CSC321. Data Collection for Project 3

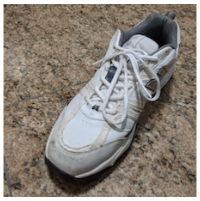
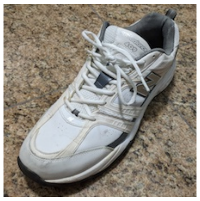
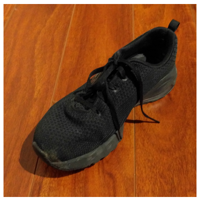
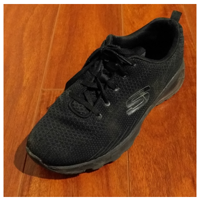
Due February 24 9:00pm

So far, we have worked with data sets that have been collected, cleaned, and curated by machine learning researchers and practitioners. In the tutorial examples, we used classic datasets like MNIST, which are often used as toy examples, both by students and by researchers testing a new machine learning model. In projects 1 and 2 our datasets were cleaned and curated by teaching assistants.

In the real world, getting a clean data set is never that easy. Over half the battle of applying machine learning is finding, gathering, cleaning, and formatting your data set. Part of this project involves gaining experience gathering your own data set, so you can understand the challenges involved in the data cleaning process.

The Data

We will collect and clean photos of left shoes and right shoes. We will each take three pictures of their left shoes, and three pictures of their right shoes. Then, we will resize these images to 224 × 224 pixels, placing the shoe in the center. We will place the shoes in a slight angle, and always have the front of the shoe point forward. Here are some examples of the kind of images we should obtain:

Right Shoe	Left Shoe
	
	

This step needs to be done **individually**. Even if you intend to work with a partner for project 3, each person needs to submit the data they collected. The images you submit will be graded separately.

Data Collection

Start by picking three pairs of shoes. You should use your own shoes, but if you do not own at least three pairs of shoes, you can use shoes of other people who are *not* in this class. (Note: If there are reasons why finding shoes might be difficult for you, please talk to Lisa.)

Then, take pictures of each of these left and right shoes. Place a shoe on the floor (or a large flat surface). Point the front of your shoe towards the camera, and angle your shoe roughly 45 degrees to the left. Ensure that the background/floor is *uniform* (the texture doesn't change) and *uncluttered* (there are no other objects in the picture). Avoid glare, large shadows, and other artifacts.

Transfer your images to your computer for cleaning. Crop a square region of the image so that the shoe is centered, then resize the image to 224 × 224 pixels. Make sure that you do not "stretch" the shoe horizontally or vertically, and not change the aspect ratio.

Resizing images

You can resize your images using any application you like. Here are some example applications that you could like:

- On a **Mac**, use Preview. Hold down CMD + Shift will keep square aspect ratio while selecting the hand area, and use CMD + K to crop. Resize to 224x224 pixels.
- On **Windows 10**, Use `Photos` app to edit and crop the image and keep the aspect ratio as Square. Use `Paint` to resize the image to the final image size of 224x224 pixels.
- On **Linux** you can use GIMP, imagemagick, or other tools of your choosing.
- You Can also use online tools such as http://picresize.com

All the above steps are illustrative only. You need not follow these steps but following these will ensure that you churn out good quality dataset. You will be judged based on the quality of the images alone.

Do not alter your photos in any other way. For example, do not use photoshop, paint, or GIMP to erase the background or erase objects. The images should be as "natural" as possible.

Finally, save your images as jpg files called `left1.jpg`, `right1.jpg`, `left2.jpg`, `right2.jpg`, `left3.jpg`, `right3.jpg`. The shoes in the corresponding images must match. That is, `left1.jpg` and `right1.jpg` should be from the same pair of shoes.

Metadata

Along with these images, we ask you to submit a metadata file `meta.txt`. This file gives us some more information about the shoes for an experiment on algorithmic fairness that we want the class to do. In particular, we're asking you to label whether each shoe was marketed for men or for women.

We're also asking for your permission to reuse the images that you collect for educational purposes in the future. No personally identifiable information will be collected; only the images.

The metadata file should be formatted like this:

```
permission: yes or no (reusing images for educational purposes)
shoe1: w (for women) or m (for men)
shoe2: w (for women) or m (for men)
shoe3: w (for women) or m (for men)
```

Here is an example metadata file:

```
permission: yes
shoe1: w
shoe2: w
shoe3: m
```

Please make sure that your metadata file `meta.txt` is in **plain text**, and only contains those 4 lines of text. If you use a software like Microsoft Word, you won't produce a plain text file.

Submission and Grading

Submit the six images plus the metadata file on Markus. Your TA will be anonymizing and combining the images that everyone submits. We will announce when the combined data set will be available for download.

The data collection step must be done individually, and accounts for 10% of your project 3 grade. The grading is based on:

1. (25%) Have you submitted a properly-formatted metadata?
2. (25%) Have you submitted 6 pictures of shoes?
3. (50%) Have you submitted shoes that are well formatted?

Examples of Unacceptable Images

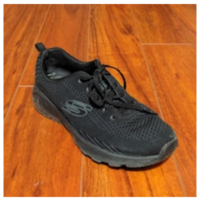
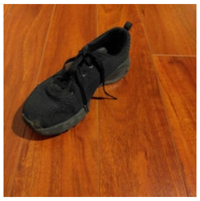
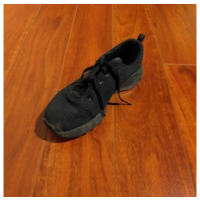
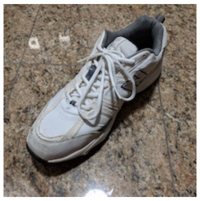
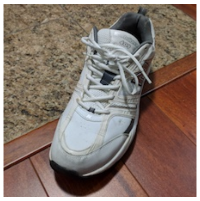
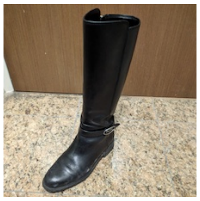
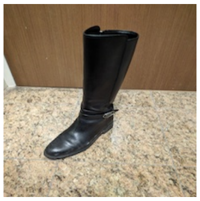
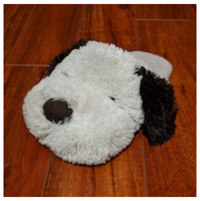
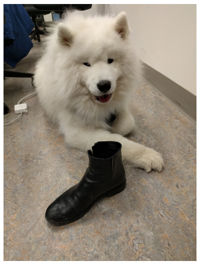
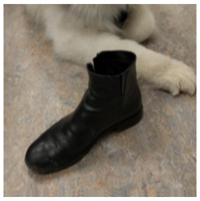
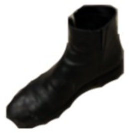
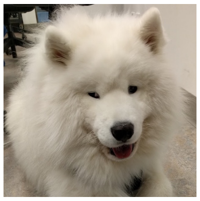
Image	Problem
	The shoe is not angled toward the left.
	The shoe is not centered in the image.
	The shoe is too small.
	There is glare in the top left and bottom left of the image.
	The background is not uniform.
	The background is not uniform. Avoid tall boots so the shoe can fit into a square image.
	The image is stretched horizontally, and the background is also still not uniform.
	Avoid slippers, especially those that don't look like shoes.

Image	Problem
	The image is not cropped and resized to 224 × 224 pixels, and there is a dog.
	There are still foreign objects in the background.
	Dog was digitally removed. Please don’t alter the image beyond cropping and resizing.
	This is still a dog.