Tutorial for Assignment 2

COMP3314
Machine Learning

Lao Yixing

Guidelines for submitting Homework 2

(Due: 23:59 Nov 13, 2022)

- Step 1: Download the files (3 Jupyter notebooks) from Moodle.
- Step 2: Visit https://colab.research.google.com/
- Step 3: Choose "Upload notebook" to upload your notebooks.
- Step 4: Answer the questions in this file.
- Step 5: Execute all the code blocks to print the results.
- Step 6: Save the **executed notebooks** in .ipynb file
- Step 7: Put the finished .ipynb files in one .zip, and name it using your uid, like 3009666.zip
- Step 8: Submit the .zip on Moodle.

Review of Homework 1

- HW1 solution is available on Moodle now
- Contact TAs if you have any questions

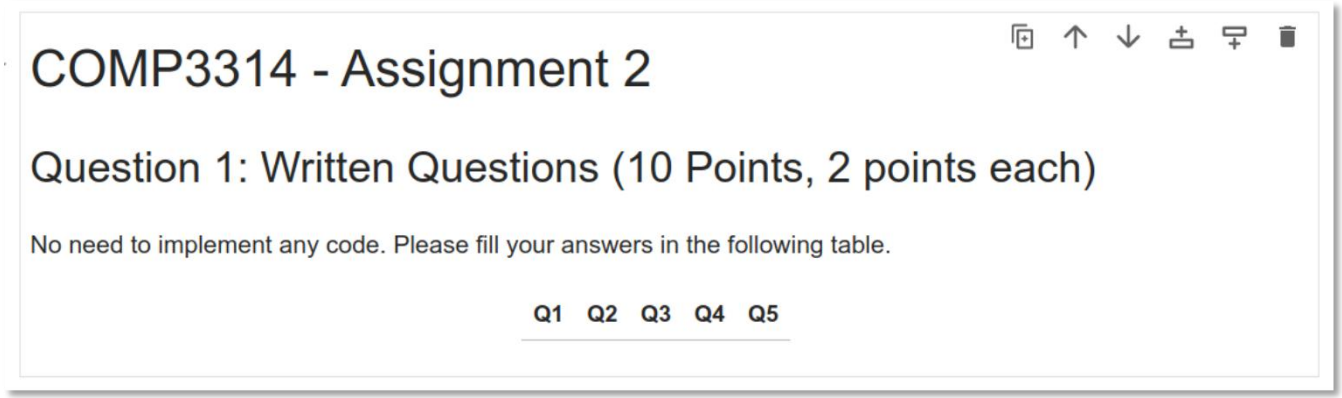
Homework 2 overview

- Total: 50 points
  - HW2-Q1: Written questions (10 points)
  - HW2-Q2: Spam classifier (20 points)
  - HW2-Q3: MNIST dimensionality reduction (20 points)

HW2-Q1 Written Questions

HW2-Q1 Written questions (10 points)

- These are all multiple choice questions
- Simply fill in the table with your answers
- Review the lecture slides carefully



HW2-Q2 Spam Classifier

HW2-Q2 Spam classifier

- Total: 20 points
  - Step 1: Download dataset
  - Step 2: Feature extraction (5 points)
    - Your code needed
  - Step 3: Train a spam classifier (5 points)
    - Your code needed
  - Step 4: Eval your classifier (5 points)
    - Your code needed
  - Step 5: Ensemble of classifiers (5 points)
    - Your code needed

```
To: <VALUED_CUSTOMERS@mandark.labs.netnoteinc.com>
From: "Jason Howard" <erienw9@wam.co.za>
Subject: Free info.  Start your Own Internet Consulting Business NTSJ
Date: Thu, 16 May 2002 07:43:37 -1700
MIME-Version: 1.0
Content-Type: text/plain; charset="Windows-1252"
Reply-To: erienw9@wam.co.za
X-Mailer: Mozilla 4.73 [en] (Win98; U)
X-Keywords:
Content-Transfer-Encoding: 7bit


Did you know 4 of the country's 10 richest people never graduated from college?

They had the courage to dream, and the wisdom to
take advantage of opportunities.

Do you have the Courage and the Wisdom to
change YOUR life?

YOU DESERVE SUCCESS!

Checking out this web site is free, and it could pay off in the form of
a dramatically improved lifestyle for you and your loved ones.

You will never know unless you check it out NOW!

Invest JUST ONE minute to check out this website right now.

http://www.22freewayexit2284.net/index7.html



If you would like to be removed from all future mailings just
send and email to erienw3943@freemail.hu
```

