Nickhil Tekwani || CS 6220 || HW3B tSNE, Feature Selection, Image HAAR Features

PROBLEM 1: tSNE dim reduction

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

from sklearn.datasets import fetch_openml, fetch_20newsgroups
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.manifold import TSNE

mnist = fetch_openml('mnist_784', version=1)
mnist_data, mnist_labels = mnist.data, mnist.target

newsgroups = fetch_20newsgroups(subset='all', remove=('headers', 'footers',
vectorizer = TfidfVectorizer(max_features=2000)  # Limit features for bette
newsgroups_data = vectorizer.fit_transform(newsgroups.data).todense()
newsgroups_labels = newsgroups.target
```

```
/usr/local/lib/python3.10/site-packages/sklearn/datasets/_openml.py:968:
FutureWarning: The default value of `parser` will change from `'liac-arf
f'` to `'auto'` in 1.4. You can set `parser='auto'` to silence this warni
ng. Therefore, an `ImportError` will be raised from 1.4 if the dataset is
dense and pandas is not installed. Note that the pandas parser may return
different data types. See the Notes Section in fetch_openml's API doc for
details.
  warn(
```

In [6]:

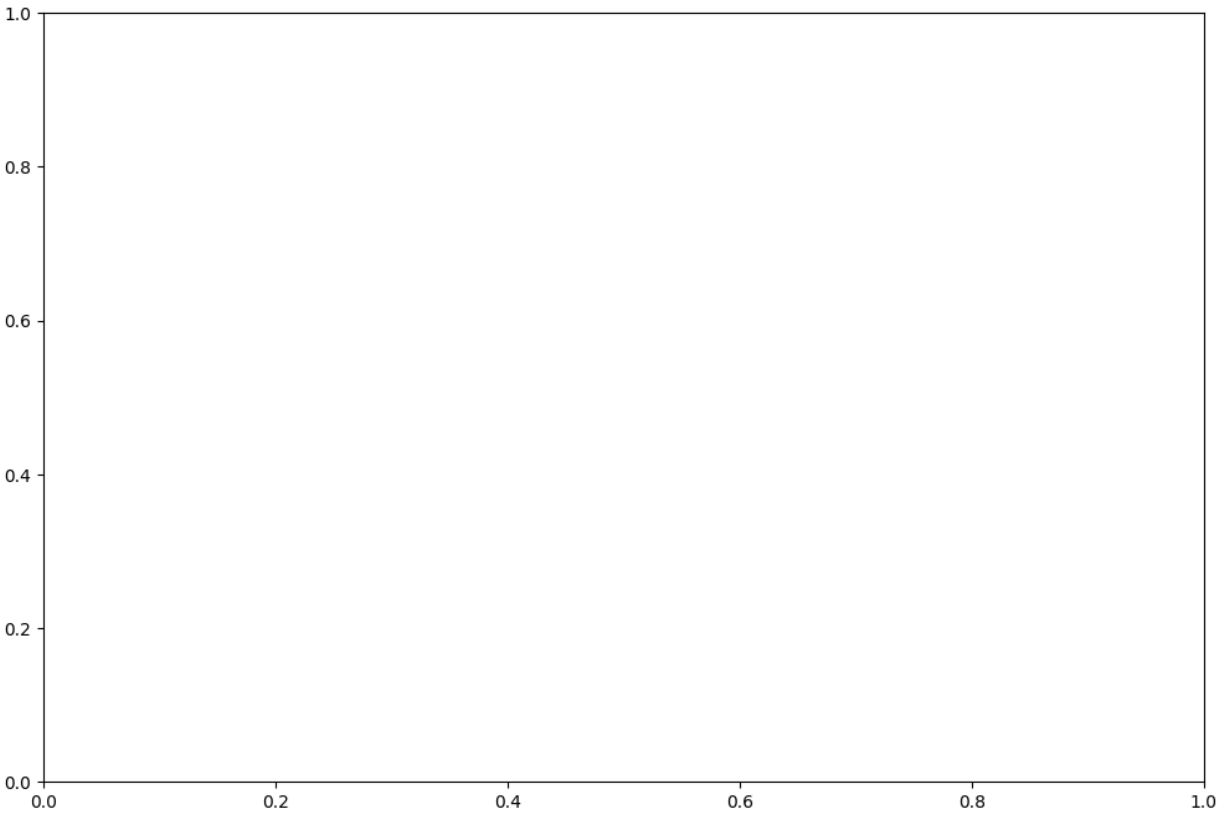
```
# run tsne
def run_tsne_and_plot(data, labels, perplexities, n_components=2, dataset_n
    for perplexity in perplexities:
        tsne = TSNE(n_components=n_components, perplexity=perplexity)
        reduced_data = tsne.fit_transform(data)

