COMP3314 Tutorial 2

Assignment 2

2024-10-04

COMP3314 TAs

Assignment 2

- When to submit
  - Sunday 23:59pm, 2 weeks+ from now
- Overview
  - Question 1: Written assignment
    - 10 multiple choice
    - 1 calculation question
  - Question 2: Digits classification
  - Question 3: Dimensionality reduction

Assignment 2 Question 1: Multiple Choice Section

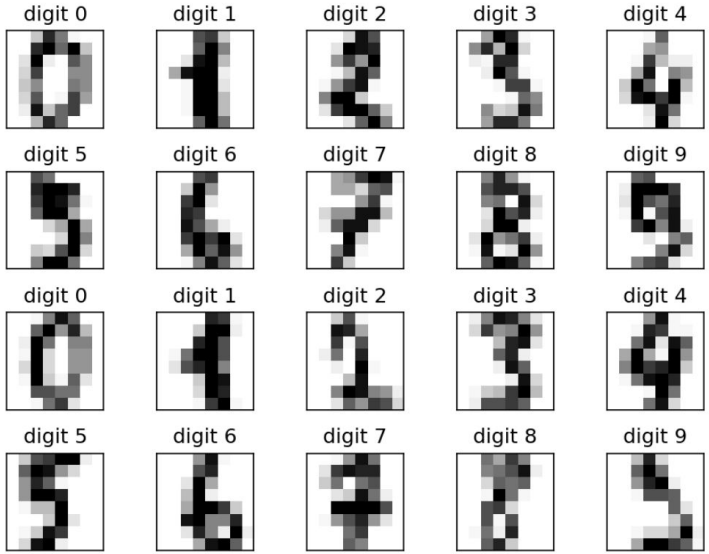
- Support Vector Machines (SVM):
  - Q1: Role of the kernel trick in transforming feature space.
  - Q2: Regularization parameter C in soft-margin SVM.
  - Q3: Effect of C on margin width and classification penalty.
- Principal Component Analysis (PCA):
  - Q4: Characteristics and properties of PCA (e.g., matrix representation).
  - Q5: Principal components as eigenvectors showing maximum variance.
  - Q6: Explained variance by principal components.
- K-Nearest Neighbors (KNN):
  - Q7: Memory requirements and the trade-off between accuracy and K value.
- Decision Trees:
  - Q8: Importance of pruning and invariance to input scaling.
  - Q9: Overfitting solutions, such as adjusting max-depth.
- Data Preprocessing:
  - Q10: When to use standardization vs normalization, especially for gradient descent with features of different scales.

Assignment 2 Question 1: Calculation Section

Genre	Age Group	Ticket Price	Preference?
Action	Adult	VIP	Yes
Drama	Adult	Regular	Yes
Drama	Adult	Regular	Yes
Action	Adult	VIP	No
Action	Teen	Regular	No
Action	Adult	Regular	Yes
Drama	Teen	Regular	Yes
Drama	Teen	VIP	Yes
Action	Teen	Regular	Yes
Drama	Adult	Regular	No
Drama	Adult	VIP	Yes
Drama	Teen	Regular	Yes
Drama	Adult	VIP	No
Drama	Teen	VIP	Yes
Action	Teen	VIP	No
Action	Teen	VIP	No

1. Compute Gini impurity at the root node.
2. For the 3 features (Genre, Age Group, and Ticket Price), compute the information gain if that feature is used to split the root node.
3. Conclude which feature is the most important feature to predict the preference as it maximizes the information gain.

Assignment 2 Question 2: Digits Classification



- Task 1: Create classifiers using scikit-learn:
  - Implement KNN classifier and scan for different numbers of neighbors.
  - Implement Decision Tree classifier and scan for different tree depths.
- Task 2: Compare training speed and performance between:
  - Naive (Linear) SVM and Kernel SVM.
  - Select a kernel function and discuss the differences in training time and accuracy.
- Task 3: Implement KNN from scratch:
  - Create your own KNN classifier.
  - Train and test it on the same dataset.
  - Compare its performance to the scikit-learn KNN implementation.

Assignment 2 Question 3: Dimensionality Reduction

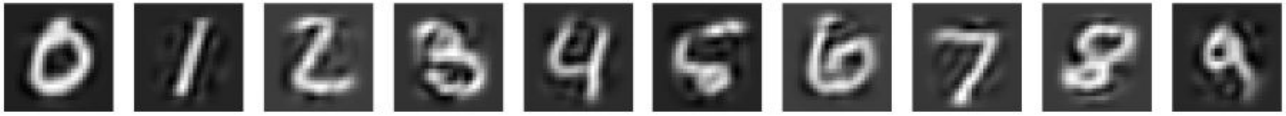
- Step 1: Dataset Preparation
  - Introduction to loading and splitting the MNIST dataset.
- Step 2: Visualize Digits
  - Demonstration of each digit's visualization from the dataset.
- Step 3: PCA Projection and Recovery
  - Application of PCA to reduce dimensions and then recover to visualize distortion.
- Step 4: PCA Evaluation
  - Discussion on evaluating PCA performance through explained variance.
- Step 5: t-SNE Visualization
  - Implementation of t-SNE to project data to 2D and visualize clusters.
- Step 6: PCA vs. t-SNE Visualization
  - Comparison between PCA and t-SNE for 2D visualization effectiveness.