Sample spam

```
On Tue, Aug 06, 2002 at 12:56:17PM +0100, kevin lyda mentioned:
> sorry, i missed this.  redhat supplies something similar called kickstart
> (guess who inspired them?).  pc hardware is dumb, so you'll need to use
> a floppy.  otoh, every jumpstart config i've seen required rarp plus
> plugging the new box's ethernet+ip into a file.  a kickstart boot can
> just use a dhcp server.

 I've just been re-aquainted with the Jumpstart stuff after a long
absence.

 There is a nice need 'add_install_client' script that you feed the
archtecture, ethernet address & ip to, and it'll setup everything from
RARP to Bootparams for you. Very simple.

 This script takes a -d option, to boot via DHCP also. On the negative
side, Sun's terminal handling leaves a lot to be desired - it won't work
properly on a Wyse 120+ for instance, no matter what emulation mode the
Wyse is trying to do.

 To do PC netbooting properly, you need an motherbard with a PXE BIOS.
Then you are flying.

 Heh, how hard would it be to get a PC with an OpenBoot prom ?

Kate


--
Irish Linux Users' Group: ilug@linux.ie
http://www.linux.ie/mailman/listinfo/ilug for (un)subscription information.
List maintainer: listmaster@linux.ie
```

Sample non-spam

HW2-Q2 Data cleanning

- Provided
  - Remove header
  - URL to word
- Your code
  - To lower case
  - Num to word
  - Remove punctuation
  - Feel free to add more

Before:

```
print(X)
```

```
From www-data@mail.virtualed.org  Tue Jul 30 22:15:36 2002
Return-Path: <www-data@mail.virtualed.org>
Delivered-To: yyyy@localhost.netnoteinc.com
Received: from localhost (localhost [127.0.0.1])
        by phobos.labs.netnoteinc.com (Postfix) with ESMTP id 109CC4406D
        for <jm@localhost>; Tue, 30 Jul 2002 17:15:36 -0400 (EDT)
Received: from mandark.labs.netnoteinc.com [213.105.180.140]
        by localhost with POP3 (fetchmail-5.9.0)
        for jm@localhost (single-drop); Tue, 30 Jul 2002 22:15:36 +0100 (IST)
Received: from mail.virtualed.org (root@ns1.virtualimpact.net [209.114.200.97])
        by mandark.labs.netnoteinc.com (8.11.6/8.11.6) with ESMTP id g6UL8Jp12675
        for <jm@netnoteinc.com>; Tue, 30 Jul 2002 22:08:20 +0100
Received: (from www-data@localhost)
        by mail.virtualed.org (8.11.1/8.9.3/Debian 8.9.3-21) id g6UL7uI30453;
        Tue, 30 Jul 2002 17:07:56 -0400
Date: Tue, 30 Jul 2002 17:07:56 -0400
Message-Id: <200207302107.g6UL7uI30453@mail.virtualed.org>
To: yyyy@netnoteinc.com
From: suddenlysusan@Stoolmail.zzn.com ()
Subject: Best Price on the netf5f8m1

 (suddenlysusan@Stoolmail.zzn.com) on Tuesday, July 30, 2002 at 17:07:56
: Why Spend upwards of $4000 on a DVD Burner when we will show you an alternative that will d
o the exact same thing for just a fraction of the cost? Copy your DVD's NOW. Best Price on th
e net. Click here: http://002@www.dvdcopyxp.com/cgi-bin/enter.cgi?marketing_id=dcx009 Click t
o remove http://003@www.spambites.com/cgi-bin/enter.cgi?spambytes_id=100115
```

