No Points]

You may test your implementation of the function in the cell below. See Using the Local Tests for more details.

In [12]:
```
##############################
### DO NOT CHANGE THIS CELL ###
##############################

from utilities.localtests import TestSVDRecommender

unittest_svd_rec = TestSVDRecommender()
unittest_svd_rec.test_recommender_svd()
```

```
UnitTest passed successfully for "recommender_svd() function"!
```

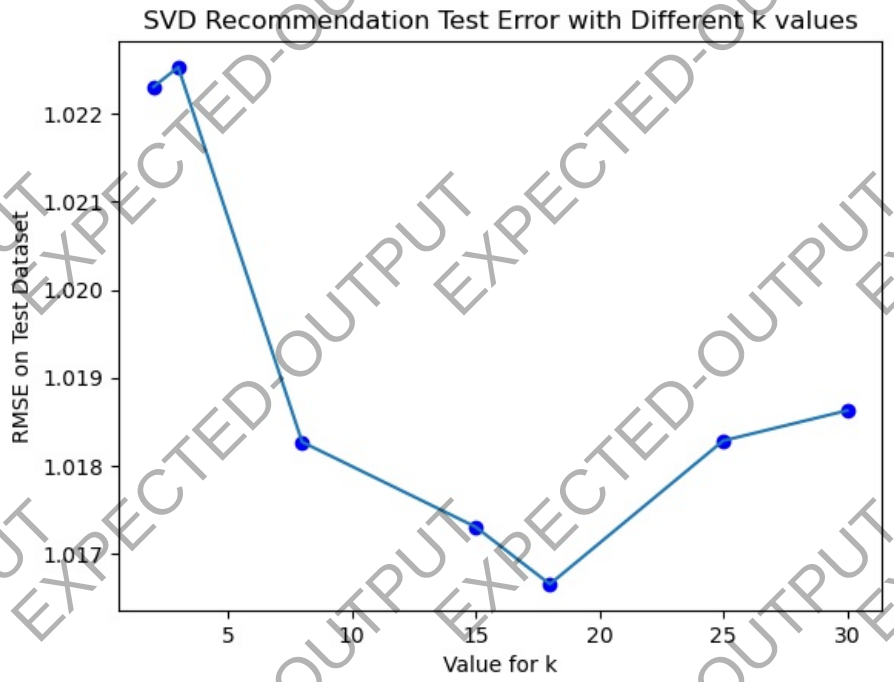
In [13]:
```
##############################
### DO NOT CHANGE THIS CELL ###
##############################

# Implement the method `recommender_svd` and run it for the following values of features
no_of_features = [2, 3, 8, 15, 18, 25, 30]
test_errors = []

for k in no_of_features:
    U_k, V_k = recommender.recommender_svd(R_filled, k)
    pred = []  # to store the predicted ratings
    for _, row in test.iterrows():
        user = row["userId"]
        movie = row["movieId"]
        u_index = users_index[user]
        # If we have a prediction for this movie, use that
        if movie in movies_index:
            m_index = movies_index[movie]
            pred_rating = np.dot(U_k[u_index, :], V_k[:, m_index]) + r_means[m_index]
        # Else, use an average of the users ratings
        else:
            pred_rating = np.mean(np.dot(U_k[u_index], V_k)) + r_means[m_index]
        pred.append(pred_rating)
    test_error = regression.rmse(test["rating"], pred)
    test_errors.append(test_error)
    print("RMSE for k = {} --> {}".format(k, test_error))
```

```
RMSE for k = 2 --> 1.0223035413708281
RMSE for k = 3 --> 1.0225266494179552
```

```
RMSE for k = 8 --> 1.0182709203352787
RMSE for k = 15 --> 1.017307118738714
RMSE for k = 18 --> 1.0166562048687975
RMSE for k = 25 --> 1.0182856984912254
RMSE for k = 30 --> 1.0186282488126601
```

Plot the Test Error over the different values of `k`

In [14]:
```
##############################
### DO NOT CHANGE THIS CELL ###
##############################
fig = plt.figure()
plt.plot(no_of_features, test_errors, "bo")
plt.plot(no_of_features, test_errors)
plt.xlabel("Value for k")
plt.ylabel("RMSE on Test Dataset")
plt.title("SVD Recommendation Test Error with Different k values")

plt.show()
```



2.4.3 Local Test for `predict` Functions [No Points]

You may test your implementation of the function in the cell below. See Using the Local Tests for more details.

In [15]:
```
##############################
### DO NOT CHANGE THIS CELL ###
##############################

unittest_svd_rec.test_predict()
```

```
Top 3 Movies the User would want to watch:
Captain America: The First Avenger (2011)
Ant-Man (2015)
Avengers: Age of Ultron (2015)
------------------------------------------------------------
UnitTest passed successfully for "predict() function"!
```

2.4.4 Comparison: SVD vs. PCA [0.70% pt]

Both PCA and SVD use the same mathematical backbone, but are used for different tasks.

Consider:

- When reconstructing the movie rating matrix with SVD, we did not center the matrix.
- When calculating principal components for the facemask images and eigenfaces for the face images, we did center the matrices.

Why does PCA require centering the data but SVD does not?

Q3 Polynomial regression and regularization [72pts: 52pts + 2.3% Bonus for All + 20pts Grad / 6% Bonus for Undergrads] **[P]** | **[W]**

3.1 About RMSE (Root Mean Square Error) [3 pts] **[W]**

Mean Squared Error (MSE) is used to provide an indication of how well a model is performing in terms of prediction accuracy.

$$MSE = \frac{1}{n}\sum_{i=1}(y_i - \hat{y}_i)^2$$

where:

- n is the number of data points
- y_i is the actual target value
- $\hat{y}_i$ is the predicted value by the model

However, Root Mean Square Error (RMSE) is preferred over MSE because it brings the error back to the same units as the target variable, providing easier comparison to the actual values.

This helps in providing better practical interpretation of how well the model is performing on both small and large errors.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}(y_i - \hat{y}_i)^2}$$

What is a good RMSE value?

If we normalize our labels such that the true labels y and the model outputs $\hat{y}$ can only be between 0 and 1, what does it mean when the RMSE = 1? Please provide a toy example with your explanation.

Answer:

3.2 Regression and regularization implementations [50pts: 30pts + 20pts Grad / 6% Bonus for Undergrad] **[P]**

We have three methods to fit linear and ridge regression models: 1) closed form solution; 2) gradient descent (GD); 3) stochastic gradient descent (SGD). Some of the functions are bonus, see the below function list on what is required to be implemented for graduate and undergraduate students. We use the term weight in the following code. Weights and parameters (θ) have the same meaning here. We used parameters (θ) in the lecture slides.

In the **regression.py** file, complete the Regression class by implementing the listed functions below. We have provided the Loss function, L, associated with the GD and SGD function for Linear and Ridge Regression for deriving the gradient update.

- **rmse**
- **construct_polynomial_feats**
- **predict**
- **linear_fit_closed**: You should use `np.linalg.pinv` in this function
- **linear_fit_GD** (bonus for undergrad, **required for grad**): $L_{linear,\ GD}(\theta) = {}^{1}2N\sum_{i=0}^{N}[y_i - \hat{}y_i(\theta)]^2$ y_i = label, $\hat{}y_i(\theta)$ = prediction
- **linear_fit_SGD** (bonus for undergrad, **required for grad**): $L_{linear,\ SGD}(\theta) = {}^{1}2[y_i - \hat{}y_i(\theta)]^2$ y_i = label, $\hat{}y_i(\theta)$ = prediction
- **ridge_fit_closed**: You should adjust your I matrix to handle the bias term differently than the rest of the terms
- **ridge_fit_GD** (bonus for undergrad, **required for grad**): $L_{ridge,\ GD}(\theta) = L_{linear,\ GD}(\theta) + {}^{c}\lambda 2N\theta^T\theta$
- **ridge_fit_SGD** (bonus for undergrad, **required for grad**):

$L_{ridge,\ SGD}(\theta) = L_{linear,\ SGD}(\theta) + {}^{c}\lambda 2N\theta^T\theta$

- **ridge_cross_validation**: Use `ridge_fit_closed` for this function

IMPORTANT NOTE:

- Use your RMSE function to calculate actual loss when coding GD and SGD, but use the loss listed above to derive the gradient update.
- In **ridge_fit_GD** and **ridge_fit_SGD**, you should avoid applying regularization to the bias term in the gradient update.

The points for each function is in the **Deliverables and Points Distribution** section.

3.2.1 Local Tests for Helper Regression Functions [No Points]

You may test your implementation of the functions contained in **regression.py** in the cell below. Feel free to comment out tests for functions that have not been completed yet. See Using the Local Tests for more details.

In [16]:
```
################################
### DO NOT CHANGE THIS CELL ###
################################

from utilities.localtests import TestRegression

unittest_reg = TestRegression()
unittest_reg.test_rmse()
unittest_reg.test_construct_polynomial_feats()
```

```
unittest_reg.test_predict()
```

```
UnitTest passed successfully for "RMSE"!
UnitTest passed successfully for "Polynomial feature construction"!
UnitTest passed successfully for "Linear regression prediction"!
```

3.2.2 Local Tests for Linear Regression Functions [No Points]

You may test your implementation of the functions contained in **regression.py** in the cell below. Feel free to comment out tests for functions that have not been completed yet. See Using the Local Tests for more details.

In [17]:
```
##############################
### DO NOT CHANGE THIS CELL ###
##############################

from utilities.localtests import TestRegression

unittest_reg = TestRegression()
unittest_reg.test_linear_fit_closed()
```

```
UnitTest passed successfully for "Closed form linear regression"!
```

3.2.3 Local Tests for Ridge Regression Functions [No Points]

You may test your implementation of the functions contained in **regression.py** in the cell below. Feel free to comment out tests for functions that have not been completed yet. See Using the Local Tests for more details.

In [18]:
```
##############################
### DO NOT CHANGE THIS CELL ###
##############################

from utilities.localtests import TestRegression

unittest_reg = TestRegression()
unittest_reg.test_ridge_fit_closed()
unittest_reg.test_ridge_cross_validation()
```

```
UnitTest passed successfully for "Closed form ridge regression"!
UnitTest passed successfully for "Ridge regression cross validation"!
```

3.2.4 Local Tests for Gradient Descent and SGD (Bonus for Undergrad Tests) [No Points]

You may test your implementation of the functions contained in **regression.py** in the cell below. Feel free to comment out tests for functions that have not been completed yet. See Using the Local Tests for more details.

In [19]:
```
##############################
### DO NOT CHANGE THIS CELL ###
##############################

from utilities.localtests import TestRegression

unittest_reg = TestRegression()
unittest_reg.test_linear_fit_GD()
unittest_reg.test_linear_fit_SGD()
unittest_reg.test_ridge_fit_GD()
unittest_reg.test_ridge_fit_SGD()
```

```
UnitTest passed successfully for "Gradient descent linear regression"!
UnitTest passed successfully for "Stochastic gradient descent linear regression"!
UnitTest passed successfully for "Gradient descent ridge regression"!
UnitTest passed successfully for "Stochastic gradient descent ridge regression"!
```

3.3 Testing: General Functions and Linear Regression [5 pts] **[P]**

In this section. we will test the performance of the linear regression. As long as your test RMSE score is close to the TA's answer (TA's answer ± 0.05), you can get full points. Let's first construct a dataset for polynomial regression.

In this case, we construct the polynomial features up to degree 5. Each data sample consists of two features [a, b]. We compute the polynomial features of both a and b in order to yield the vectors $[1, a, a^2, a^3, ..., a^{degree}]$ and $[1, b, b^2, b^3, ..., b^{degree}]$. We train our model with the cartesian product of these polynomial features. The cartesian product generates a new feature vector consisting of all polynomial combinations of the features with degree less than or equal to the specified degree.

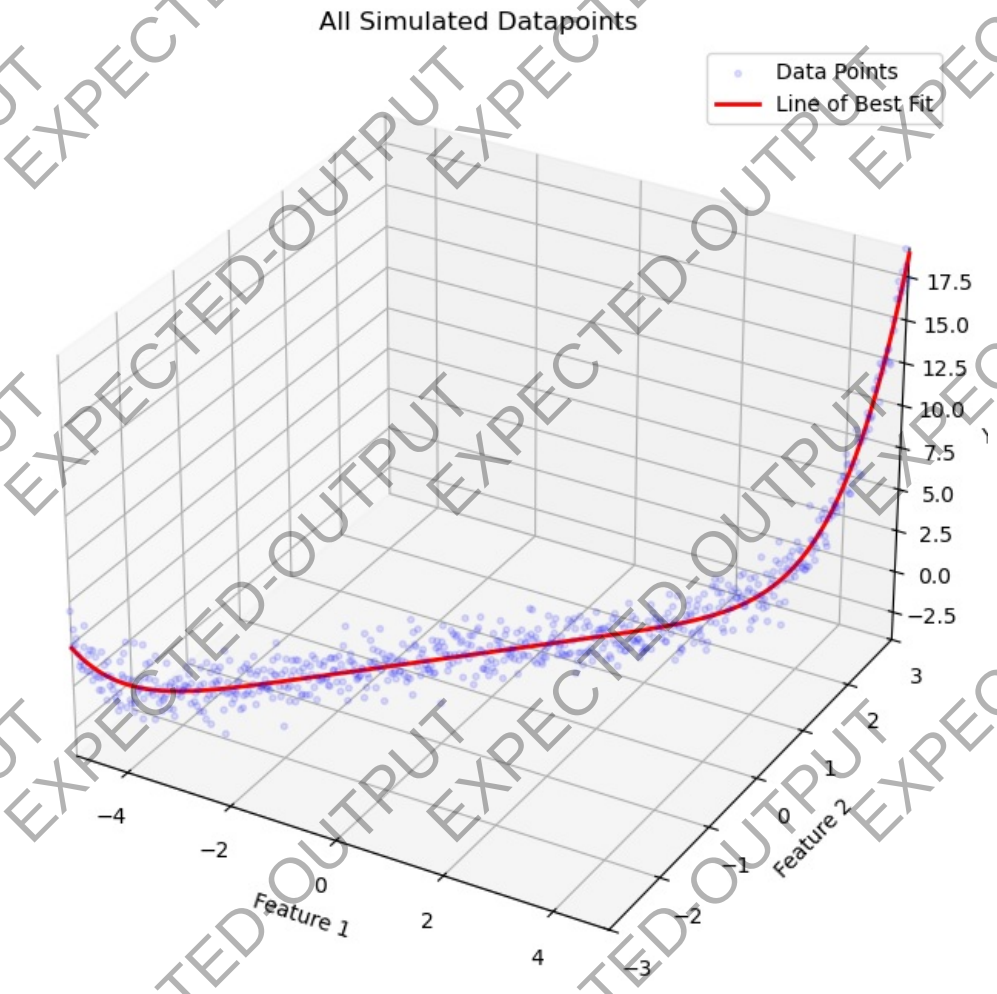
For example, if degree = 2, we will have the polynomial features $[1, a, a^2]$ and $[1, b, b^2]$ for the datapoint [a, b]. The cartesian product of these two vectors will be $[1, a, b, ab, a^2, b^2]$. We do not generate a^3 and b^3 since their degree is greater than 2 (specified degree).

