Homework 4 Written Questions

Template Instructions

This document is a template with specific answer regions and a fixed number of pages. Given large class sizes and limited TA time, the template helps the course staff to grade efficiently and still focus on the content of your submissions. Please help us in this task:

- Make this document anonymous.
- Questions are in the orange boxes. Provide answers in the green boxes.
- Use the footer to check for correct page alignment.
- **Do NOT remove the answer box.**
- **Do NOT change the size of the answer box.**
- **Extra pages are not permitted unless otherwise specified.**
- **Template edits or page misalignment will lead to a 10 point deduction.**

Gradescope Submission

- Compile this document to a PDF and submit it to Gradescope.
- Pages will be automatically assigned to the right questions on Gradescope.

Declaration of Generative AI Use

Reminder of Course Policy

- The use of GenAI tools (e.g., ChatGPT, Grammarly, Bard) for completing any part of this course is discouraged.
- Using these tools is not needed to be successful in the class and could be detrimental to your learning experience.
- If you use them, you must cite the tool you used and explain how you used it.
- If you do not cite the tool, it is an academic code violation.
- We will be using special tools to detect cases of unattributed GenAI use.

Student Declaration

Have you used generative AI tools to complete this assignment:

YES ☐ NO ■

If you answered YES above, describe what tools you used and what parts of the assignment you used them for below:

Example: I used ChatGPT to debug my convolution implementation

This Homework

- 5 questions **[10 + 12 + 14 + 6 + 10 = 52 points]**.
- Include code, images, and equations where appropriate.

Q1: [10 points]

(a) **[4 points]**

Define these two types of error in machine learning and how to measure them:

(i) **[2 points]** Bias **[2–3 sentences]**

TODO: Your answer for (a) (i) here

(ii) **[2 points]** Variance **[2–3 sentences]**

TODO: Your answer for (a) (ii) here

(b) **[4 points]**

Define each terms in the context of evaluating a classifier, and describe how to measure each one.

(i) **[2 points]** Overfitting **[2–3 sentences]**

TODO: Your answer for (b) (i) here

(ii) **[2 points]** Underfitting **[2–3 sentences]**

TODO: Your answer for (b) (ii) here

(c) **[2 points]**

How might we mitigate high bias? How might we mitigate high variance? **[2–4 sentences]**

TODO: Your answer for (c) here

Q2: **[12 points]** Suppose we are given a scene classification task and a dataset. Suppose also that we pick bag of words as the image representation for our classifier. 'Bag of words' originated in text processing; for computer vision, a 'word' is a visual feature.

The bag of words representation handles the spatial layout of information in a way that can be an advantage or a disadvantage in different cases.

(a) **[2 points]**

Describe two example scenarios when dealing with images: one that illustrates an advantage of the bag of words representation, and another that shows a disadvantage of the bag of words representation. **[5–6 sentences]**

TODO: Your answer for (a) here

(b) **[1 point]**

Describe additional features or other information that you can add to your representation that might overcome the disadvantage of bag of words you stated above. **[2–4 sentences]**

TODO: Your answer for (b) here

Next, we define a feature transform for each image, and pick SIFT extracted in a sparse grid pattern. To reduce the size of our problem, we create a dictionary of visual words using k-means clustering upon all the features extracted from our images.

Examining the SIFT features generated from our training database tells us that many features are almost equidistant from two or more visual words.

(c) **[1 points]**

Why might this affect classification accuracy? **[2–4 sentences]**

TODO: Your answer for (c) here

(d) **[2 points]**

Given the situation, describe *two* methods to improve classification accuracy, and explain why they would help. (*These can be for k-means, or otherwise.*) **[4–6 sentences]**

TODO: Your answer for (d) here

(e) **[4 points]**

How might we determine whether our classifier is a good model? Discuss technical solutions. **[5–6 sentences]**

TODO: Your answer for (e) here

(f) **[2 points]**

After training, suppose we find that our model shows bad performance on our test data. When is it appropriate to pick a new test dataset for which our model achieves better performance? **[3–4 sentences]**

TODO: Your answer for (f) here

Q3: [14 points] The performance of a machine learning system is determined by the data we train it upon and the feature transforms and classifier optimizations we execute. Our perception of that performance is determined by how we evaluate and deploy machine learning systems. Performance limitations from the bias we discussed in Q1 is a *learning bias*, and there are many other kinds of bias throughout the machine learning life cycle.