After:

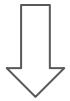
```
print(EmailCleaner().fit([X]).transform([X]))
```

```
['suddenlysusanstoolmailzzncom on tuesday july 30 NUM at 170756  why spend upwards of 4000 on
a dvd burner when we will show you an alternative that will do the exact same thing for just
a fraction of the cost copy your dvds now best price on the net click here URL click to remov
e URL']
```

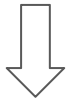
HW2-Q2 Feature extraction (text to vector)

```
print(EmailCleaner().fit([X]).transform([X]))
```

```
['suddenlysusanstoolmailzzncom on tuesday july 30 NUM at 170756  why spend upwards of 4000 on
a dvd burner when we will show you an alternative that will do the exact same thing for just
a fraction of the cost copy your dvds now best price on the net click here URL click to remov
e URL']
```



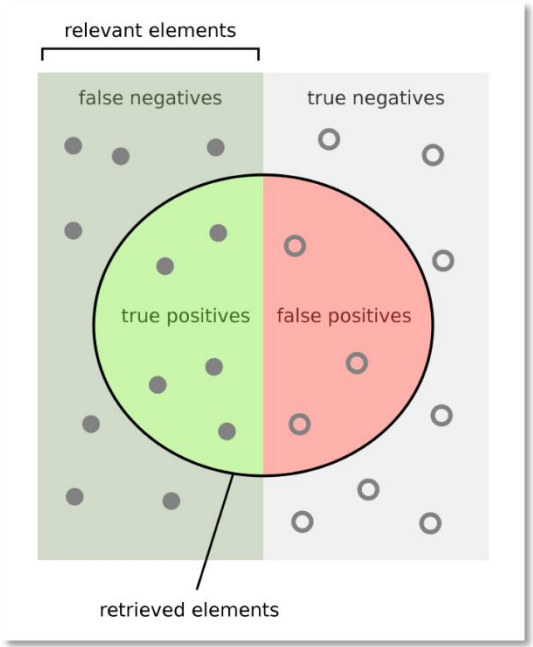
CountVectorizer (constructed from train set)

