Mini-Project 1: Image Classification with Caltech-101 Dataset

09/09/2025

E-mail Contact: mengyu_wang@meei.harvard.edu; tobias_elze@meei.harvard.edu

1. Warm-Up Homeworks (Optional but Recommended)

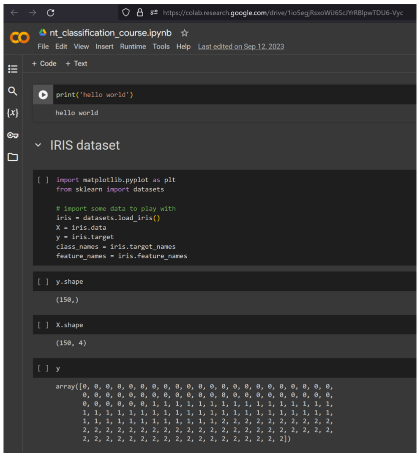
Before starting the main project, complete these warm-ups to refresh coding and ML basics. Three Jupyter Notebooks are shared as homework through the portal, including examples and to-dos. You can upload and use Google Colab.

Homework 1: Linear Regression, Unsupervised Pattern Analysis, and Dimension Reduction

Dataset: Pima Indians Diabetes Dataset Diagnostic dataset with patient features such as pregnancies, glucose, blood pressure, BMI, insulin, and age, used to predict diabetes onset.

Tasks:

- Linear Regression predict glucose levels from other patient features.
- Extra TODO: predict blood pressure instead of glucose.
- Exploratory Analysis: histograms, error computation, and visualization.
- Unsupervised Learning: k-means clustering and PCA for dimension reduction.


https://colab.research.google.com/drive/1io5egjRsxoWiJ6ScJYrRBlpwTDU6-Vyc
nt_classification_course.ipynb
File
Edit
View
Insert
Runtime
Tools
Help
Last edited on Sep 12, 2023
+ Code
+ Text
IRIS dataset

```
print('hello world')
```

```
hello world
```

```
import matplotlib.pyplot as plt
from sklearn import datasets

# import some data to play with
iris = datasets.load_iris()
X = iris.data
y = iris.target
class_names = iris.target_names
feature_names = iris.feature_names
```

```
y.shape
```

```
(150,)
```

```
X.shape
```

```
(150, 4)
```

```
y
```

```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
       2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
       2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2])
```

Homework 2: Supervised Classification

Dataset: Iris Dataset Classic dataset with 150 flower samples from 3 species (*setosa*, *versicolor*, *virginica*), each described by 4 features (sepal length/width, petal length/width).

Tasks:

- Train and evaluate classifiers: Decision Tree, Naive Bayes, SVM, and XGBoost.
- Generate confusion matrices and classification reports.
- Explore feature importance and model accuracy.

Homework 3: Supervised Classification with Deep Learning (simple CNN)

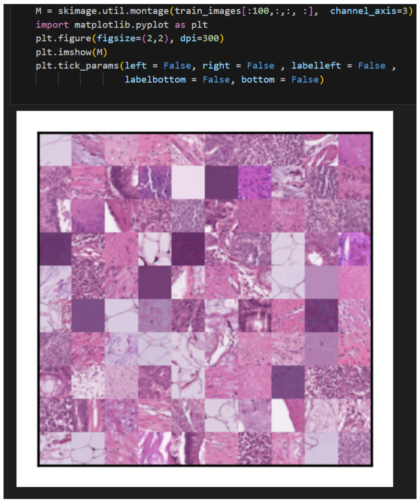
Datasets:

- PathMNIST (MedMNIST collection) histology images (9 tissue classes) for pathology classification.

- DermaMNIST (MedMNIST collection) 10,015 dermatoscopic images of pigmented skin lesions across 7 disease categories.

Tasks:

- Try classical ML methods (Decision Trees, KNN) with flattened image features.
- Train a CNN for medical image classification.
- Ablation-style experiment: compare two CNN architectures (filters = [32,32,32] vs [32,64,128]) and assess performance.
- Evaluate using confusion matrices, classification reports, and training curves.



```
M = skimage.util.montage(train_images[:100,:,:, :],  channel_axis=3)
import matplotlib.pyplot as plt
plt.figure(figsize=(2,2), dpi=300)
plt.imshow(M)
plt.tick_params(left = False, right = False , labelleft = False ,
                labelbottom = False, bottom = False)
```

2. Mini-Project 1: Image Classification with Caltech-101 Dataset

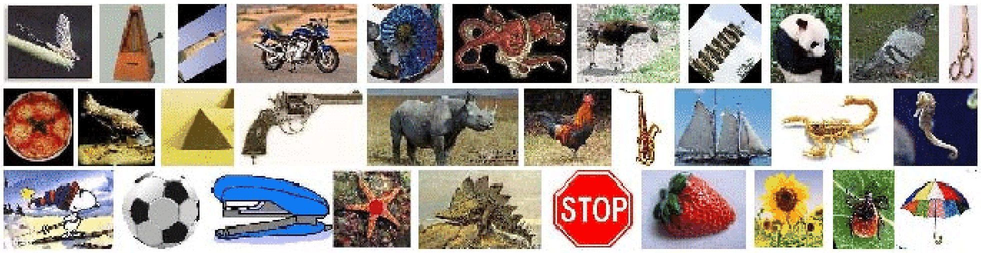
You will work with the **Caltech-101 dataset** (101 object categories, ∼9,000 images).

Your Goals

- Train and evaluate **at least three different methods** for image classification. For example ResNet, EfficientNet, ViT, and newer ones.
- Compare classical machine learning and deep learning methods.
- Write a report including:
  - Methods
  - Results (metrics + plots)
  - Observations, and/or Ablation stidues
  - Interpretations and Lessons learned
- Deliver Report, Notebooks, and/or Scripts, figures.

Dataset Preparation

- Download Caltech-101 Dataset https://www.kaggle.com/datasets/imbikramsaha/caltech-101.
- Split: Use 70% train, 15% validation, 15% test (stratified).



Evaluation Metrics

Each model must be evaluated using:

- Accuracy (overall performance).
- Per-class accuracy (to see class imbalance effects).
- Confusion matrix (visualize misclassification).
- Precision, Recall, F1-Score (macro & weighted averages).
- Top-k accuracy (optional, e.g., Top-5).

Ablation Studies

To deepen learning, run at least two small ablation experiments:

- Image size: compare 64×64 vs. 128×128
- Data augmentation: train with vs without augmentation.
- Feature extractor choice: HOG vs CNN features.
- Optimizer: SGD vs Adam for CNN.

Timeline

Week 2: Assignment released.
Week 6: Submission deadline.