In [20]:
```
###############################
### DO NOT CHANGE THIS CELL ###
###############################

from plotter import Plotter
from regression import Regression
```

In [21]:
```
###############################
### DO NOT CHANGE THIS CELL ###
###############################

# Generate a sample regression dataset with polynomial features
# using the student's regression implementation.

POLY_DEGREE = 5

reg = Regression()
plotter = Plotter(regularization=reg, poly_degree=POLY_DEGREE)

x_all, y_all, p, x_all_feat = plotter.create_data()
```

```
x_all: 700 (rows/samples) 2 (columns/features)
y_all: 700 (rows/samples) 1 (columns/features)
```

In [22]:
```
###############################
### DO NOT CHANGE THIS CELL ###
###############################

# Visualize simulated regression data

plotter.plot_all_data(x_all, y_all, p)
```

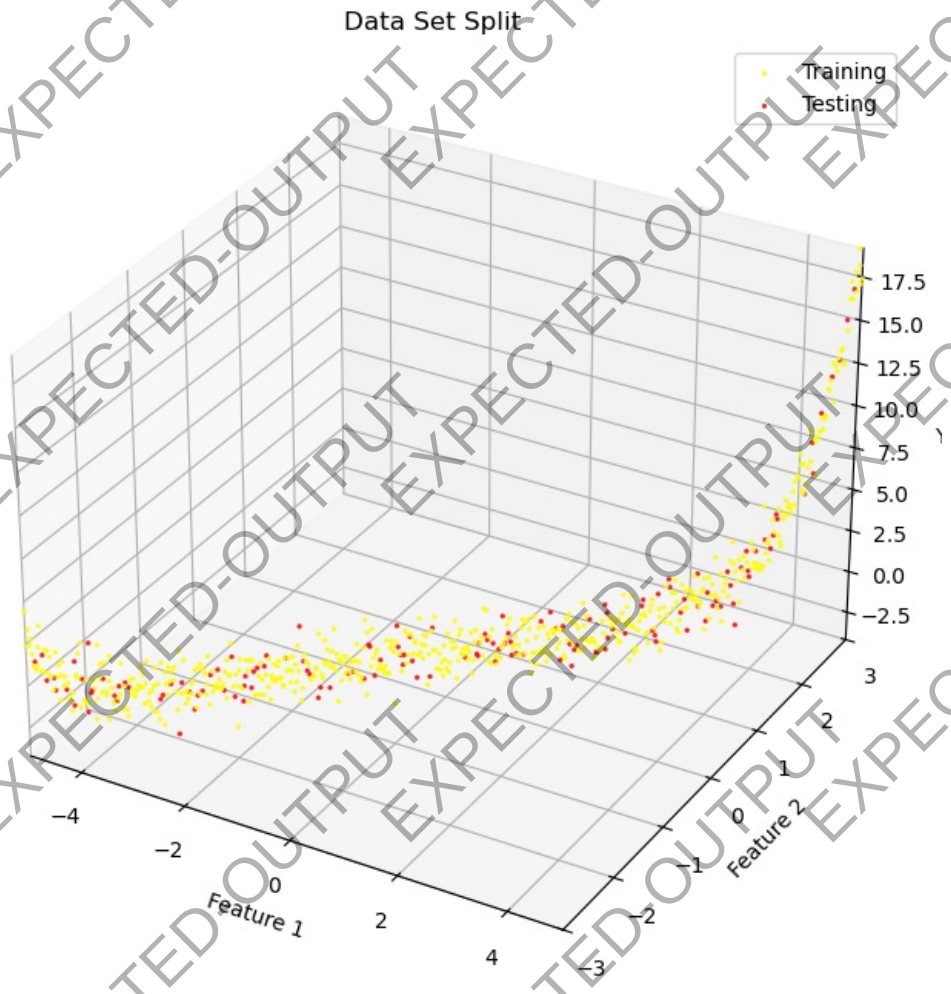


In the figure above, the red curve is the true fuction we want to learn, while the blue dots are the noisy data points. The data points are generated by $Y = X\theta + \epsilon$, where $\epsilon_i \sim N(0, 1)$ are i.i.d. generated noise.

Now let's split the data into two parts, the training set and testing set. The yellow dots are for training, while the red dots are for testing.

In [23]:

```
################################
### DO NOT CHANGE THIS CELL ###
################################

xtrain, ytrain, xtest, ytest, train_indices, test_indices = plotter.split_data(
    x_all, y_all
)

plotter.plot_split_data(xtrain, xtest, ytrain, ytest)
```

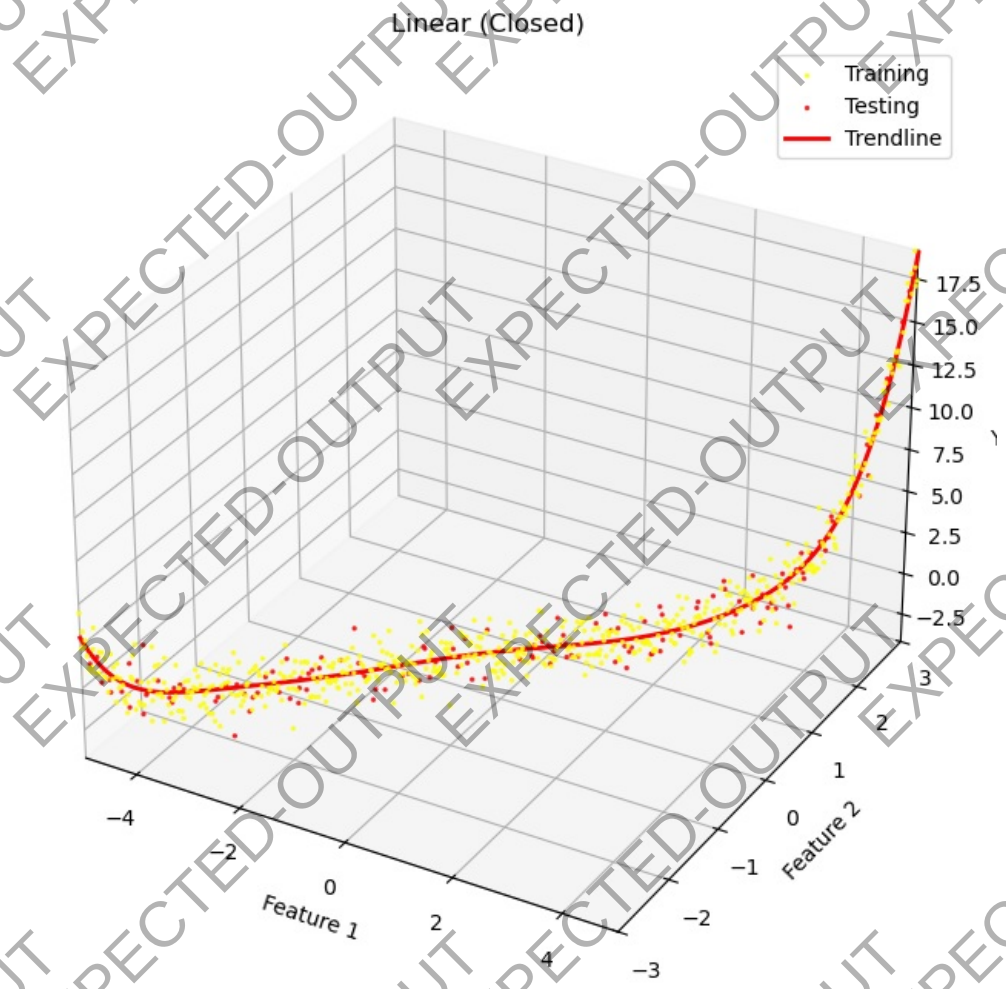


Now let us train our model using the training set and see how our model performs on the testing set. Observe the red line, which is our model's learned function.

In [24]:

```
################################
### DO NOT CHANGE THIS CELL ###
################################

# Required for both Grad and Undergrad

weight = reg.linear_fit_closed(x_all_feat[train_indices], y_all[train_indices])
y_test_pred = reg.predict(x_all_feat[test_indices], weight)
test_rmse = reg.rmse(y_test_pred, y_all[test_indices])
y_pred = reg.predict(x_all_feat, weight)
print("Linear (closed) RMSE: %.4f" % test_rmse)

plotter.plot_linear_closed(xtrain, xtest, ytrain, ytest, x_all, y_pred)
```

```
Linear (closed) RMSE: 1.0273
```



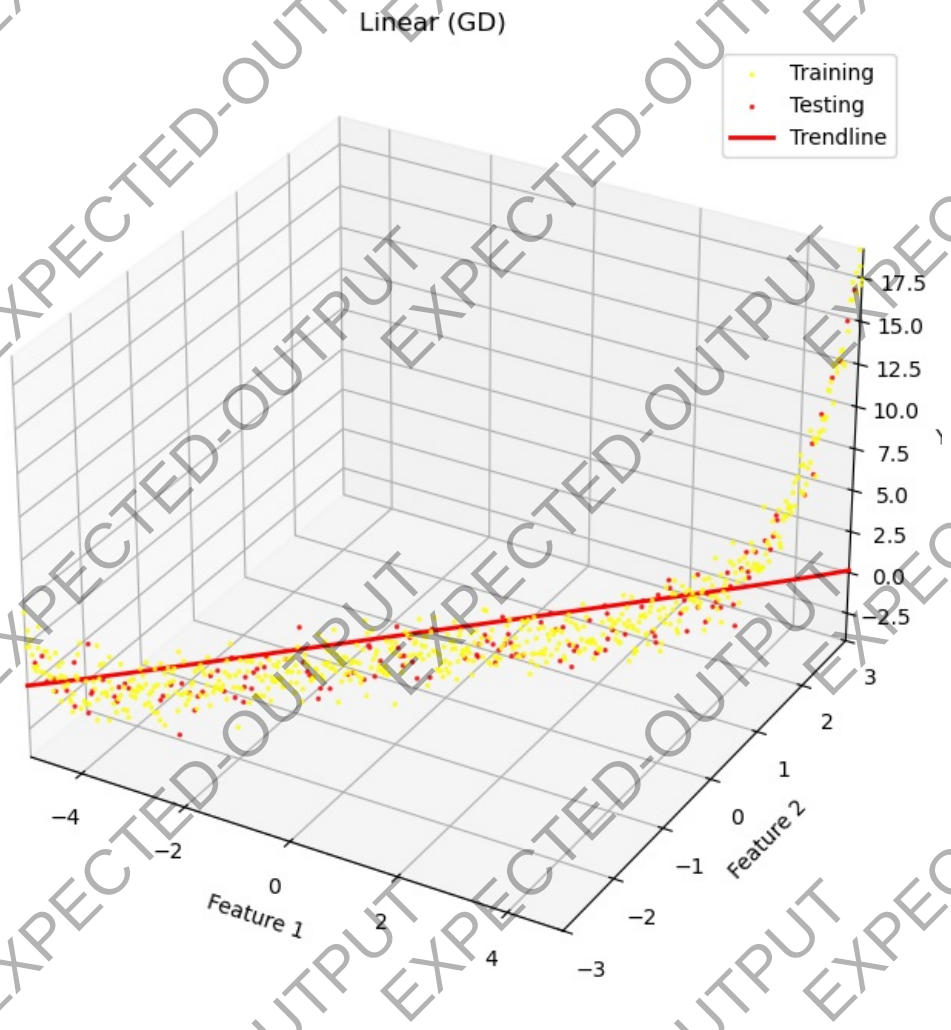
HINT: If your RMSE is off, make sure to follow the instruction given for `linear_fit_closed` in the list of functions to implement above.

Now let's use our linear gradient descent function with the same setup. Observe that the trendline is now less optimal, and our RMSE increased. Do not be alarmed.

In [25]:
```
################################
### DO NOT CHANGE THIS CELL ###
################################

# Required for Grad Only
# This cell may take more than 1 minute
weight, _ = reg.linear_fit_GD(
    x_all_feat[train_indices], y_all[train_indices], epochs=50000, learning_rate=1e-8
)
y_test_pred = reg.predict(x_all_feat[test_indices], weight)
test_rmse = reg.rmse(y_test_pred, y_all[test_indices])
print("Linear (GD) RMSE: %.4f" % test_rmse)
```

```
y_pred = reg.predict(x_all_feat, weight)
y_pred = np.reshape(y_pred, (y_pred.size,))

plotter.plot_linear_gd(xtrain, xtest, ytrain, ytest, x_all, y_pred)
```

```
Linear (GD) RMSE: 3.1861
```



We must tune our epochs and learning_rate. As we tune these parameters our trendline will approach the trendline generated by the linear closed form solution. Observe how we slowly tune (increase) the epochs and learning_rate below to create a better model.

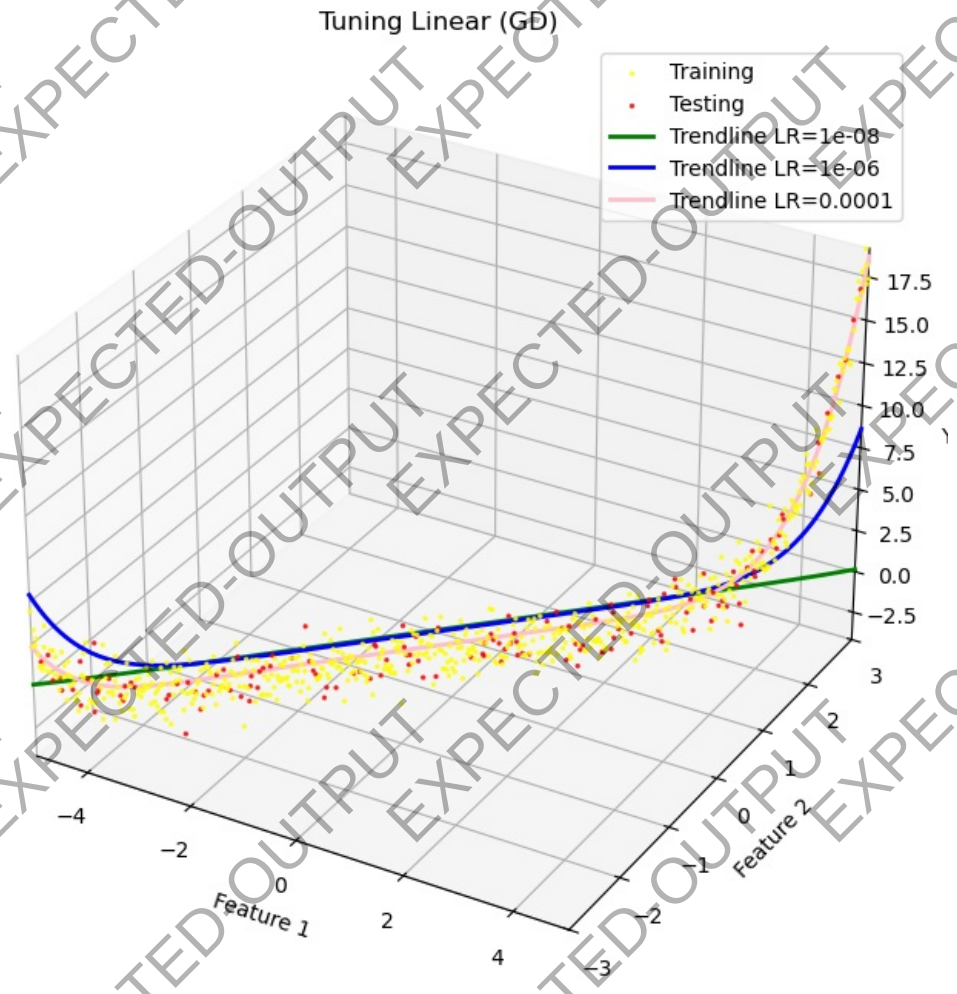
Note that the closed form solution will always give the most optimal/overfit results. We cannot outperform the closed form solution with GD. We can only approach closed forms level of optimality/overfitness. We leave the reasoning behind this as an exercise to the reader.