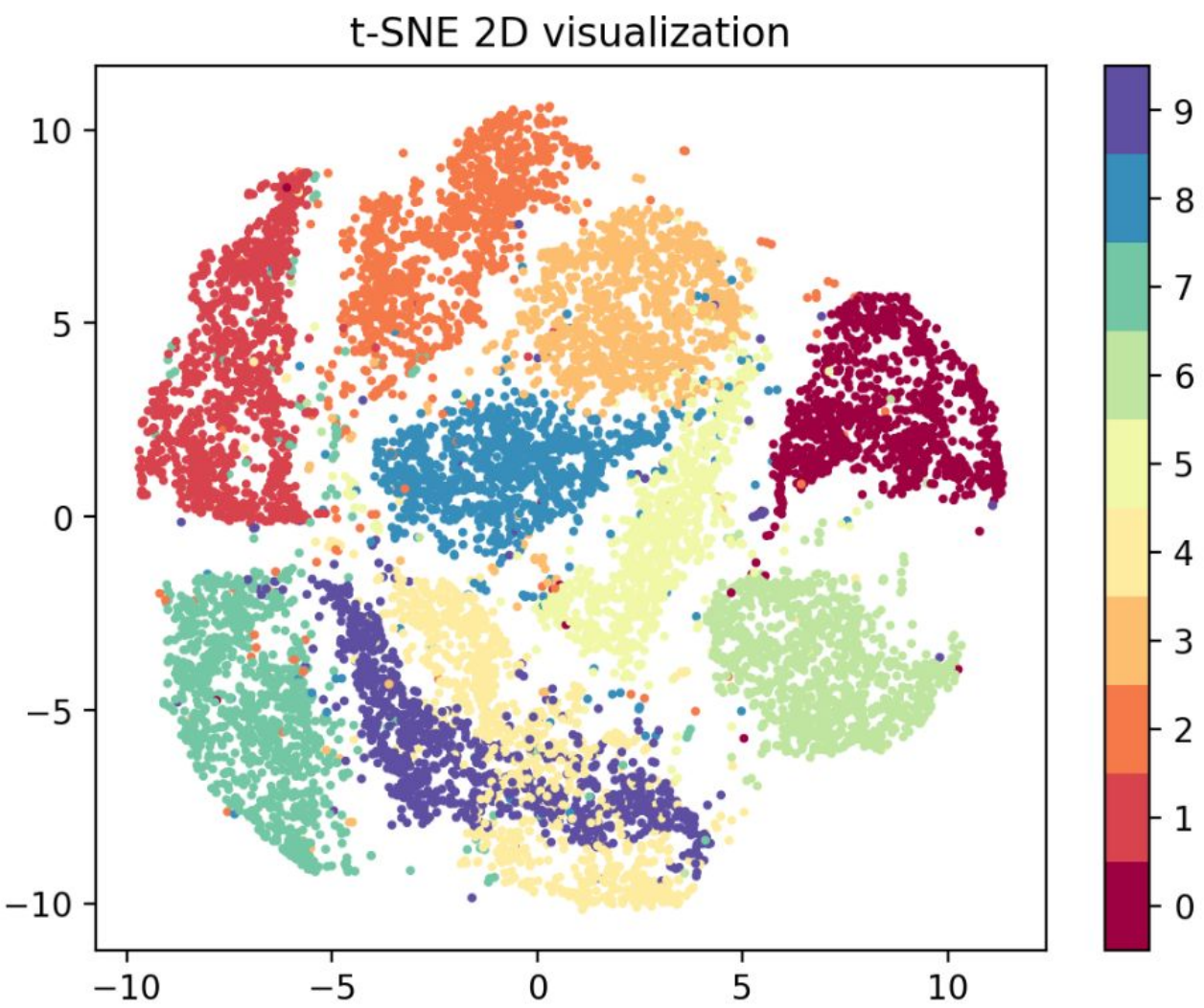
PCA 784->400 dims, recovered to 784 dims