        plt.figure(figsize=(12, 8))
        scatter = plt.scatter(reduced_data[:, 0], reduced_data[:, 1], c=lab
        plt.colorbar(scatter, boundaries=np.arange(len(np.unique(labels))+1
        plt.title(f'{dataset_name} t-SNE with perplexity {perplexity}')
        plt.show()

# Run t-SNE for MNIST and 20NG:
perplexities = [5, 20, 100]
run_tsne_and_plot(mnist_data, mnist_labels, perplexities, dataset_name='MNI
run_tsne_and_plot(newsgroups_data, newsgroups_labels, perplexities, dataset
```

```
---------------------------------------------------------------------------
--
ValueError                                Traceback (most recent call las
t)
File /usr/local/lib/python3.10/site-packages/matplotlib/axes/_axes.py:443
9, in Axes._parse_scatter_color_args(c, edgecolors, kwargs, xsize, get_ne
xt_color_func)
   4438 try:  # Is 'c' acceptable as PathCollection facecolors?
-> 4439     colors = mcolors.to_rgba_array(c)
   4440 except (TypeError, ValueError) as err:

File /usr/local/lib/python3.10/site-packages/matplotlib/colors.py:487, in
to_rgba_array(c, alpha)
    486 else:
--> 487     rgba = np.array([to_rgba(cc) for cc in c])
    489 if alpha is not None:

File /usr/local/lib/python3.10/site-packages/matplotlib/colors.py:487, in
<listcomp>(.0)
    486 else:
--> 487     rgba = np.array([to_rgba(cc) for cc in c])
    489 if alpha is not None:

File /usr/local/lib/python3.10/site-packages/matplotlib/colors.py:299, in
to_rgba(c, alpha)
    298 if rgba is None:  # Suppress exception chaining of cache lookup f
ailure.
--> 299     rgba = _to_rgba_no_colorcycle(c, alpha)
    300     try:

File /usr/local/lib/python3.10/site-packages/matplotlib/colors.py:370, in
_to_rgba_no_colorcycle(c, alpha)
    369 if not (0 <= c <= 1):
--> 370     raise ValueError(
    371         f"Invalid string grayscale value {orig_c!r}. "
    372         f"Value must be within 0-1 range")
    373 return c, c, c, alpha if alpha is not None else 1.

ValueError: Invalid string grayscale value '5'. Value must be within 0-1
range

The above exception was the direct cause of the following exception:

ValueError                                Traceback (most recent call las
t)
Cell In [6], line 15
     13 # Run t-SNE for MNIST and 20NG:
     14 perplexities = [5, 20, 100]
---> 15 run_tsne_and_plot(mnist_data, mnist_labels, perplexities, dataset
_name='MNIST')
     16 run_tsne_and_plot(newsgroups_data, newsgroups_labels, perplexitie
s, dataset_name='20 Newsgroups')

Cell In [6], line 8, in run_tsne_and_plot(data, labels, perplexities, n_c
omponents, dataset_name)
      5 reduced_data = tsne.fit_transform(data)
      7 plt.figure(figsize=(12, 8))
```

```
----> 8 scatter = plt.scatter(reduced_data[:, 0], reduced_data[:, 1], c=l
abels, cmap='viridis', alpha=0.6)
      9 plt.colorbar(scatter, boundaries=np.arange(len(np.unique(labels))
+1)-0.5).set_ticks(np.arange(len(np.unique(labels))))
     10 plt.title(f'{dataset_name} t-SNE with perplexity {perplexity}')

