CAP 4630 - Artificial Intelligence

Instructor: Dr. Chen Chen
Homework 5: Support Vector Machines (SVM)

Total Points: 60

Objective

Use scikit-learn to build and evaluate SVM classifiers (i.e. NO need to implement SVM from scratch, use scikit-learn) on the LFW dataset:

- Linear SVM and effect of the regularization parameter C.
- Kernelized SVM with RBF kernel; tune C and γ.
- (Optional) PCA dimension reduction before SVM.

Dataset & Splits

We use the LFW people dataset via scikit-learn: `https://scikit-learn.org/stable/auto_examples/applications/plot_face_recognition.html`.
We **restrict to identities with at least 50 images**. In code:

```
from sklearn.datasets import fetch_lfw_people
lfw = fetch_lfw_people(min_faces_per_person=50, resize=0.4)
```

The **starter notebook includes** this loading code and a reproducible, stratified **60% / 15% / 25%** train/validation/test split (fixed seed). You must reuse these splits across all comparisons.

Tasks (what to do and what to report)

1. **Linear SVM (20 pts)**
   Build a pipeline with `StandardScaler` and `SVC(kernel="linear")`. Tune C on the **validation set** (e.g., $C \in \{0.01, 0.1, 1, 10, 100\}$). Select the best C by validation accuracy. Retrain on train+val and report **test accuracy**.

2. **RBF SVM (25 pts)**
   Build a pipeline with `StandardScaler` and `SVC(kernel="rbf")`. Tune both C and γ on the **validation set** (e.g., $C \in \{0.1, 1, 10, 100\}$ and $\gamma \in \{\text{"scale"}, 10^{-3}, 10^{-2}, 10^{-1}\}$). Select best pair by validation accuracy. Retrain on train+val and report **test accuracy**.

3. **(Optional) PCA → SVM**
   Insert `PCA` before SVM (Linear and/or RBF). Choose k by a target variance ratio (e.g., 90/95/99%). Report the chosen k, validation-selected hyperparameters, and test accuracy.

4. **Short write-up (15 pts)**
   In 1–2 pages, include:
   (a) Final test accuracies for Linear SVM and RBF SVM (and PCA variants if attempted (optional));
   (b) The chosen hyperparameters (C, and γ for RBF);
   (c) Discussion of how C and γ affected performance;
   (d) (Optional) If PCA was used, effect on accuracy and runtime;
   (e) (Optional) Confusion matrices for best models.

Hints

- **Scaling:** Always scale features (e.g., `StandardScaler`) before SVM.
- **Validation:** Tune hyperparameters on the validation set only; then retrain on train+val before the final test evaluation.
- **PCA:** If you use PCA, fit PCA on the training split only; apply the same projection to val/test.
- **Efficiency:** Evaluating RBF SVM grids can take time; start with a coarse grid, then refine near the best region.

Deliverables

- A Jupyter notebook with code, results, and clear section headers.
- A short PDF write-up summarizing results and observations (1–2 pages).

You may use scikit-learn for SVM, pipelines, scaling, and PCA. Your analysis/tuning must be your own work.