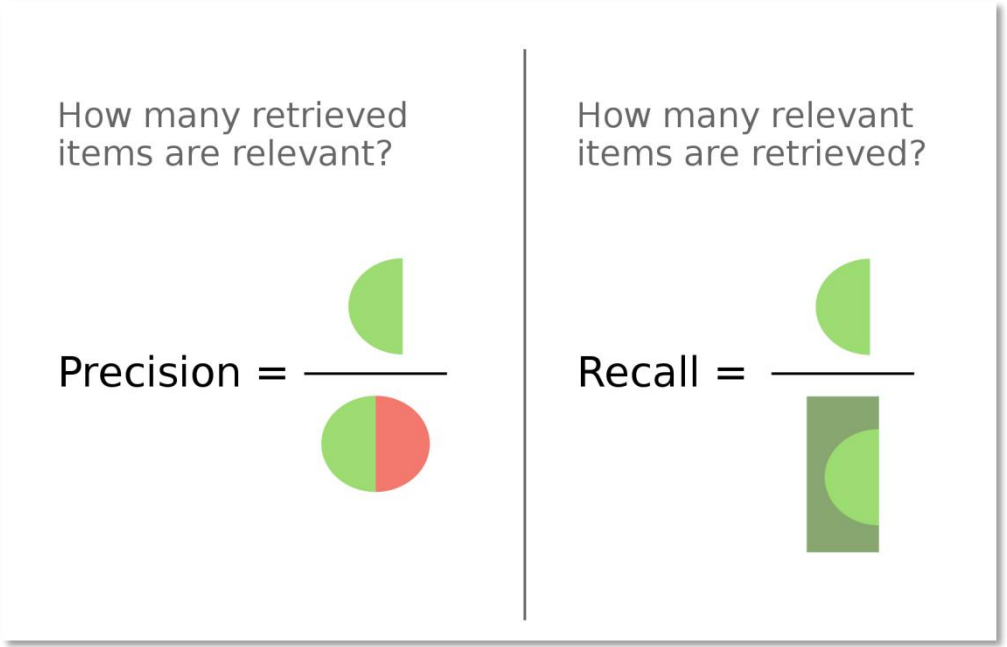


Feature vector: [..., 0, 0, 2, 0, 1, ...]

HW2-Q2 Evaluation metrics

- You need to report 2 metrics: precision and recall.





https://en.wikipedia.org/wiki/Precision_and_recall#/media/File:Precisionrecall.svg

HW2-Q2 Ensemble of classifiers

- Implement more classifiers
- Use voting for final predictions
- Compare your ensemble classifier's performance with the individual classifier

HW2-Q3 MNIST Dimensionality Reduction

HW2-Q3 MNIST dimensionality reduction

- Total: 20 points
  - Step 1: Download dataset
  - Step 2: Visualize digits (5 points)
    - Your code needed
  - Step 3: PCA projection and recovery (5 points)
    - Your code needed
  - Step 4: t-SNE 2D visualization (5 points)
    - Your code needed
  - Step 5: PCA 2D visualization (5 points)
    - Your code needed

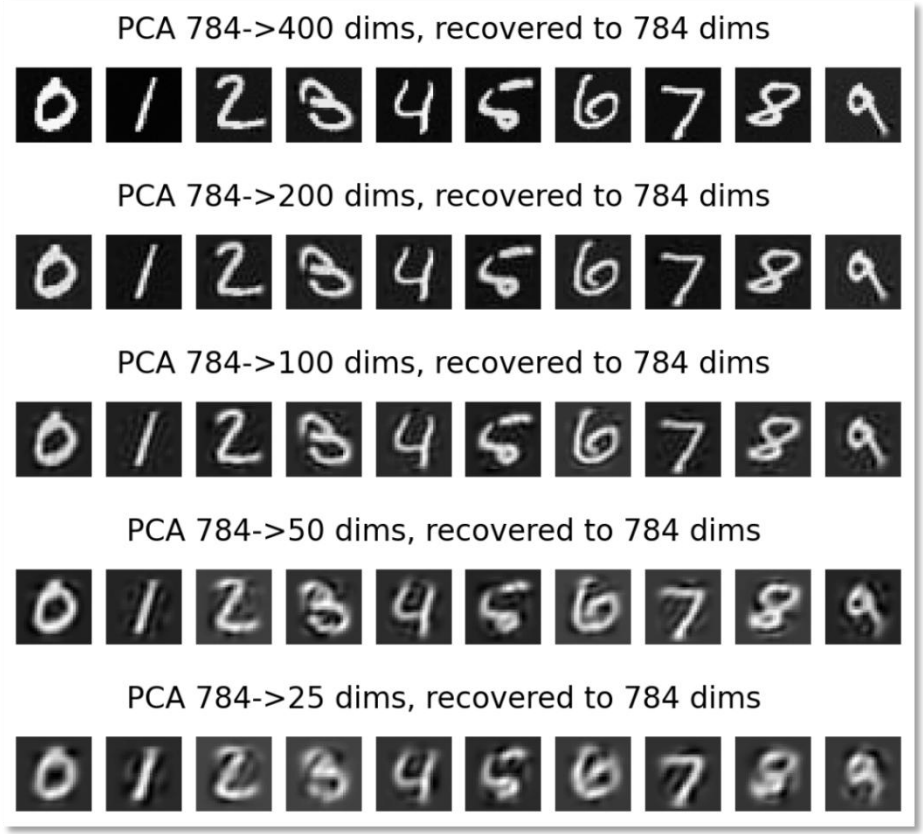
HW2-Q3 Visualize digits

- We are only going to use the test set of the MNIST dataset
  - This homework does not require training
- MNIST digits are 28x28 grayscale images
  - Reshape to proper shape for visualization
  - Reshape to proper shape for training/testing
- For each digit 0-9, pick one to visualize