In [26]:

```
##############################
### DO NOT CHANGE THIS CELL ###
##############################

# Required for Grad Only
# This cell may take more than 1 minute

learning_rates = [1e-8, 1e-6, 1e-4]
```

```
weights = np.zeros((3, POLY_DEGREE**2 + 2))

for ii in range(len(learning_rates)):
    weights[ii, :] = reg.linear_fit_GD(
        x_all_feat[train_indices],
        y_all[train_indices],
        epochs=50000,
        learning_rate=learning_rates[ii],
    )[0].ravel()
    y_test_pred = reg.predict(
        x_all_feat[test_indices], weights[ii, :].reshape((POLY_DEGREE**2 + 2, 1))
    )
    test_rmse = reg.rmse(y_test_pred, y_all[test_indices])
    print("Linear (GD) RMSE: %.4f (learning_rate=%s)" % (test_rmse, learning_rates[ii]))

plotter.plot_linear_gd_tuninglr(
    xtrain, xtest, ytrain, ytest, x_all, x_all_feat, learning_rates, weights
)
```

```
Linear (GD) RMSE: 3.1861 (learning_rate=1e-08)
Linear (GD) RMSE: 2.2901 (learning_rate=1e-06)
Linear (GD) RMSE: 1.1099 (learning_rate=0.0001)
```



And what if we just use the first 10 data points to train?

Linear Closed 10 Samples

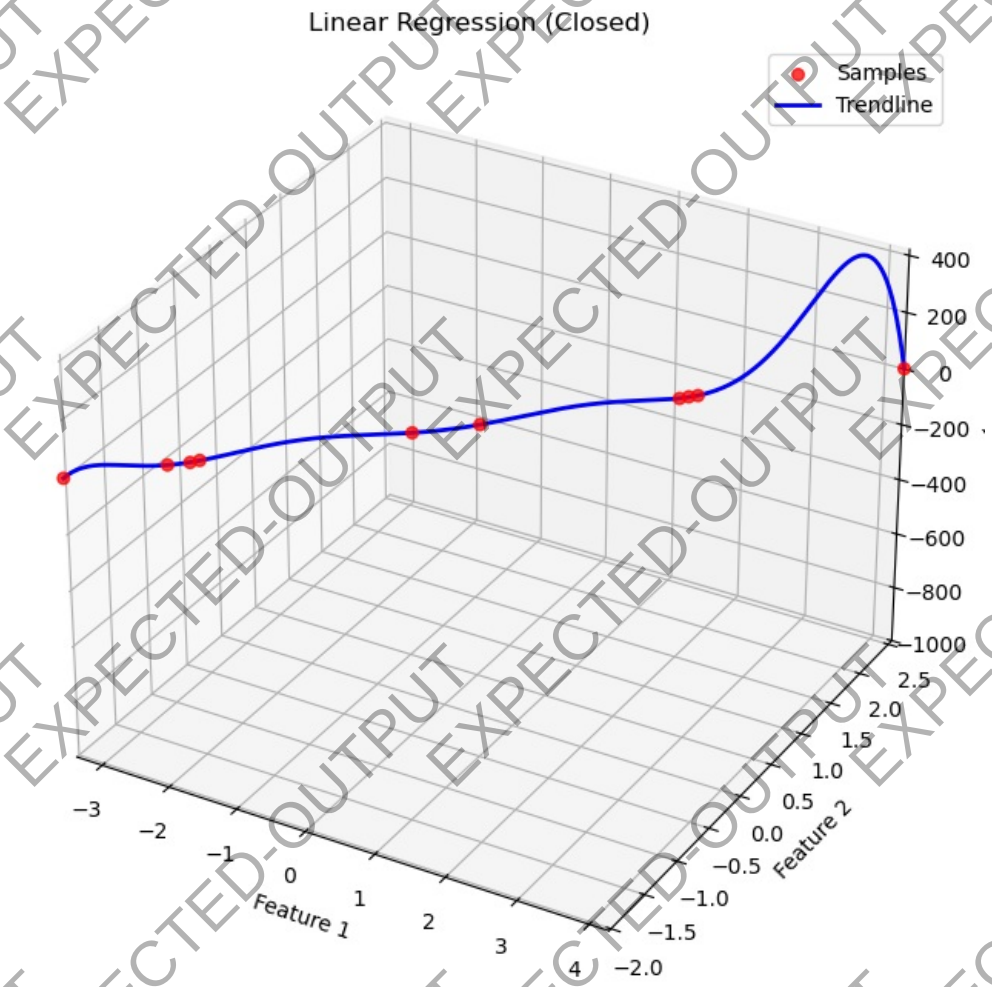
In [27]:
```
################################
### DO NOT CHANGE THIS CELL ###
################################
rng = np.random.RandomState(seed=3)
y_all_noisy = np.dot(x_all_feat, np.zeros((POLY_DEGREE**2 + 2, 1))) + rng.randn(
    x_all_feat.shape[0], 1
)
sub_train = train_indices[10:20]
```

Due to the large RMSE values, rounding errors may result in larger than normal differences from the TA solution. Here, we will accept RMSE values ± 0.1 from the TA solution.

In [28]:
```
################################
### DO NOT CHANGE THIS CELL ###
```

```
##############################

# Required for both Grad and Undergrad

weight = reg.linear_fit_closed(x_all_feat[sub_train], y_all_noisy[sub_train])
y_pred = reg.predict(x_all_feat, weight)
y_test_pred = reg.predict(x_all_feat[test_indices], weight)
test_rmse = reg.rmse(y_test_pred, y_all_noisy[test_indices])
print("Linear (closed) 10 Samples RMSE: %.4f" % test_rmse)

plotter.plot_10_samples(
    x_all, y_all_noisy, sub_train, y_pred, title="Linear Regression (Closed)"
)
```

```
Linear (closed) 10 Samples RMSE: 1816.3828
```



Did you see a worse performance? Let's take a closer look at what we have learned.

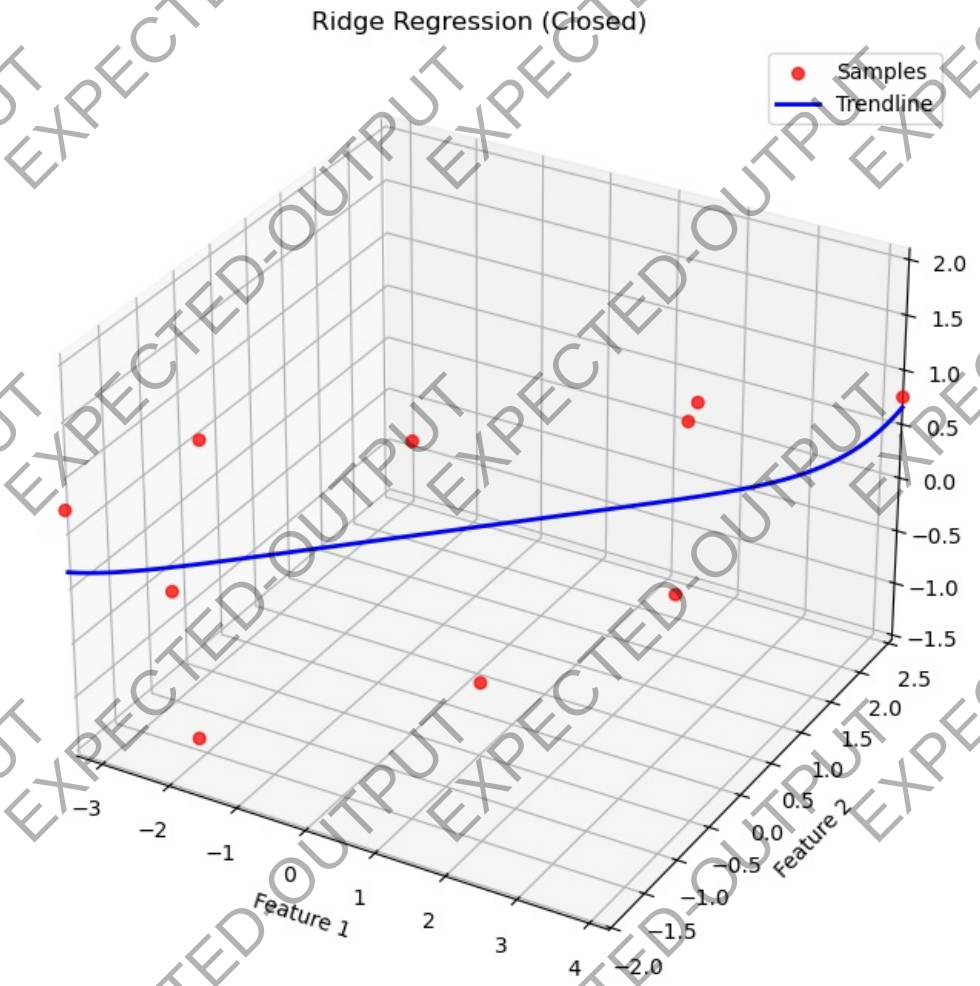
3.4 Testing: Ridge Regression [5 pts] **[P]**

Again, let's see what we have learned. **You only need to run the cell corresponding to your specific implementation.**

Note: The content covered in these cells are going to be autograded separately. The figure outputs of these cells should match the expected outputs, but the figures themselves will not be graded. This section acts as a local test, and your grade will be determined by the test cases in the autograder.

In [29]:
```
################################
### DO NOT CHANGE THIS CELL ###
################################
# Required for both Grad and Undergrad
weight = reg.ridge_fit_closed(
    x_all_feat[sub_train], y_all_noisy[sub_train], c_lambda=10
)
y_pred = reg.predict(x_all_feat, weight)
y_test_pred = reg.predict(x_all_feat[test_indices], weight)
test_rmse = reg.rmse(y_test_pred, y_all_noisy[test_indices])
print("Ridge Regression (closed) RMSE: %.4f" % test_rmse)

plotter.plot_10_samples(
    x_all, y_all_noisy, sub_train, y_pred, title="Ridge Regression (Closed)"
)
```

```
Ridge Regression (closed) RMSE: 1.1193
```

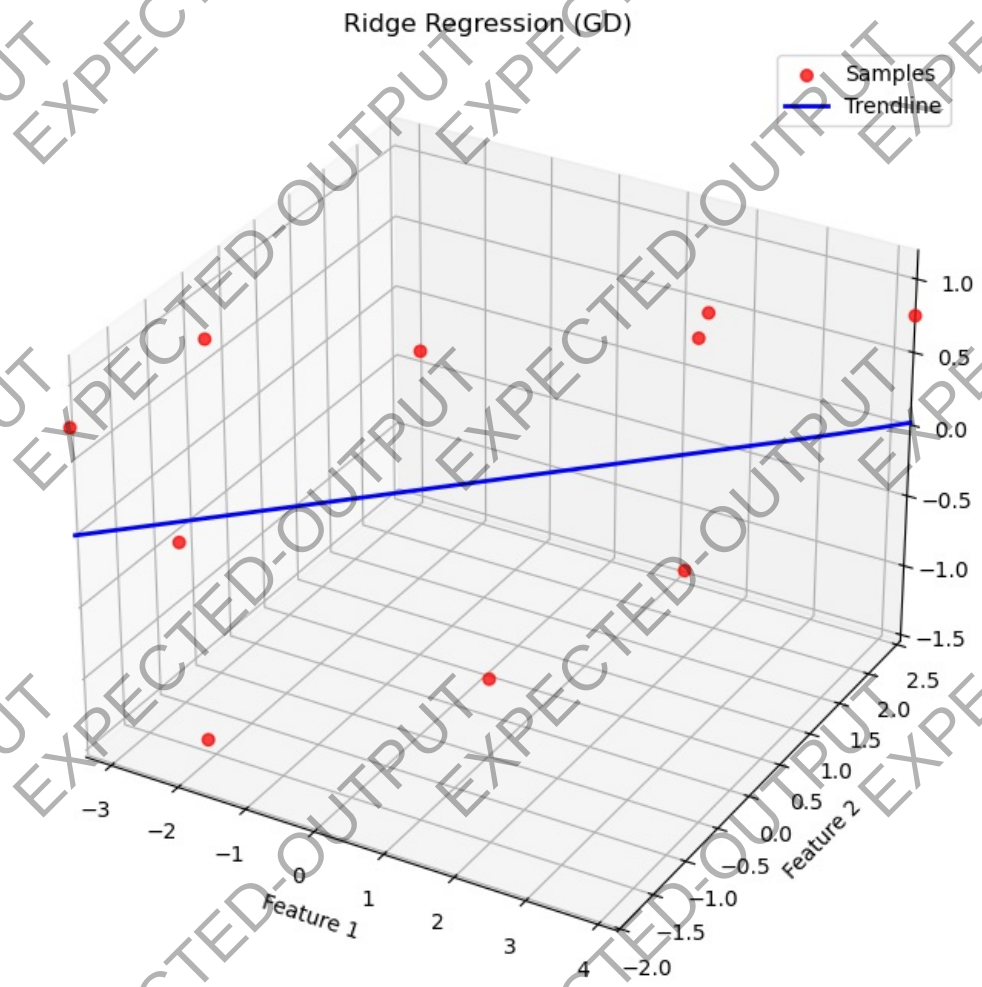


HINT: Make sure to follow the instruction given for `ridge_fit_closed` in the list of functions to implement above.