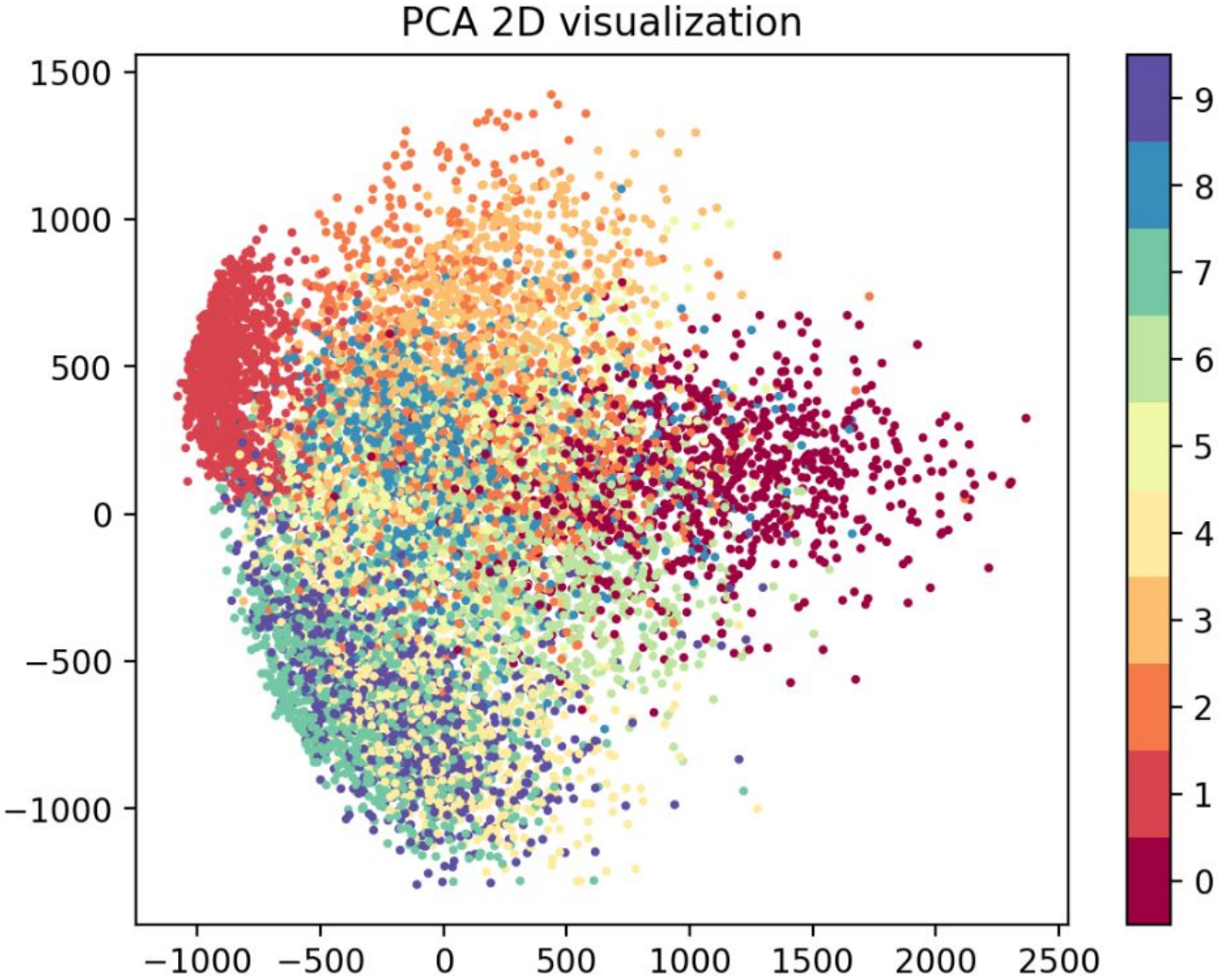
PCA 784->200 dims, recovered to 784 dims

PCA 784->100 dims, recovered to 784 dims

PCA 784->50 dims, recovered to 784 dims




t-SNE 2D visualization
10
5
0
−5
−10
−10
−5
0
5
10
9
8
7
6
5
4
3
2
1
0


PCA 2D visualization
1500
1000
500
0
−500
−1000
−1000
−500
0
500
1000
1500
2000
2500
9
8
7
6
5
4
3
2
1
0

Guidelines for Assignment Submission with Jupyter

- Step 1: Download the assignment files (e.g., .ipynb files) from Moodle
- Step 2:
  - Option 1: Open the ipynb file with **jupyter notebook**
  - Option 2: Use **Google Colab** and upload the ipynb file
- Step 4: Answer the questions in this file
- Step 5: Execute all the code blocks to print the results
- Step 6: Save and download this **executed** .ipynb file
- Step 7: Complete all your questions, put your answer files in one .zip,
- Step 8: Rename it using your uid, like 3009666000.zip
- Step 9: Submit the .zip on Moodle

Remember to save the executed notebook file!