Example output

HW2-Q3 PCA projection and recovery

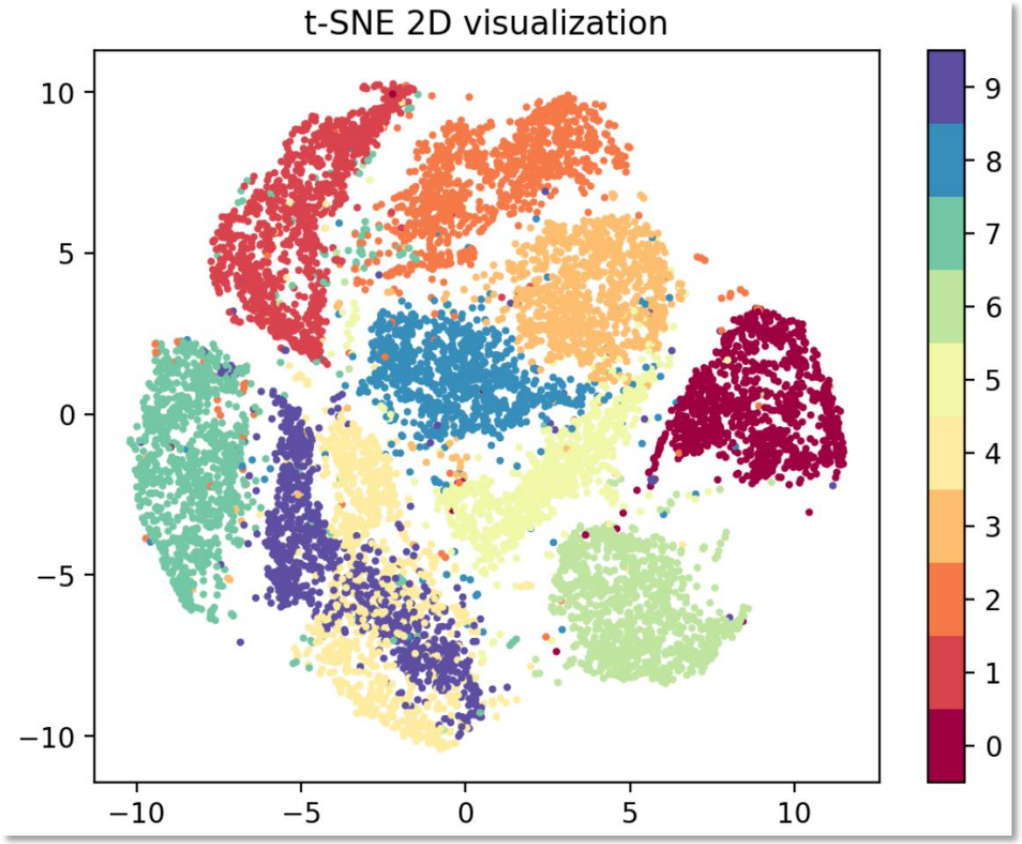
- This visualization shows how much information is lost during PCA projection to lower dimensions.
- You need to figure out how to "reverse" the PCA transformation.
- Hint: reuse the same visualization code you implemented in step 2. Try avoiding code duplication.