In [30]:

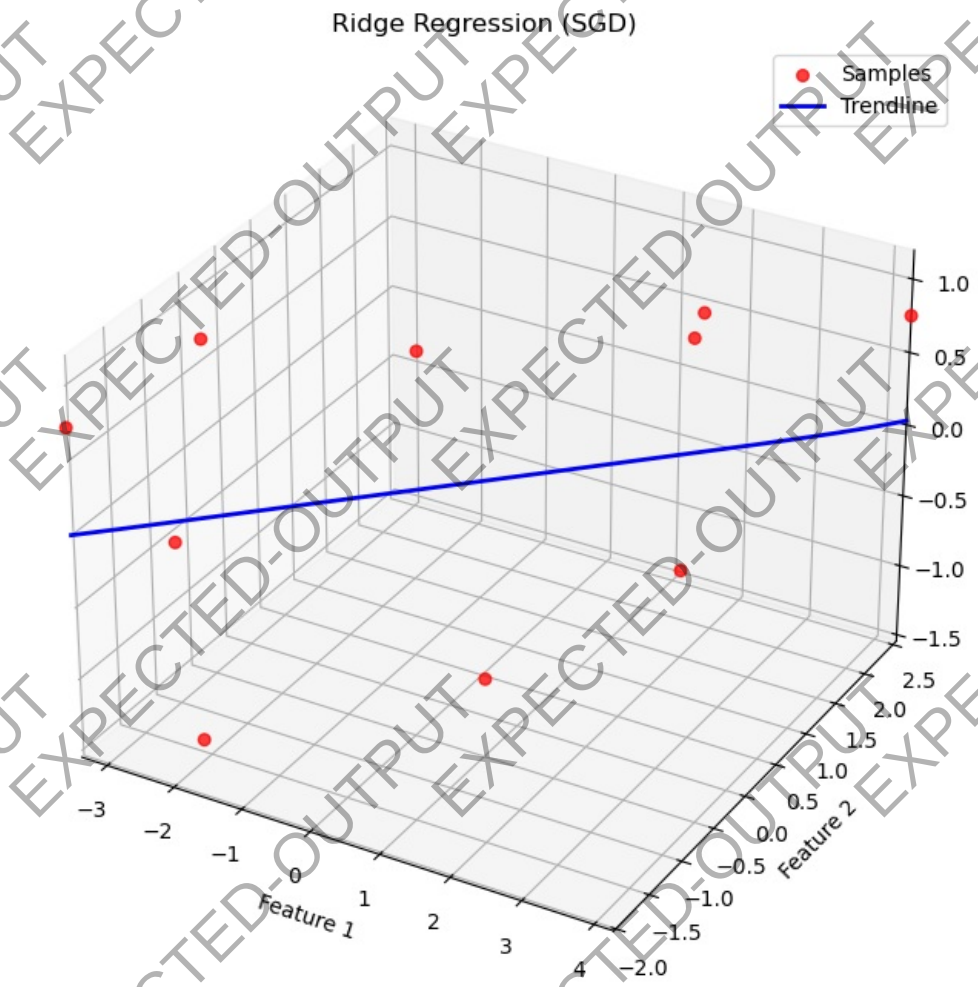
```
################################
### DO NOT CHANGE THIS CELL ###
################################

# Required for Grad Only

weight, _ = reg.ridge_fit_GD(
    x_all_feat[sub_train], y_all_noisy[sub_train], c_lambda=20, learning_rate=1e-5
)
y_pred = reg.predict(x_all_feat, weight)
y_test_pred = reg.predict(x_all_feat[test_indices], weight)
test_rmse = reg.rmse(y_test_pred, y_all_noisy[test_indices])
print("Ridge Regression (GD) RMSE: %.4f" % test_rmse)

plotter.plot_10_samples(
    x_all, y_all_noisy, sub_train, y_pred, title="Ridge Regression (GD)"
)
```

```
Ridge Regression (GD) RMSE: 1.0413
```



In [31]:
```
##############################
### DO NOT CHANGE THIS CELL ###
##############################

# Required for Grad Only

weight, _ = reg.ridge_fit_SGD(
    x_all_feat[sub_train], y_all_noisy[sub_train], c_lambda=20, learning_rate=1e-5
)
y_pred = reg.predict(x_all_feat, weight)
y_test_pred = reg.predict(x_all_feat[test_indices], weight)
test_rmse = reg.rmse(y_test_pred, y_all_noisy[test_indices])
print("Ridge Regression (SGD) RMSE: %.4f" % test_rmse)

plotter.plot_10_samples(
    x_all, y_all_noisy, sub_train, y_pred, title="Ridge Regression (SGD)"
)
```

```
Ridge Regression (SGD) RMSE: 1.0410
```



3.5 Linear vs. Ridge Regression Analysis [4pts] **[W]**

Analyze the difference in performance between the linear and ridge regression methods given the output RMSE **from the testing on 10 samples** and their corresponding approximation plots.

1. Why does ridge regression achieve a lower RMSE than linear regression on 10 sample points? **(1pts)**
2. Describe and contrast two scenarios (real life applications): One where linear is more suitable than ridge, and one in which ridge is better choice than linear. Explain why. **(1 pts)**
3. What is the impact of having some highly correlated features on the data set in terms of linear algebra? Mathematically explain (include expressions) how ridge has an advantage on this in comparison to linear regression. Include the idea of numerical stability. **(2pts)**

Hint: Think about the closed form solution for the weights

1. **Answer** ...

2. **Answer** ...

3. **Answer** ...

3.6 Cross Validation Hyperparameter Search [5 pts] **[P]**

Let's use Cross Validation to search for the best value for `c_lambda` in ridge regression.

Imagine we have a dataset of 10 points [1,2,3,4,5,6,7,8,9,10] and we want to do 5-fold cross validation.

- The first iteration we would train with [3,4,5,6,7,8,9,10] and test (validate) with [1,2]
- The second iteration we would train with [1,2,5,6,7,8,9,10] and test (validate) with [3,4]
- The third iteration we would train with [1,2,3,4,7,8,9,10] and test (validate) with [5,6]
- The fourth iteration we would train with [1,2,3,4,5,6,9,10] and test (validate) with [7,8]
- The fifth iteration we would train with [1,2,3,4,5,6,7,8] and test (validate) with [9,10]

We provided a list of possible values for λ, and you will complete the `ridge_cross_validation` method to perform 5-fold cross-validation on the training data (we already use `train_indices` to get training data in the cell below). Split the training data into 5 folds, where 20 percent of the data will be used to test and 80 percent will be used to train. For each λ, you will have calculated 5 RMSE values. We provide a function `hyperparameter_search` that takes the average of the RMSE values for each λ and picks the λ with the lowest mean RMSE. (Please look at hints for more information),

HINTS:

- `np.concatenate` is your friend
- Make sure to follow the instruction given for `ridge_fit_closed` in the list of functions to implement above.
- To use the 5-fold method, loop over all the data 5 times, where we split a different 20% of the data at every iteration. The first iteration extracts the first 20% for testing and the remaining 80% for training. The second iteration splits the second 20% of data for testing and the (different) remaining 80% for testing. If we have the array of elements 1 - 10, the second iteration would extract the numbers "3" and "4" because that's in the second 20% of the array.
- The `hyperparameter_search` function will handle averaging the errors, so don't average the errors in `ridge_cross_validation`. We've done this so you can see your error across every fold when using the gradescope tests.

In [32]:
```
###############################
### DO NOT CHANGE THIS CELL ###
###############################
lambda_list = [0.0001, 0.001, 0.1, 1, 5, 10, 50, 100, 1000, 10000]
kfold = 5

best_lambda, best_error, error_list = reg.hyperparameter_search(
    x_all_feat[train_indices], y_all[train_indices], lambda_list, kfold
)
for lm, err in zip(lambda_list, error_list):
    print("Lambda: %.4f" % lm, "RMSE: %.6f" % err)

print("Best Lambda: %.4f" % best_lambda)
weight = reg.ridge_fit_closed(
    x_all_feat[train_indices], y_all_noisy[train_indices], c_lambda=best_lambda
)
y_test_pred = reg.predict(x_all_feat[test_indices], weight)
```

```
test_rmse = reg.rmse(y_test_pred, y_all_noisy[test_indices])
print("Best Test RMSE: %.4f" % test_rmse)
```

```
Lambda: 0.0001 RMSE: 0.986072
Lambda: 0.0010 RMSE: 0.987209
Lambda: 0.1000 RMSE: 0.989441
Lambda: 1.0000 RMSE: 0.987945
Lambda: 5.0000 RMSE: 0.986684
Lambda: 10.0000 RMSE: 0.986821
Lambda: 50.0000 RMSE: 0.989110
Lambda: 100.0000 RMSE: 0.994419
Lambda: 1000.0000 RMSE: 1.289583
Lambda: 10000.0000 RMSE: 2.544557
Best Lambda: 0.0001
Best Test RMSE: 1.0528
```

3.7 Noisy Input Samples in Linear Regression [2.3% Bonus for All] **[W]**

Consider a linear model of the form:

$$y(x_n, \theta) = \theta_0 + \sum_{d=1}^{D} \theta_d x_{nd}$$

where $x_n = (x_{n1}, ..., x_{nD}) \in R^D$ and weights $\theta = (\theta_0, ..., \theta_D) \in R^D$. Given the the D-dimension input sample set $x = \{x_1, ..., x_n\}$ with corresponding target value $y = \{y_1, ..., y_n\}$, the sum-of-squares error function is:

$$E_D(\theta) = \frac{1}{2} \sum_{n=1}^{N} \Big[y(x_n, \theta) - y_n \Big]^2$$

Now, suppose that Gaussian noise $\epsilon_n \in R^D$ is added independently to each of the input sample x_n to generate a new sample set $x' = \{x_1 + \epsilon_1, ..., x_n + \epsilon_n\}$. Here, ϵ_{ni} (an entry of ϵ_n) has zero mean and variance σ^2. For each sample x_n, let $x'_n = (x_{n1} + \epsilon_{n1}, ..., x_{nD} + \epsilon_{nd})$, where n and d is independent across both n and d indices.

1. (0.7% Bonus) Show that $y(x'_n, \theta) = y(x_n, \theta) + \sum_{d=1}^{D} \theta_d \epsilon_{nd}$

2. (1.6% Bonus) Assume the sum-of-squares error function of the noise sample set $x' = \{x_1 + \epsilon_1, ..., x_n + \epsilon_n\}$ is $E_D(\theta)'$. Prove the expectation of $E_D(\theta)'$ is equivalent to the sum-of-squares error $E_D(\theta)$ for noise-free input samples with the addition of a weight-decay regularization term (e.g. ℓ_2 norm) , in which the bias parameter θ_0 is omitted from the regularizer. In other words, show that $E[E_D(\theta)'] = E_D(\theta) + \text{Regularizer}$.

N.B. You should be incorporating your solution from the first part of this problem into the given sum of squares equation for the second part.

Write your responses below using LaTeX in Markdown.

HINT:

- During the class, we have discussed how to solve for the weight θ for ridge regression, the function looks like this: $E(\theta) = \frac{1}{N} \sum_{i=1}^{N} \Big[y(x_i, \theta) - y_i \Big]^2 + \frac{\lambda}{N} \sum_{i=1}^{d} |\theta_i|^2$ where the first term is the sum-of-squares error and the second term is the regularization term. N is the number of samples. In this question, we use another form of the ridge regression, which is: $E(\theta) = \frac{1}{2} \sum_{i=1}^{N} \Big[y(x_i, \theta) - y_i \Big]^2 + \frac{\lambda}{2} \sum_{i=1}^{d} |\theta_i|^2$

- For the Gaussian noise ϵ_n, we have $E[\epsilon_n] = 0$

- Assume the noise $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ are **independent** to each other, we have $E[\epsilon_n \epsilon_m] = \begin{cases} \sigma^2 & m = n \\ 0 & m \neq n \end{cases}$

1. **Answer:**

2. **Answer:**

Q4: Naive Bayes and Logistic Regression [40pts] [P] | [W]

In Bayesian classification, we're interested in finding the probability distribution of the label space given some observed feature vector $x = [x_1, ..., x_d]$, which we can write as $P(y \mid x_1, ..., x_d)$. Bayes's theorem tells us how to express this in terms of quantities we can compute more directly:

$P(y \mid x_1, ..., x_d) = P(x_1, ..., x_d \mid y)P(y)P(x_1, ..., x_d)$

The main assumption in Naive Bayes is that, given the label, the observed features are conditionally independent i.e.

$P(x_1, ..., x_d \mid y) = P(x_1 \mid y) \times ... \times P(x_d \mid y)$

Now, let's apply this to a real-life scenario.

4.1 Profile Screening [7pts] [W]

4.1.1 Profile Screening using Naive Bayes [5pts] [W]

In the rapidly evolving job market of Techlanta, a leading tech company has devised an automated resume screening system to assist in the hiring process for three distinct roles: Machine Learning Engineer, Data Analyst, and Product Manager. This system is designed to evaluate applicants based on five key binary attributes extracted from their resumes: {coding proficiency (1 for high, 0 for low), data analysis skills (1 for strong, 0 for weak), leadership experience (1 for yes, 0 for no), product design experience (1 for yes, 0 for no), marketing skills (1 for strong, 0 for weak)}.

Aiming to ensure that the screening process is both fair and effective, the company is mindful of avoiding biases that could disadvantage any applicant. They have compiled a dataset of 12 anonymized resumes, with each position having 4 applicants, alongside feedback from previous interviews to serve as the ground truth.

Machine Learning Engineer applicants have demonstrated attributes such as: {1, 1, 0, 0, 1}, {1, 1, 1, 0, 0}, {1, 0, 0, 1, 0}, {0, 1, 0, 1, 0}

Data Analyst applicants are characterized by: {0, 1, 1, 0, 0}, {1, 0, 1, 0, 1}, {0, 1, 1, 0, 1}, {0, 1, 0, 1, 0}

Product Manager applicants display: {0, 1, 0, 1, 0}, {0, 0, 1, 0, 1}, {1, 0, 1, 1, 1}, {0, 1, 0, 1, 1}.

A new applicant's resume has been screened, and identified to **not have** leadership, **have** data analysis skills **without** product design experience but **does have** coding experience and marketing skills.

Now is the time to test your method!

Using a multiclass Naive Bayes classifier, determine the most suitable job position for the new applicant.

NOTE: We expect students to show their work (prior probabilities, likelihood, and resulting posterior probabilites) and not just the final answer.

Don't divide your answer by the marginal probability P(x). Since this is not a function of the label, it will be the same for every label, thus this additional division is not necessary when using Naive Bayes and should be skipped. Doing so will result in a deduction during grading.

4.1. **Answer:**

4.1.2 AI-Driven Profile Screening [2pts] **[W]**

Traditionally, human HR personnel review resumes to identify candidates who meet the minimum qualifications for a job. This process is subjective and can be influenced by conscious or unconscious biases. More recently, many organizations use Applicant Tracking Systems (ATS) to manage and screen resumes. ATS systems parse resumes to extract information like education, experience, skills, and other relevant details. They rank candidates based on how well their resume matches the job description and criteria set by the employer.

More advanced systems incorporate AI to not only parse and match resumes but also to predict a candidate's job performance, cultural fit, and even retention likelihood. These systems can use machine learning models trained on historical hiring data, and they often employ a combination of keyword matching, machine learning models (such as decision trees, support vector machines, and neural networks), and natural language processing (NLP) techniques.

In recent years, several high-profile cases have highlighted the challenges of bias in AI-driven resume screening processes. For example:

- In 2018, Amazon had to abandon its AI recruitment tool because it was trained on a decade's worth of resumes, predominantly from men, leading to a bias against women's resumes, such as penalizing resumes that mentioned "women's" clubs or activities.[1]
- UnitedHealth Group faced allegations of racial bias in their AI-driven hiring tool. The system reportedly favored white applicants over black applicants. The company discontinued the tool after these concerns were raised.[2].
- HireVue's AI tool analyzes video interviews, and has been used by more than a 100 companies on over a million applicants, according to the Washington Post[3]. Notice that the algorithm can learn historical patterns in the data (e.g., gender, race, socioeconomic status) and be more likely to mark "traditional" applicants (white, male, able-bodied) as more employable.[4] As a result, applicants who deviate from the "traditional"—including people don't speak English as a native language or who are disabled—are likely to get lower scores.[5]
- LinkedIn's job-matching AI was found to exhibit gender biases in job recommendations, a consequence of the training data's inherent patterns. The algorithms ranked candidates based on their likelihood to apply for a position or respond to a recruiter. As a result, more men were referred for open roles due to their proactive approach in seeking new opportunities.[6]

Sources:

1: Guardian. 2: The Wall street Journal. 3: Washington Post. 4: MIT Technology Review. 5: Brookings. 6: MIT Technology Review.

Given the context above, which of the following approaches is most effective for mitigating bias in AI-driven resume screening systems? (Choose all that apply.)