File /usr/local/lib/python3.10/site-packages/matplotlib/pyplot.py:2862, i
n scatter(x, y, s, c, marker, cmap, norm, vmin, vmax, alpha, linewidths,
edgecolors, plotnonfinite, data, **kwargs)
   2857 @_copy_docstring_and_deprecators(Axes.scatter)
   2858 def scatter(
   2859         x, y, s=None, c=None, marker=None, cmap=None, norm=None,
   2860         vmin=None, vmax=None, alpha=None, linewidths=None, *,
   2861         edgecolors=None, plotnonfinite=False, data=None, **kwarg
s):
-> 2862     __ret = gca().scatter(
   2863         x, y, s=s, c=c, marker=marker, cmap=cmap, norm=norm,
   2864         vmin=vmin, vmax=vmax, alpha=alpha, linewidths=linewidths,
   2865         edgecolors=edgecolors, plotnonfinite=plotnonfinite,
   2866         **({"data": data} if data is not None else {}), **kwargs)
   2867     sci(__ret)
   2868     return __ret

File /usr/local/lib/python3.10/site-packages/matplotlib/__init__.py:1442,
in _preprocess_data.<locals>.inner(ax, data, *args, **kwargs)
   1439 @functools.wraps(func)
   1440 def inner(ax, *args, data=None, **kwargs):
   1441     if data is None:
-> 1442         return func(ax, *map(sanitize_sequence, args), **kwargs)
   1444     bound = new_sig.bind(ax, *args, **kwargs)
   1445     auto_label = (bound.arguments.get(label_namer)
   1446                   or bound.kwargs.get(label_namer))

File /usr/local/lib/python3.10/site-packages/matplotlib/axes/_axes.py:460
2, in Axes.scatter(self, x, y, s, c, marker, cmap, norm, vmin, vmax, alph
a, linewidths, edgecolors, plotnonfinite, **kwargs)
   4599 if edgecolors is None:
   4600     orig_edgecolor = kwargs.get('edgecolor', None)
   4601 c, colors, edgecolors = \
-> 4602     self._parse_scatter_color_args(
   4603         c, edgecolors, kwargs, x.size,
   4604         get_next_color_func=self._get_patches_for_fill.get_next_c
olor)
   4606 if plotnonfinite and colors is None:
   4607     c = np.ma.masked_invalid(c)

File /usr/local/lib/python3.10/site-packages/matplotlib/axes/_axes.py:444
8, in Axes._parse_scatter_color_args(c, edgecolors, kwargs, xsize, get_ne
xt_color_func)
   4445             raise invalid_shape_exception(c.size, xsize) from err
   4446         # Both the mapping *and* the RGBA conversion failed: pret
ty
   4447         # severe failure => one may appreciate a verbose feedbac
k.
-> 4448         raise ValueError(
   4449             f"'c' argument must be a color, a sequence of colors,
"
```

```
   4450             f"or a sequence of numbers, not {c!r}") from err
   4451 else:
   4452     if len(colors) not in (0, 1, xsize):
   4453         # NB: remember that a single color is also acceptable.
   4454         # Besides *colors* will be an empty array if c == 'none'.
```

```
ValueError: 'c' argument must be a color, a sequence of colors, or a sequ
ence of numbers, not 0        5
1         0
2         4
3         1
4         9
         ..
69995     2
69996     3
69997     4
69998     5
69999     6
Name: class, Length: 70000, dtype: category
Categories (10, object): ['0', '1', '2', '3', ..., '6', '7', '8', '9']
```



In [ ]:

```
# visualize in 3d
from mpl_toolkits.mplot3d import Axes3D

def run_tsne_and_plot_3d(data, labels, perplexities, dataset_name=''):
    for perplexity in perplexities:
        tsne = TSNE(n_components=3, perplexity=perplexity)
        reduced_data = tsne.fit_transform(data)

        fig = plt.figure(figsize=(12, 8))
        ax = fig.add_subplot(111, projection='3d')
        scatter = ax.scatter(reduced_data[:, 0], reduced_data[:, 1], reduce
        plt.colorbar(scatter, boundaries=np.arange(len(np.unique(labels))+1
        plt.title(f'{dataset_name} t-SNE 3D with perplexity {perplexity}')
        plt.show()

# For 3D visualization:
run_tsne_and_plot_3d(mnist_data, mnist_labels, perplexities, dataset_name='
run_tsne_and_plot_3d(newsgroups_data, newsgroups_labels, perplexities, data
```