Example output

HW2-Q3 t-SNE visualization

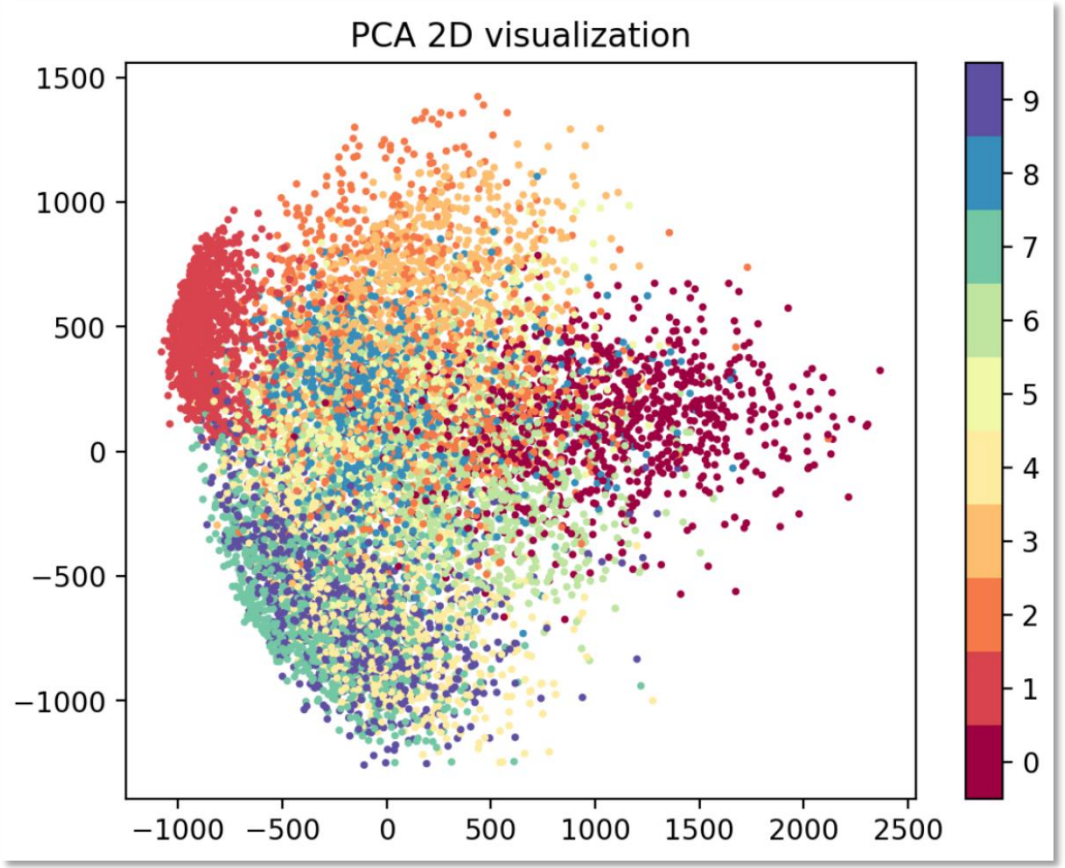
- Use t-SNE from sklearn to embed the dataset into 2D.
- Your visualization shall clearly show that the same digits are organized in clusters, after t-SNE embedding.



Example output

HW2-Q3 PCA visualization

- Finally, we'll compare PCA and t-SNE for their dimensionality reduction to 2D.
- Compare PCA with t-SNE, which one is better?
- Hint: reuse the same code you implemented for t-SNE visualization. Try avoiding code duplication.



Example output

Q & A

Further Questions

- Xi Chen (email: xchen2@cs.hku.hk)
- Office: HW-RSC or zoom (https://hku.zoom.us/my/xavier.xichen)
- Consultation hours : Wednesday, 10:00 am - 12:00 pm

- Yixing Lao (email: laoyx@connect.hku.hk)
- Office: HW-RSC or zoom (https://hku.zoom.us/my/laoyixing)
- Consultation hours: Tuesday, 2:00 pm - 4:00 pm
- Please send an email before our meeting