A. Increasing the size of the training dataset

B. Conducting regular audits of the AI system's decisions

C. Utilizing differential privacy techniques (introducing "noise" or subtle alterations to the data in a way that protects individual privacy while allowing the aggregate patterns to remain intact)

D. Relying solely on keyword matching to ensure objectivity

4.1.2 **Answer:**

4.2 News Data Sentiment Classification via Logistic Regression [30pts] **[P]**

This dataset contains the sentiments for financial news headlines from the perspective of a retail investor. The sentiment of news has 3 classes, negative, positive and neutral. In this problem, we only use the negative (class label = 0) and positive (class label = 1) classes for binary logistic regression. For data preprocessing, we remove the duplicate headlines and remove the neutral class to get 1967 unique news headlines. Then we randomly split the 1967 headlines into training set and evaluation set with 8:2 ratio. We use the training set to fit a binary logistic regression model.

The code which is provided loads the documents, preprocess the data, builds a "bag of words" representation of each document. Your task is to complete the missing portions of the code in **logisticRegression.py** to determine whether a news headline is negative or positive.

In **logistic_regression.py** file, complete the following functions:

- **sigmoid**: transform s = xθ to probability of being positive using sigmoid function, which is $11+e^{-s}$.
- **bias_augment**: augment x with 1's to account for bias term in θ
- **predict_probs**: predicts the probability of positive label P(y = 1 | x)
- **predict_labels**: predicts labels
- **loss**: calculates binary cross-entropy loss
- **gradient**: calculate the gradient of the loss function with respect to the parameters θ.
- **accuracy**: calculate the accuracy of predictions
- **evaluate**: gives loss and accuracy for a given set of points
- **fit**: fit the logistic regression model on the training data.

Logistic Regression Overview:

1. In logistic regression, we model the conditional probability using parameters θ, which includes a bias term b. $p(y_i = 1 \mid x_i; \theta) = h_\theta(x_i) = \sigma(x\theta)$
   $p(y_i = 0 \mid x_i; \theta) = 1 - h_\theta(x_i)$

where $\sigma(\cdot)$ is the sigmoid function as follows: $\sigma(s) = 11 + e^{-s}$

2. The conditional probabilities of the positive class (y = 1) and the negative class (y = 0) of the sample x_i attributes are combined into one equation as follows:

$p(y * i \mid x_i; \theta) = (h * \theta(x * i))^{y_i}(1 - h * \theta(x_i))^{1-y_i}$

3. Assuming that the samples are independent of each other, the likelihood of the entire dataset is the product of the probabilities of all samples. We use maximum likehood estimation to estimate the model parameters θ. The negative log likelihood (scaled by the dataset size N) is given by:
   $L(\theta \mid X, Y) = -1N N\sum i=1 y\_i log h_\theta(x * i) + (1 - y\_i) log(1 - h * \theta(x\_i))$

where:

N = number of training samples
x_i = *bag of words* features of the i-th training sample
y_i = label of the i-th training sample

Note that this will be our model's loss function

4. Then calculate the gradient $\nabla_\theta L$ and use gradient descent to optimize the loss function: $\theta_{t+1} = \theta_t - \eta \cdot \nabla_\theta L(\theta_t \mid X, Y)$

where η is the learning rate and the gradient $\nabla_\theta L$ is given by:

$$\nabla_\theta L(\theta \mid X, Y) = \frac{1}{N} \sum_{i=1}^{N} x_i^T \left(h_\theta(x_i) - y_i \right)$$

4.2.1 Local Tests for Logistic Regression [No Points]

You may test your implementation of the functions contained in **logistic_regression.py** in the cell below. Feel free to comment out tests for functions that have not been completed yet. See Using the Local Tests for more details.

In [33]:
```
##############################
### DO NOT CHANGE THIS CELL ###
##############################

from utilities.localtests import TestLogisticRegression

unittest_lr = TestLogisticRegression()
unittest_lr.test_sigmoid()
unittest_lr.test_bias_augment()
unittest_lr.test_loss()
unittest_lr.test_predict_probs()
unittest_lr.test_predict_labels()
unittest_lr.test_loss()
unittest_lr.test_accuracy()
unittest_lr.test_evaluate()
unittest_lr.test_fit()
```

```
UnitTest passed successfully for "Logistic Regression sigmoid"!
UnitTest passed successfully for "Logistic Regression bias_augment"!
UnitTest passed successfully for "Logistic Regression loss"!
UnitTest passed successfully for "Logistic Regression predict_probs"!
UnitTest passed successfully for "Logistic Regression predict_labels"!
UnitTest passed successfully for "Logistic Regression loss"!
UnitTest passed successfully for "Logistic Regression accuracy"!
UnitTest passed successfully for "Logistic Regression evaluate"!
Epoch 0:
        train loss: 0.675       train acc: 0.7
        val loss:   0.675       val acc:   0.7
UnitTest passed successfully for "Logistic Regression fit"!
```

4.2.2 Logistic Regression Model Training [No Points]

In [34]:
```
##############################
### DO NOT CHANGE THIS CELL ###
##############################

from logistic_regression import LogisticRegression as LogReg
from logistic_regression import hyperparameter_tuning
```

In [35]:
```
##############################
### DO NOT CHANGE THIS CELL ###
##############################

news_data = pd.read_csv("./data/news-data.csv", encoding="cp437", header=None)

class_to_label_mappings = {"negative": 0, "positive": 1}
```

```
label_to_class_mappings = {0: "negative", 1: "positive"}

news_data.columns = ["Sentiment", "News"]
news_data.drop_duplicates(inplace=True)

news_data = news_data[news_data.Sentiment != "neutral"]

news_data["Sentiment"] = news_data["Sentiment"].map(class_to_label_mappings)

vectorizer = text.CountVectorizer(stop_words="english")

X = news_data["News"].values
y = news_data["Sentiment"].values.reshape(-1, 1)

RANDOM_SEED = 5
BOW = vectorizer.fit_transform(X).toarray()
indices = np.arange(len(news_data))
X_train, X_test, y_train, y_test, indices_train, indices_test = train_test_split(
    BOW, y, indices, test_size=0.2, random_state=RANDOM_SEED
)
```

Fit the model to the training data Try different learning rates `lr` and number of `epochs` to achieve >80% test accuracy.

In [36]:
```
################################
### DO NOT CHANGE THIS CELL ###
################################

model = LogReg()
lr = 0.05
epochs = 10000
threshold = 0.5
theta = model.fit(X_train, y_train, X_test, y_test, lr, epochs, threshold)
```

```
Epoch 0:
        train loss: 0.69        train acc: 0.7
        val loss:   0.691       val acc:   0.665
Epoch 1000:
        train loss: 0.436       train acc: 0.794
        val loss:   0.532       val acc:   0.701
Epoch 2000:
        train loss: 0.364       train acc: 0.846
        val loss:   0.484       val acc:   0.746
Epoch 3000:
        train loss: 0.318       train acc: 0.873
        val loss:   0.456       val acc:   0.761
Epoch 4000:
        train loss: 0.286       train acc: 0.896
        val loss:   0.438       val acc:   0.772
Epoch 5000:
        train loss: 0.262       train acc: 0.914
        val loss:   0.425       val acc:   0.782
Epoch 6000:
        train loss: 0.242       train acc: 0.926
        val loss:   0.416       val acc:   0.789
Epoch 7000:
        train loss: 0.226       train acc: 0.933
        val loss:   0.409       val acc:   0.797
Epoch 8000:
        train loss: 0.212       train acc: 0.943
        val loss:   0.404       val acc:   0.802
```

```
Epoch 9000:
      train loss: 0.2 train acc: 0.95
      val loss:   0.4 val acc:   0.799
```

4.2.3 Logistic Regression Model Evaluation [No Points]

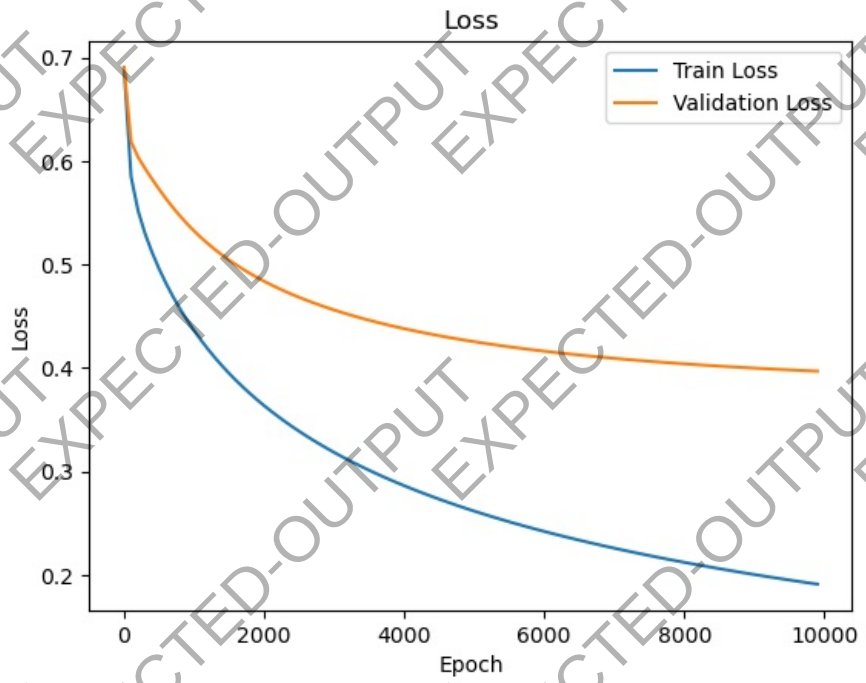
Evaluate the model on the test dataset

In [37]:
```
##############################
### DO NOT CHANGE THIS CELL ###
##############################

test_loss, test_acc = model.evaluate(X_test, y_test, theta, threshold)
print(f"Test Dataset Accuracy: {round(test_acc, 3)}")
```

```
Test Dataset Accuracy: 0.807
```

Plotting the loss function on the training data and the test data for every 100th epoch

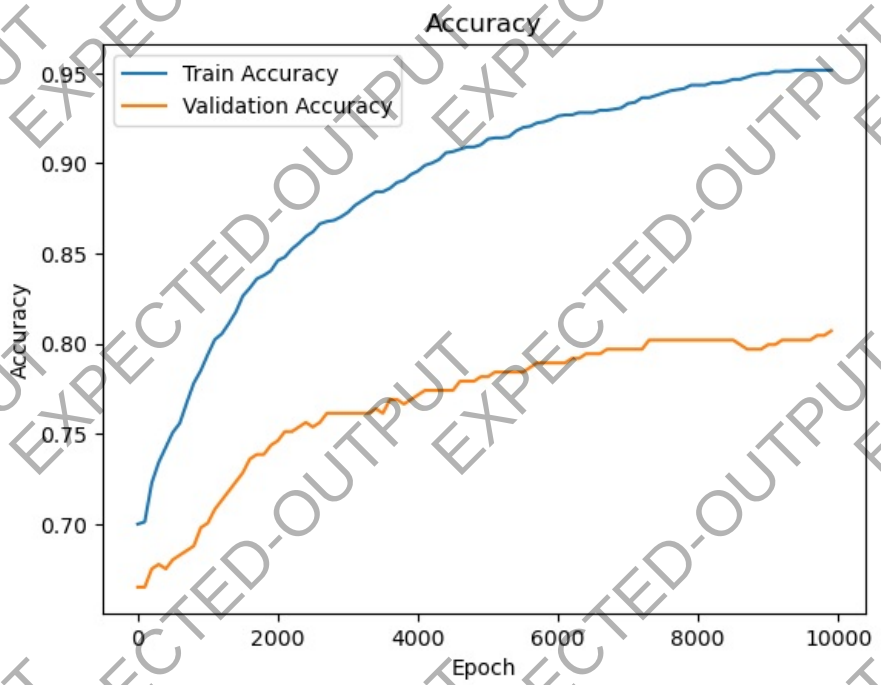
In [38]:
```
##############################
### DO NOT CHANGE THIS CELL ###
##############################

model.plot_loss()
```



Plotting the accuracy function on the training data and the test data for each epoch

In [39]:
```
##############################
### DO NOT CHANGE THIS CELL ###
##############################
```

```
model.plot_accuracy()
```



Check out sample evaluations from the test set.

In [40]:
```
################################
### DO NOT CHANGE THIS CELL ###
################################

num_samples = 10
for i in range(10):
    rand_index = np.random.randint(0, len(X_test))
    x_test = np.reshape(X_test[rand_index], (1, X_test.shape[1]))
    prob = model.predict_probs(model.bias_augment(x_test), theta)
    pred = model.predict_labels(prob, threshold)
    print(f"Input News: {X[indices_test[rand_index]]}\n")
    print(f"Predicted Sentiment: {label_to_class_mappings[pred[0][0]]}")
    print(f"Actual Sentiment: {label_to_class_mappings[y_test[rand_index][0]]}\n")
```

```
Input News: Metso said it has won an order worth around 40 mln eur to supply a kraftliner board machine to China 's Lee & Man Paper Co. .

Predicted Sentiment: positive
Actual Sentiment: positive

Input News: Xerox and Stora Enso have teamed up to tailor the iGen3 to the short-run , on-demand packaging market .

Predicted Sentiment: negative
Actual Sentiment: positive

Input News: In addition to Russia , we now seek additional growth in Ukraine .

Predicted Sentiment: positive
Actual Sentiment: positive

Input News: `` Stonesoft sees great promise in the future of IPv6 .

Predicted Sentiment: positive
Actual Sentiment: positive

Input News: These financing arrangements will enable the company to ensure , in line with its treasury policy , that it has sufficient financial instruments
at its disposal for its potential capital requirements .

Predicted Sentiment: positive
Actual Sentiment: positive

Input News: Via the move , the company aims annual savings of some EUR3m , the main part of which are expected to be realized this year .

Predicted Sentiment: positive
Actual Sentiment: positive

Input News: Ragutis , which is based in Lithuania 's second-largest city Kaunas , boosted its sales last year 22.3 per cent to 36.4 million liters .

Predicted Sentiment: positive
Actual Sentiment: positive

Input News: According to Arokarhu , some of the purchases that had been scanned into the cash register computer disappeared when the total sum key was press
ed .

Predicted Sentiment: negative
Actual Sentiment: negative

Input News: `` Stonesoft sees great promise in the future of IPv6 .

Predicted Sentiment: positive
Actual Sentiment: positive

Input News: However , the suspect stole his burgundy Nissan Altima .

Predicted Sentiment: positive
Actual Sentiment: negative
```

4.3 Logistic Regression Model Threshold Experiments [3pts] **[P]**

Recall that the sigmoid function in a logistic regression model outputs a decimal value between 0 and 1. For a classification problem, we need a threshold to determine which outputs are considered as positive and which are considered as negatives.

Please implement the hyperparameter_tuning method in logistic_regression.py to perform hyperparameter tuning on the thresholds.

In [41]:
```
##############################
### DO NOT CHANGE THIS CELL ###
##############################

unittest_lr.test_thresholding()
```

```
Epoch 0:
        train loss: 0.675       train acc: 0.7
        val loss:   0.675       val acc:   0.7
UnitTest passed successfully for "Hyperparameter Tuning"!
```

Q5 Feature Selection Implementation [30 pts] **[P]**

Feature selection is an integral aspect of machine learning. It is the process of selecting a subset of relevant features that are to be used as the input for the machine learning task. Feature selection may lead to simpler models for easier interpretation, shorter training times, avoidance of the curse of dimensionality, and better generalization by reducing overfitting.

5.1 Feature Reduction [30pts] **[P]**

In the **feature_reduction.py** file, complete the following functions:

- **forward_selection**
- **backward_elimination**

Reminder: A p-value is known as the observed significance value for a null hypothesis. In our case, the p-value of a feature is associated with the hypothesis $H_0 : \beta_j = 0$. If $\beta_j = 0$, then this feature contributes no predictive power to our model and should be dropped. We reject the null hypothesis if the p-value is smaller than our significance level. In short, a p-value is a measure of how much the given feature significantly represents an observed change. **A lower p-value represents higher significance.** Some more information about p-values can be found here: https://www.youtube.com/watch?v=vemZtEM63GY

Forward Selection:

In forward selection, we start with a null model and fit the model with one individual feature at a time. We then select the most significant feature with the lowest p-value. We continue to do this until we try to select a feature with a p-value >= significance level. This implies that all remaining features are insignificant with p-values < significance level and that no more features should be added.

Steps to implement it:

1. Choose a significance level (provided to you)
2. Start with an empty list of selected features
3. For each feature NOT yet included in the selected features:
   - Fit a simple regression model using the the selected features AND the feature under consideration
   - Record the p-value of the feature under consideration
4. Find the feature with the minimum p-value.
   - If the feature's p-value < significance level, ADD the feature to the selected features and repeat from Step 2.
   - Otherwise, stop and return the selected features

Backward Elimination:

In backward elimination, we start with a full model and then remove the most insignificant feature with the highest p-value. We continue to do this until we try to remove a feature with p-value < significance level. This implies that all of the remaining features are significant with pvalues < significance level, and should therefore be kept.

Steps to implement it:

1. Choose a significance level (provided to you)
2. Start with a full list of ALL features as selected features.
3. Fit a simple regression model using the selected features
4. Find the feature with the maximum p-value.
    - If the feature's p-value >= significance level, REMOVE the feature from the selected features and repeat from Step 3.
    - Otherwise, stop and return the selected features.

HINT: Use `sm.OLS` as your regression model (documentation here). Be sure to add bias to your regression model by augmenting your data using the `sm.add_constants` function

In [42]:
```
###############################
### DO NOT CHANGE THIS CELL ###
###############################

from utilities.localtests import TestFeatureReduction

unittest_feature_reduction = TestFeatureReduction()
unittest_feature_reduction.test_forward_selection()
unittest_feature_reduction.test_backward_elimination()
```

```
UnitTest passed successfully for "Forward Selection"!
UnitTest passed successfully for "Backward Elimination"!
```

In [43]:
```
###############################
### DO NOT CHANGE THIS CELL ###
###############################
from feature_reduction import FeatureReduction

bc_dataset = load_breast_cancer()
bc = pd.DataFrame(bc_dataset.data, columns=bc_dataset.feature_names)
print("Dataset Features: ", bc.columns.tolist())
bc["Diagnosis"] = bc_dataset.target

X = bc.drop("Diagnosis", axis=1)
y = bc["Diagnosis"]
featureselection = FeatureReduction()
# Run the functions to make sure two feature lists are generated, one for each method
forward_selection_feature_list = FeatureReduction.forward_selection(X, y, 0.1)
backward_selection_feature_list = FeatureReduction.backward_elimination(X, y, 0.1)
```

```
Dataset Features:  ['mean radius', 'mean texture', 'mean perimeter', 'mean area', 'mean smoothness', 'mean compactness', 'mean concavity', 'mean concave poi
nts', 'mean symmetry', 'mean fractal dimension', 'radius error', 'texture error', 'perimeter error', 'area error', 'smoothness error', 'compactness error',
'concavity error', 'concave points error', 'symmetry error', 'fractal dimension error', 'worst radius', 'worst texture', 'worst perimeter', 'worst area', 'w
orst smoothness', 'worst compactness', 'worst concavity', 'worst concave points', 'worst symmetry', 'worst fractal dimension']
```

In [44]:
```
###############################
### DO NOT CHANGE THIS CELL ###
###############################

# View selected features and perform linear regression using the selected features

print("Forward Feature Selection")
FeatureReduction.evaluate_features(X, y, forward_selection_feature_list)
print("Backward Feature Elimination")
FeatureReduction.evaluate_features(X, y, backward_selection_feature_list)
```

```
Forward Feature Selection
Significant Features: ['worst concave points', 'worst radius', 'worst texture', 'worst area', 'smoothness error', 'worst symmetry', 'compactness error', 'ra
dius error', 'worst fractal dimension', 'mean compactness', 'mean concave points', 'worst concavity', 'concavity error', 'area error']
RMSE: 0.23821726582785563

Backward Feature Elimination
Significant Features: ['mean radius', 'mean compactness', 'mean concave points', 'radius error', 'smoothness error', 'concavity error', 'concave points erro
r', 'worst radius', 'worst texture', 'worst area', 'worst concavity', 'worst symmetry', 'worst fractal dimension']
RMSE: 0.23745896004435194
```

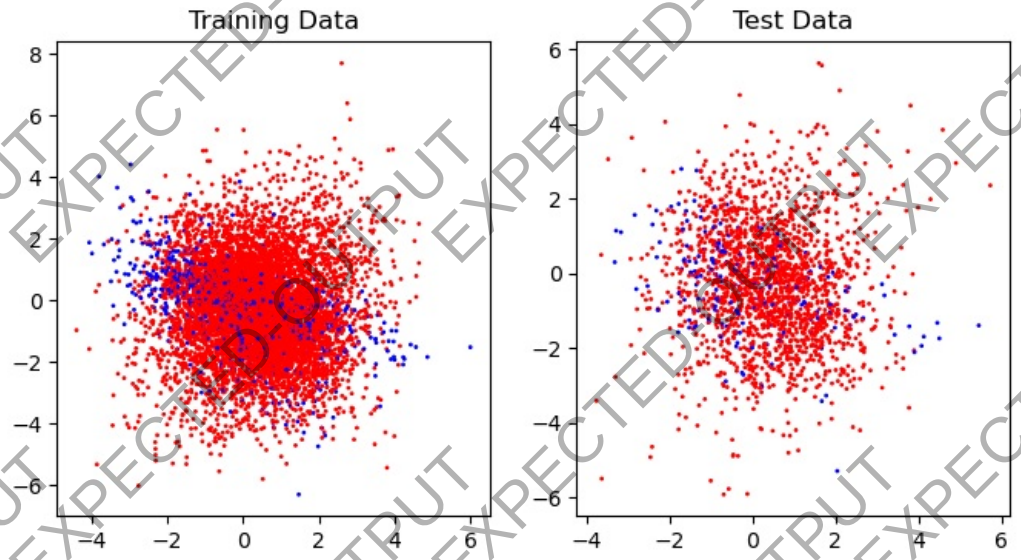
Q6: Imbalanced Classes in Classification Tasks [5.6% Bonus For All] **[P]**

In many datasets, the representation of classes in the training data is unequal. For example, most transactions in a banking system are legitimate, most images of manufactured parts show no visual defect, most email attachments are entirely benign. However, when you have highly imbalanced data, your model may converge to a highly biased estimator, classifying most inputs to the majority class. This is a valid solution, after all, the model is simplying converging based on the a priori distribution of classes in the training data, but we still want our model to have accurate prediction on the minority class.

To illustrate this point, take a look at the following 2D artificial dataset with a high class imbalance, and the results of training a classifier on it.

In [45]:
```python
# Set higher imbalance and lower separability
SEED = 42
imb = 0.9  # More imbalance (90% in one class)
class_sep = 0.4  # Lower separability (more overlap)
flip_y = 0

# Generate a more imbalanced and less separable dataset
X, y = make_classification(
    n_samples=10000,
    weights=[imb],
    n_features=10,
    n_informative=5,
    n_redundant=5,
    n_repeated=0,
    n_classes=2,
    n_clusters_per_class=2,
    flip_y=flip_y,
    class_sep=class_sep,
    random_state=SEED,
)

# Split the dataset
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=SEED
)

# Visualizing the dataset with two features
a, b = 1, 4  # Select two features for visualization
fig, axs = plt.subplots(1, 2, figsize=(8, 4))
axs[0].scatter(
    X_train[:, a],
    X_train[:, b],
    c=y_train,
    cmap=matplotlib.colors.ListedColormap(["red", "blue"]),
    s=1,
)
axs[0].title.set_text(f"Training Data")
```

```
axs[1].scatter(
    X_test[:, a],
    X_test[:, b],
    c=y_test,
    cmap=matplotlib.colors.ListedColormap(["red", "blue"]),
    s=1,
)
axs[1].title.set_text(f"Test Data")
plt.show()
```



This data is higher than 2 dimensional, but we can visualize any dim-2 subspace, just for understanding purposes. If you want, you can tinker with `a` and `b` to choose different subspaces, though most won't be particularly informative. Note that while the data doesn't look even vaguely separable in 2D, distances stack up as you add dimensions, so in 10D, this data is separable.

Let's run a basic classifier on the data!

In [46]:
```
################################
### DO NOT CHANGE THIS CELL ###
################################

classifier = SVC(
    C=0.1, probability=True, random_state=SEED
)  # sklearn's Support Vector Classifier

classifier.fit(X_train, y_train)
y_predicted = classifier.predict(X_test)

# y_pred has the probabilities of the positive class.
y_pred = classifier.predict_proba(X_test)[:, 1]

# Display the accuracy of this prediction.
print(f"Accuracy: {100*accuracy_score(y_test, y_predicted)}%")
```

```
Accuracy: 91.75%
```

This classifier achieved very good accuracy, but as we'll see, accuracy is not always a perfect measure for determining the goodness of a model, since it can hide a bias.

6.1 A More Comprehensive Measure [1.75% Bonus] **[P]**

To get a better view of our model's accuracy, we need to generate a confusion matrix. Each data point in the test set has a true class value and a predicted class value. A confusion matrix counts the instances of every possible combination of test and prediction values.

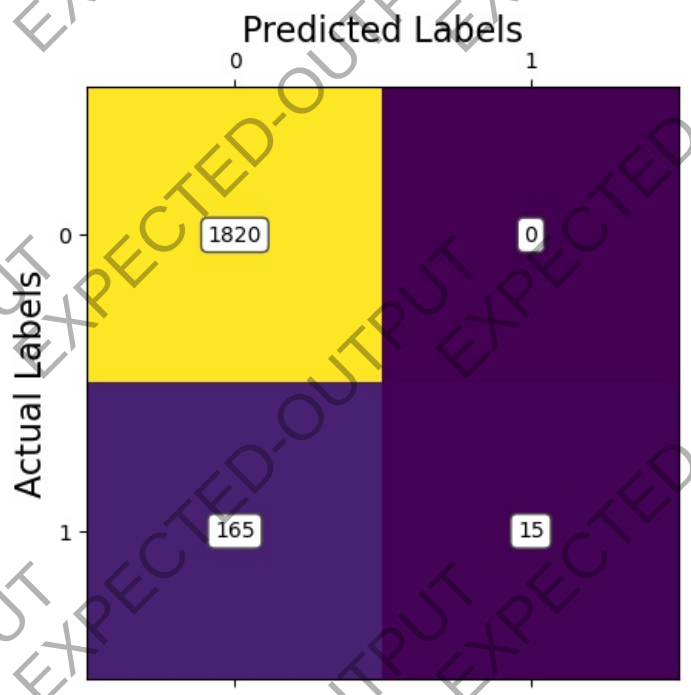
In the **smote.py** file, complete the following:

1. **generate_confusion_matrix**: Given the true test labels and the predicted labels from a model, generate a confusion matrix. C[i, j] should denote the number of instances where a sample from class i was predicted to be in class j. Even though our example is binary classification, your code should work for an arbitrary number of classes.

In [47]:
```
from smote import SMOTE
from utilities.localtests import TestSMOTE

unittest_sm = TestSMOTE()

sm = SMOTE()

unittest_sm.test_simple_confusion_matrix()
unittest_sm.test_complex_confusion_matrix()
```

```
UnitTest passed successfully for "simple confusion matrix"!
UnitTest passed successfully for "multiclass confusion matrix"!
```

In [48]:
```
# Display the confusion matrix of the prediction we made.
from smote import confusion_matrix_vis

confusion_matrix_vis(sm.generate_confusion_matrix(y_test, y_predicted))
```



Just from a first look, this relatively high accuracy is clearly not a good measure of the model's performance, as the model is highly biased. This (albeit naive) model is

very accurate when given a point from class 0. However, when given a point from class 1, it's much less inaccurate.

Depending on the application, e.g., cancer screening or facial recognition, you may want to prioritize minimizing false negatives or false positives depending on your application's objectives, in which case you might have to accept a low degree of accuracy for some classes, but in this exploration, we just want a balanced performance across our labels. We need a measure of test performance that not only conveys the accuracy of the model, but is robust against bias. A common metric is Macro-F1 average, but we'll explore another: ROC-AUC.

Receiver Operation Characteristic Area Under the Curve (ROC AUC)

The Receiver Operating Characteristic (ROC) curve is a tool which is used to evaluate the performance of a classification model. Classification models often give a probability distribution across the labels. In a binary setting, like with our SVC, this can be represented with a single number, which, WLOG, we choose to refer to the probability of predicting the positive class. Then, we can adjust the classification threshold, which is the probability cutoff for deciding between classes, allowing us to generate several different predictions from the model.

Each such binary prediction from the cutoff can be evaluated with normal metrics. In ROC-AUC, we want to look at False Positive Rate (FPR) and True Positive Rate (TPR).Mathematically, the True Positive Rate (TPR) and False Positive Rate (FPR) are defined as:

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

where:

- **TP** (True Positives) are correctly predicted positive cases,
- **FN** (False Negatives) are actual positive cases wrongly predicted as negative,
- **FP** (False Positives) are actual negative cases wrongly predicted as positive,
- **TN** (True Negatives) are correctly predicted negative cases.

For every prediction we can generate a (FPR, TPR) pair, and plot that on a graph of TPR (y-axis) vs. FPR (x-axis). The curve connecting these pairs forms the Receiver Operating Characteristic. Here is a reference to ROC AUC.

Here is an example walking through that calculation.

We consider 10 samples where y_{true} contains the actual class labels, and y_{pred} contains the predicted probabilities from our classifier.

$$y_{true} = [0, 0, 0, 0, 0, 1, 1, 1, 1, 1]$$

$$y_{pred} = [0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 1.0]$$

First, we threshold y_{pred} to make it into a binary vector. Let us consider the calculation at threshold value 0.5, where $\hat{y}_{pred} = y_{pred} \geq 0.5$.

$$\hat{y}_{pred} = [0, 0, 0, 1, 1, 1, 1, 1, 1, 1]$$

Now we can compute the True Positive (TP), False Negative (FN), False Positive (FP) and True Negative (TN):

$$TP = 5 \; FN = 0 \; FP = 2 \; TN = 3$$

Based on the above values, we can compute the FPR and TPR:

$$FPR = \frac{2}{2 + 3} = \frac{2}{5} = 0.4 \; TPR = \frac{5}{5 + 0} = \frac{5}{5} = 1$$

This gives us the point (0.4, 1.0). But now, we must do this for several thresholds. In this question, we will consider evenly spaced thresholds, though, there are other techniques for choosing the threshold values. Let's choose 11 thresholds:

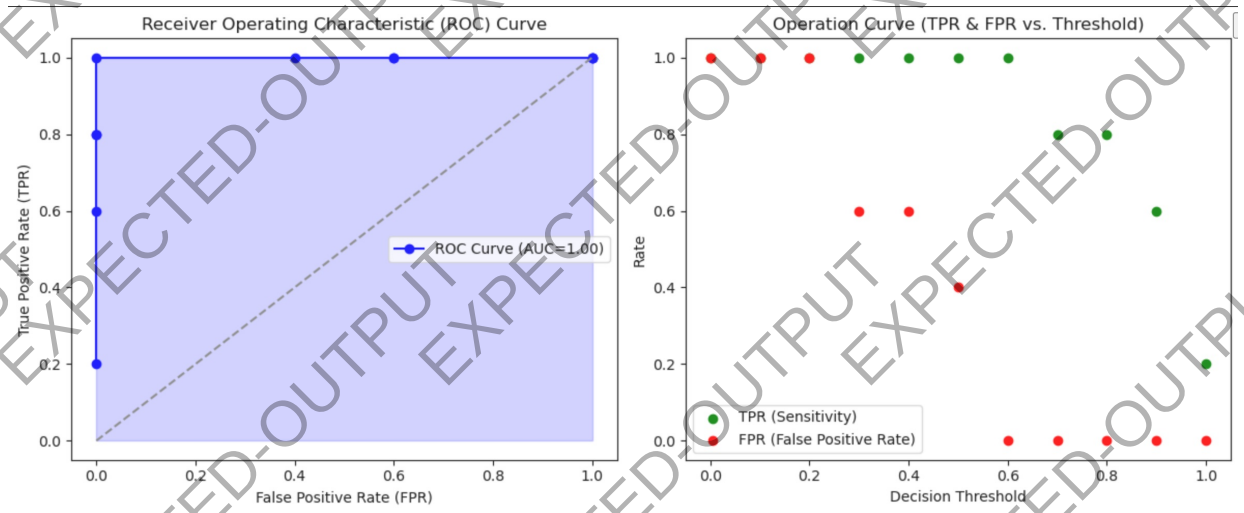
threshold = [0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]

After doing this for all the thresholds, we get:

FPR = [0.0, 0.0, 0.0, 0.0, 0.0, 0.4, 0.6, 0.6, 1.0, 1.0, 1.0] TPR = [0.2, 0.6, 0.8, 0.8, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0]

We need to sort the values such that FPR is in ascending order. However, when multiple points have the same FPR, we must sort them by TPR in ascending order to ensure that the ROC curve correctly moves upwards and does not jag back down. Sorting properly is important for accurately calculating AUC. The indices of the sorted values must also be correctly aligned so that each TPR and threshold corresponds to its respective FPR. This alignment preserves the correct order of points for plotting the ROC curve and ensures the correct integration for computing AUC.

After that, we get the ROC plot as follows:



Based on the above ROC curve, the area under that curve (AUC) will tell us how well our model has differentiated between the two classes.

We calculate the AUC as the integral of TPR with respect to FPR:

$AUC = \int_{10} TPR(FPR) d(FPR)$

Since the ROC is a plot of TPR versus FPR, the area under the curve represents how much the model's TPR improves as we allow more FPR.

- If AUC = 1, the model is classifying perfectly, and there exists a threshold at which the model perfectly discriminates the classes
- If AUC = 0.5, the model is doing no better than randomly guessing

Higher value of AUC represents better classification performance.

In this example, the value of AUC is 1 which means that the classifier model is perfect in distinguishing between positive and negative classes.

In the **smote.py** file, complete the following:

1. **compute_tpr_fpr**: Input y_true and y_pred calculated using predict_proba() which contains the probability of class 1 (true class). You will need to calculate the True Positive Rate (TPR) and False positive Rate (FPR) by creating the thresholds in the function and compute the TPR and FPR according to the corresponding thresholds.

The output should contain the value of TPR, FPR and threshold arrays. Please be careful with the order of the output. Ensure that the FPR is sorted, and the TPR and threshold values are corresonding to the values of FPR elements.

2. **compute_roc_auc**: Input the TPR and FPR array that you get from the compute_tpr_fpr function to calculate the Area under the curve of the Receiver Operating Characteristic Curve. The return value should be a float which is between 0 and 1.

In [49]:
```
unittest_sm.test_tpr_fpr()
```

```
UnitTest passed successfully for "TPR, FPR and threshold"!
```

In [50]:
```
unittest_sm.test_roc_auc()
```

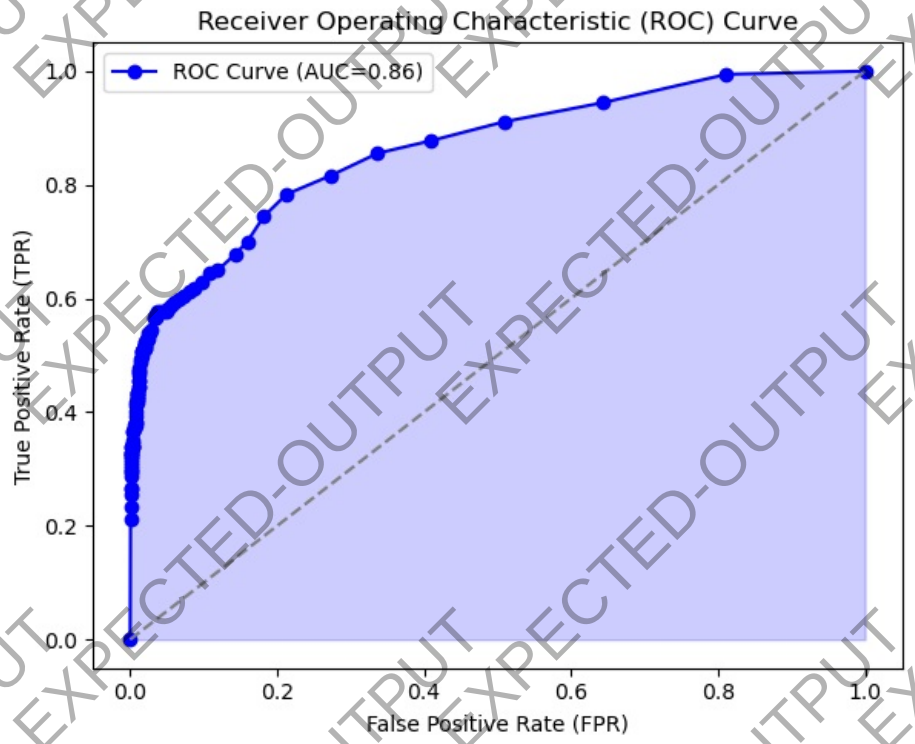
```
UnitTest passed successfully for "The implementation of compute_roc_auc"!
```

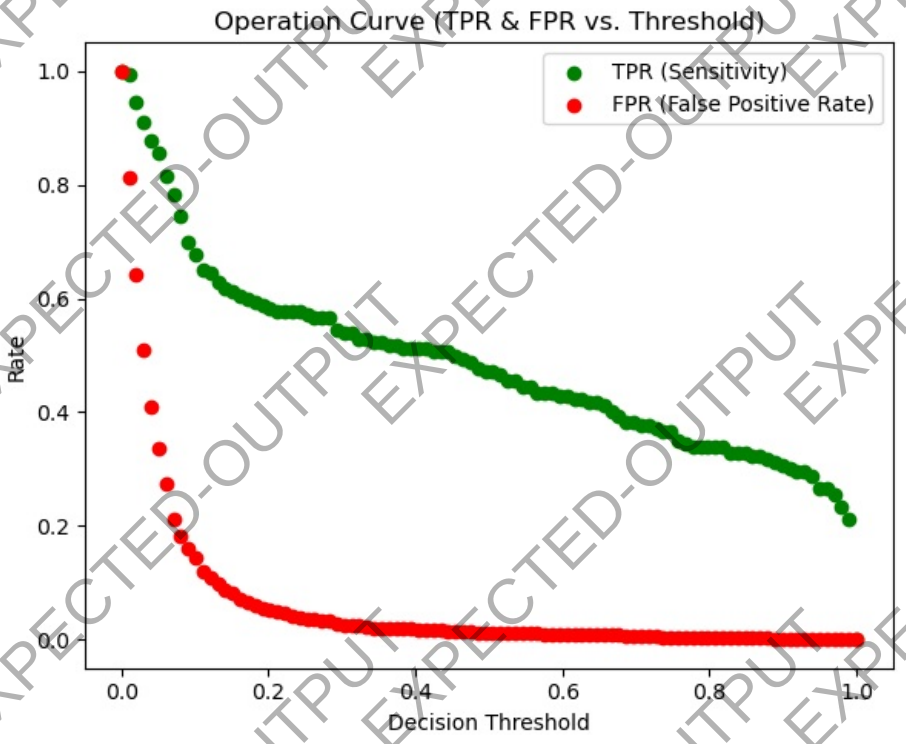
Let's take a look at our new metric.

In [51]:
```
conf = sm.generate_confusion_matrix(y_test, y_predicted)
print(f"Accuracy: {100*accuracy_score(y_test, y_predicted)}%")
# print(f"F1 Scores: {np.round(sm.f1_scores(conf), 3)}")
tpr, fpr, threshold = sm.compute_tpr_fpr(y_test, y_pred)
auc = sm.compute_roc_auc(tpr, fpr)
print(f"AUC: {auc}")
```

```
Accuracy: 91.75%
AUC: 0.863992673992674
```

In [52]:
```
sm.plot_roc_auc(auc, tpr, fpr, threshold)
```





To improve the performance bias, many models employ a technique called class weighting. Essentially, when the model fit or iterative training is performed, the

loss/gain/effect (varies from model to model) caused by each point in the training set can be weighted by the inverse of the proportion of that point's class. Applied to what you just implemented (logistic regression), that might look like this: $\theta_{t+1} = \theta_t - \eta \cdot 1N\sum i=1x_{\top i}\left(h_\theta(x_i) - y_i\right) \downarrow \theta_{t+1} = \theta_t - \eta \cdot 1N\sum i=1x_{\top i}\left(h_\theta(x_i) - y_i\right) \cdot N|$

$\{y \in Y \mid y = y_i\}\left(\text{ total \# of points\# of points in } y_i\text{'s class}\right)$

Class weighting is generally applicable, and is often the preferred solution, but on some models it's inaccessible, unwieldy, or produces undesirable results. There is an alternative solution: instead of changing the model, eliminate the problem. If our training data had balanced classes, we could just run the model. This is the idea behind oversampling. Oversampling refers to the practice of populating the input space with points from the input space itself.

Though, if you just sampled minority classes with replacement, you would get duplicates of the same point in the training data. This can genuinely help (since it approximates the class weight technique), but there is a more sophisticated solution: the Synthetic Minority Oversampling TEchnique (SMOTE). You can read the original paper on the arxiv for your own edification, but we'll be implementing a slightly different version, so the details in this Notebook and the function docstrings are sufficient to complete this HW.

6.2 SMOTE [3.85% Bonus] **[P]**

Instead of directly sampling the points from the minority class to bring it up to size, SMOTE samples a training point from the minority, samples another training point from the minority class that is "within a neighborhood around the first point," then randomly linearly interpolates between them to generate a new "synthetic" point lying on the line segment drawn between those two points. In this section, we're going to focus in on our binary classification problem, so we will only be oversampling the points from the one minority class. Additionally, our algorithm will only oversample to equality (the point at which the two classes are of equal size). Though, in practice, the amount you oversample becomes a hyperparameter to your model, which you can sweep with cross-validation.

In the **smote.py** file, complete the following:

1. **interpolate**: Given a start point, an end point, and an interpolation coefficient, return a linearly interpolated point.
2. **k_nearest_neighbors**: Given some set of points (N, D) and a parameter k, generate an (N, k) array of indices such that `output[i]` contains the k indices corresponding to the k nearest neighbors of point i.
3. **smote**: Given some data `X` (|maj|+|min|, D) and their binary labels `y` (|maj|+|min|,), generate |maj|-|min| new synthetic points from the minority class and return only those new synthetic points.

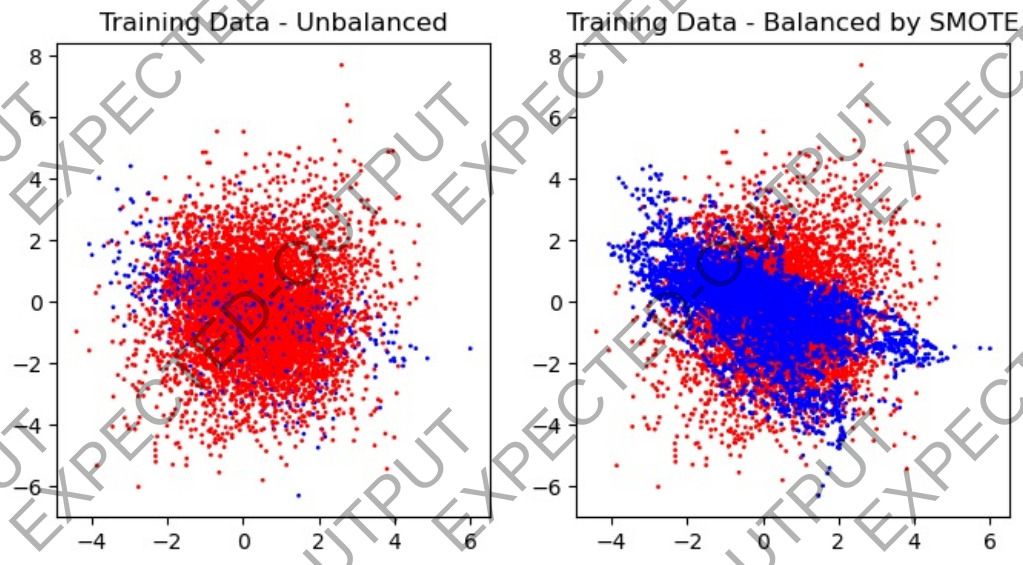
In [53]:
```
unittest_sm.test_interpolate()
unittest_sm.test_knn()
unittest_sm.test_smote()
```

```
UnitTest passed successfully for "interpolation"!
UnitTest passed successfully for "k nearest neighbors"!
UnitTest passed successfully for "SMOTE"!
```

Let's apply SMOTE to our dataset and see the downstream effect on our classification task!

In [54]:
```
##############################
### DO NOT CHANGE THIS CELL ###
##############################

# Run SMOTE!
X_train_synth, y_train_synth = sm.smote(X_train, y_train, k=5, inter_coeff_range=(0, 1))
# Combine synthetic data with original data.
```

```
X_train_balanced = np.vstack((X_train, X_train_synth))
y_train_balanced = np.hstack((y_train, y_train_synth))

# Visualize
a, b = 1, 4
fig, axs = plt.subplots(1, 2, figsize=(8, 4))
axs[0].scatter(
    X_train[:, a],
    X_train[:, b],
    c=y_train,
    cmap=matplotlib.colors.ListedColormap(["red", "blue"]),
    s=1,
)
axs[0].title.set_text("Training Data - Unbalanced")
axs[1].scatter(
    X_train_balanced[:, a],
    X_train_balanced[:, b],
    c=y_train_balanced,
    cmap=matplotlib.colors.ListedColormap(["red", "blue"]),
    s=1,
)
axs[1].title.set_text("Training Data - Balanced by SMOTE")
plt.show()
```

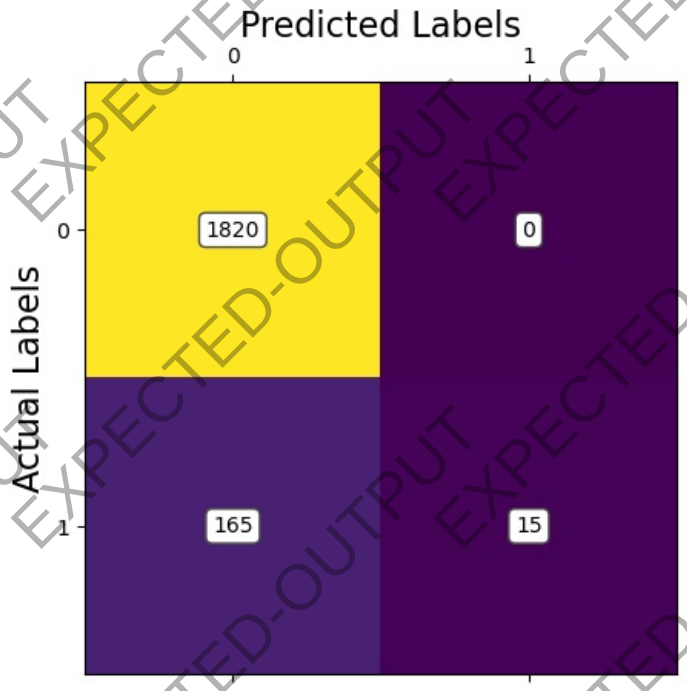


In [55]:

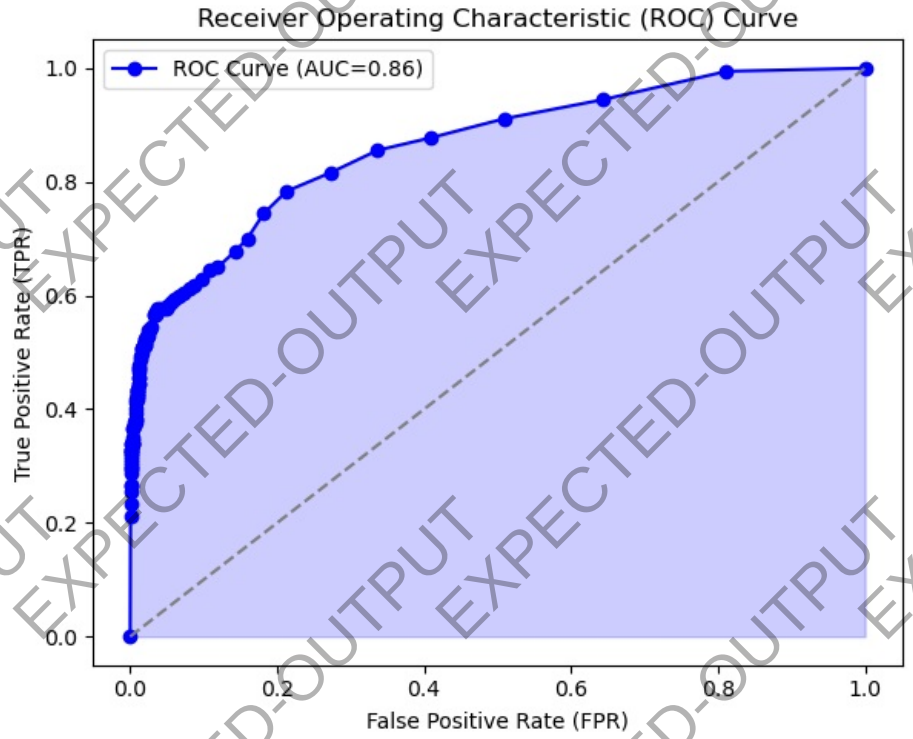
```
################################
### DO NOT CHANGE THIS CELL ###
################################

# Original
classifier.fit(X_train, y_train)
y_predicted1 = classifier.predict(X_test)
y_pred1 = classifier.predict_proba(X_test)[:, 1]

conf1 = sm.generate_confusion_matrix(y_test, y_predicted1)
print("Original performance:")
confusion_matrix_vis(conf1)
print(f"Accuracy: {100*accuracy_score(y_test, y_predicted1)}%")
```

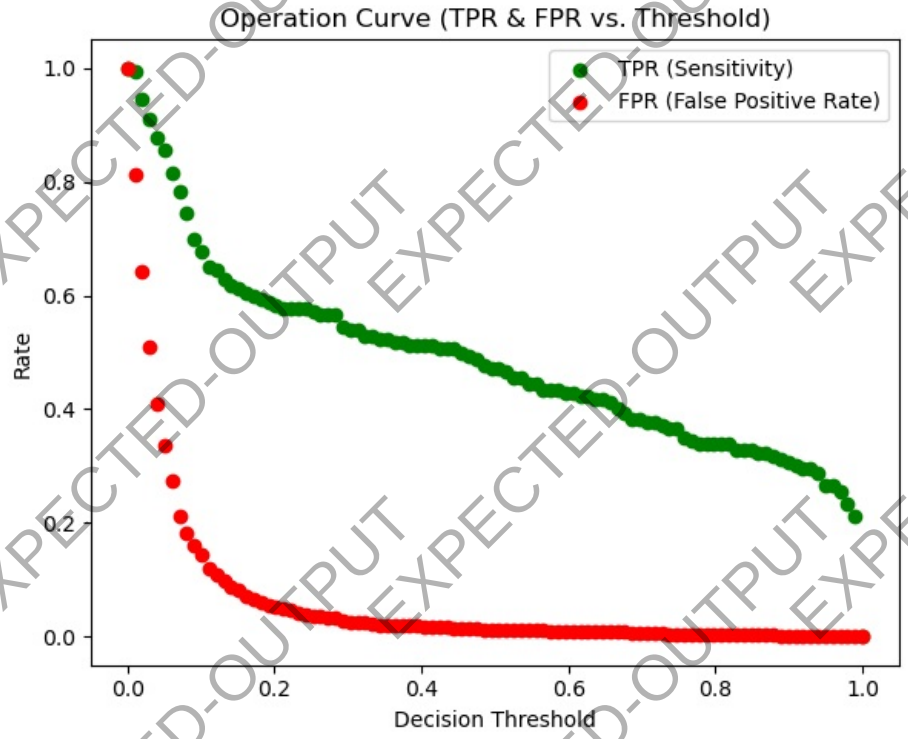
```
tpr1, fpr1, thresholds1 = sm.compute_tpr_fpr(y_test, y_pred)
auc1 = sm.compute_roc_auc(tpr1, fpr1)
print(f"AUC: {auc1}")
sm.plot_roc_auc(auc1, tpr, fpr, thresholds1)


# Balanced
classifier.fit(X_train_balanced, y_train_balanced)
y_predicted2 = classifier.predict(X_test)
y_pred2 = classifier.predict_proba(X_test)[:, 1]

print("\n\n\nPerformance after SMOTE:")
conf2 = sm.generate_confusion_matrix(y_test, y_predicted2)
confusion_matrix_vis(conf2)
print(f"SMOTEd Accuracy: {100*accuracy_score(y_test, y_predicted2)}%")
tpr2, fpr2, thresholds2 = sm.compute_tpr_fpr(y_test, y_pred2)
auc2 = sm.compute_roc_auc(tpr2, fpr2)
print(f"SMOTEd AUC: {auc2}")
sm.plot_roc_auc(auc2, tpr2, fpr2, thresholds2)


# If you're getting errors running this cell, check your implementation, then regenerate the SMOTE data by rerunning the previous cell!
```

```
Original performance:
```

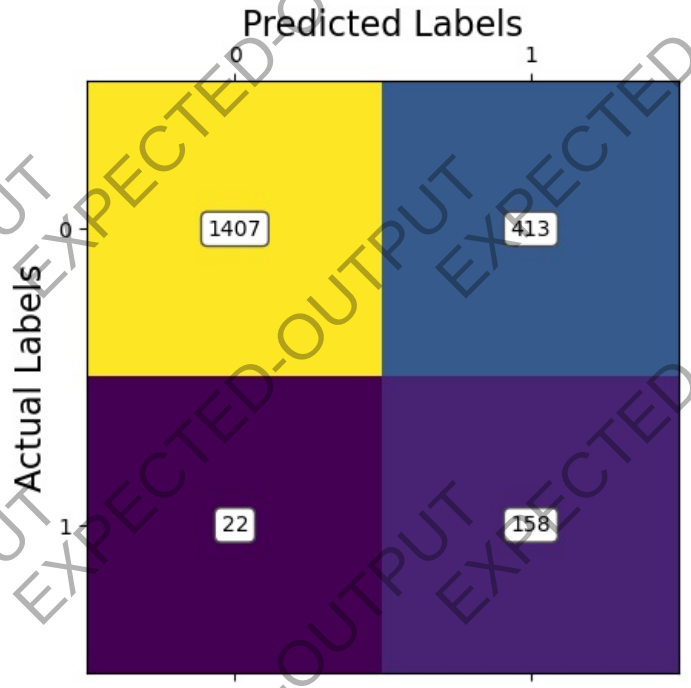



```
Accuracy: 91.75%
AUC: 0.863992673992674
```

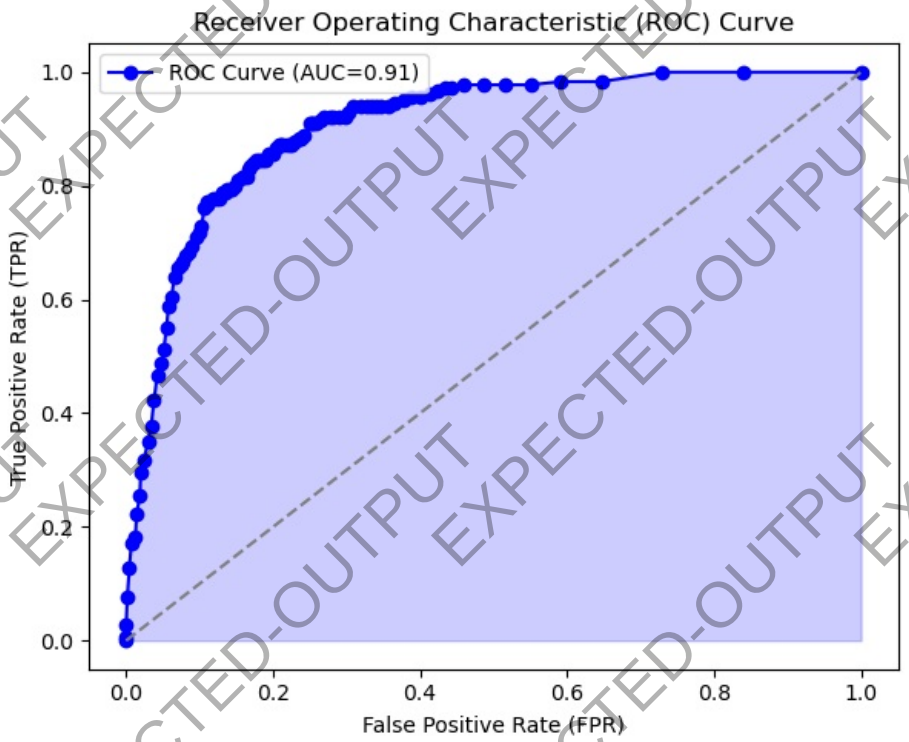


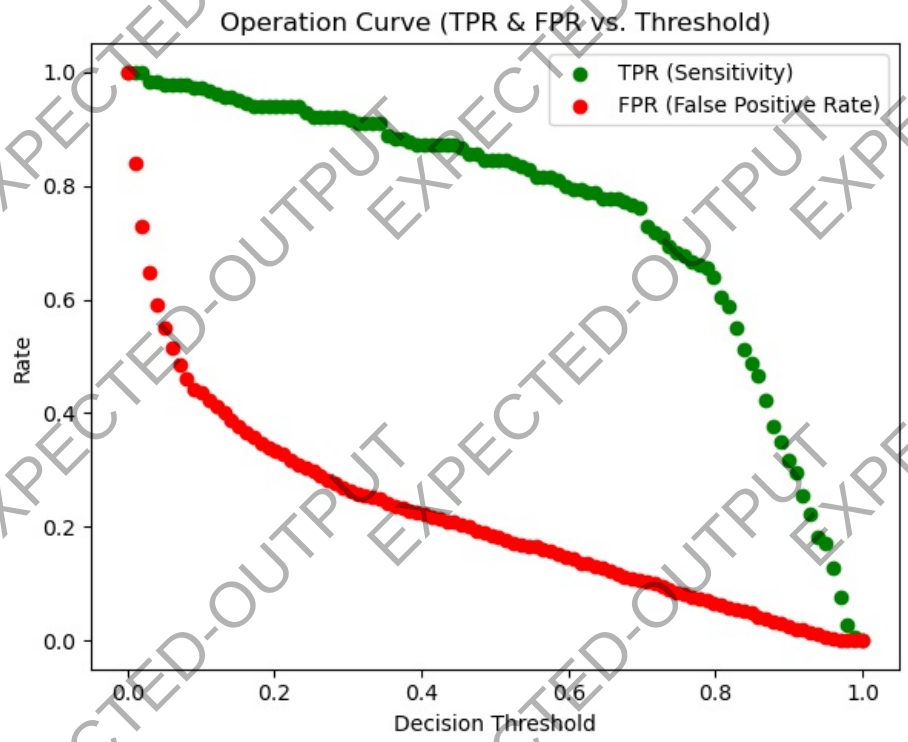



```
Performance after SMOTE:
```



SMOTEd Accuracy: 78.25%
SMOTEd AUC: 0.9062652625152625





With our new synthetic data (if done correctly), the value of ROC-AUC has improved, which means that the model is better able to distinguish between the classes. If you're unsatisfied with the accuracy dip, you could sample more or fewer points or tune the neighborhood cutoff. This technique has several hyperparameters, and there's no one way to do it, just pick the hyperparameters which result in the best performance.

Of course, SMOTE has limitations.

- Since SMOTE only works by interpolating on a line segment, SMOTE generates points within the convex hull of the input data. For classes with non-convex or multi-modal spatial distributions, SMOTE can actually ruin the quality of the class's representation in the input space for certain choices of k.
- For other types of data, like images, linear interpolation in general will not produce any meaningful data. You may have to use some semantic or generative combination instead.
- Additionally, SMOTE is very susceptible to outliers. Without undergoing some filtering, SMOTE will generate unreasonable points when interpolating with an outlier.
- Often, improving performance on the minority class can decrease performance on the majority class. Thus, if you want to reduce bias, sometimes you have to accept lower accuracy as well.

In general, just consider this another tool in your toolbox. It won't work on everything, but it will work exceptionally well on a few.

The class imbalance problem shows up all across ML and when creating human-facing ML models, the class imbalance problem has an additional ethical element tied on; since a lot of data collection is heavily biased to people of higher socio-economic status in a society, the resulting models trained from it may have higher performance quality for high SES individuals. Moreover, this isn't the only kind of data collection bias that currently exists. In general, a model that is performant to only a particular class of people will create an inherently unjust system that has the potential to relatively worsen the lives of those who are less represented in the training data, especially since some users will use your model as though it is foolproof.

If you're doing classification, This class imbalance problem will also probably show up in your project! Remember that accuracy isn't everything, and analysis of your system's bias is also important.