Another is *evaluation bias*: that evaluation using overall accuracy can mask poor performance in specific subgroups. Buolamwini and Gebru's 2018 paper Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification found that Microsoft's Face API trained gender classifier achieved 94% overall accuracy, with 100% accuracy on lighter-skinned male faces but only 79.2% accuracy on darker-skinned female faces. In response to the report, Microsoft significantly improved their classifier's performance by making changes across their machine learning system. This occurred through "expanded and revised training and benchmark datasets, new data collection efforts to further improve the training data by focusing specifically on skin tone, gender, and age, and an improved classifier to produce higher precision results."

Suresh and Guttag describe seven kinds of bias in machine learning system life cycles. To help you gain the language to talk about these issues, please read their article up to Section 4 (log-in through Brown to ACM Digital Library or PDF here) [15 minutes].

In this homework, we will train a scene classifier using data from Lazebnik et al. 2006. Please review the data to check for potential biases: look at images in the data/train and data/test directories and consider their class labels.

For example, if the street images in the dataset were used for training today, new buildings or the variation in the styles of cars and pedestrian clothings might bias the machine learning system. This could be a data *historical bias* caused by out-of-date real-world sampling, among other potential biases.

(a) **[6 points]**

Please list at least two potential issues in the dataset and name their biases following Suresh and Guttag. There may be more than one bias per issue. For each issue, please describe a potential consequence for an application that trained with this data. Is there one bias that seems to be more relevant to the dataset? **[2–3 sentences for each]**

TODO: Your answer here

Dataset curators might merge multiple sources to fill in gaps in sampling a data distribution. Upon limitations in Microsoft's Face API being revealed, its engineers were required to "expand and revise training datasets"—perhaps in a hurry and under pressure.

Please find additional scene data to add to the Lazebnik et al. 2006 dataset to reduce one of your issues. Here are two places to find datasets: Google Dataset Search and Kaggle.

(b) **[2 points]**

Find a dataset and provide a URL to it. How was this new data collected? Is it clear how this new data was collected? If not, what might be helpful to know? (Recall some of the values in our ethics resource) Additionally, given one of your issues identified in (a), describe how the new data addresses it. **[4–5 sentences]**

TODO: Your answer to (b) here

Microsoft's engineers also conducted "new data collection efforts." With many potential pitfalls, data collection can be a daunting task. One approach that researchers and companies have used to ease cost and time investment is web scraping, which downloads data across many websites to more easily create large datasets.

At the same time, the 15-scene dataset was collected using a combination of personal images, licensed photo collections, and web scraping. Web scraping has come under increased scrutiny as computer vision systems have been deployed in real-world applications, because the technical capability to download publicly-accessible data does not imply consent for the use of that data. In 2021, France found that Clearview AI, a company that operates a facial recognition platform for law enforcement, violated privacy laws by scraping 10 billion images of people's faces from Facebook, YouTube, and other websites without consent.

(c) **[6 points]**

Suppose you were a machine learning engineer tasked with collecting new scene data yourself. You are not given clear specifications as to how you collect the images. Name three stakeholders you would engage to increase your success and explain why. (The ethics primer may also be helpful here) **[6–8 sentences]**

TODO: Your answer to (c) here

Q4: [6 points] Given a linear classifier such as SVM which separates two classes (binary decision), how might we use multiple linear classifiers to create a new classifier which separates k classes?

Below, we provide simplified pseudocode for a linear classifier. It trains a model on a training set, and then classifies a new test example into one of two classes. Please edit the pseudo-code to convert this into a multi-class classifier.

Note: A more efficient software application would separate the classifier training and testing into two different functions so that the model could be reused without retraining. Feel free to ignore this for now.

You can take either the one vs. all (or one vs. others) approach or the one vs. one approach; please declare which approach you take.

TODO: Select the implementation you chose.

One vs One	☐
One vs Many	☐

Be aware that:

1. The input labels in the multi-class case are different, and you will need to match the expected label input for the `train_linear_classifier` function.
2. You need to make a new decision on how to aggregate or decide on the most confident prediction.

```
# Inputs
#   train_feats: N x d matrix of N features each d descriptor
                                    long
#   train_labels: N x 1 array containing values of either -1
#                (class 0) or 1 (class 1)
#   test_feat: 1 x d image for which we wish to predict a label
#
#   Outputs: -1 (class 0) or 1 (class 1)
#
# Inputs:
#    As before, except
#    train_labels: N x 1 array of class label integers from 0 to
                                    k-1
# Outputs:
#    A class label integer from 0 to k-1
#

# TODO: Turn this into a multi-class classifier for k classes.
def classify(train_feats, train_labels, test_feat)
    # Train classification hyperplane
    weights, bias = train_linear_classifier(train_feats,
                                        train_label)
    # Compute distance from hyperplane
    test_score = weights * test_feats + bias

    return 1 if test_score > 0 else -1
```

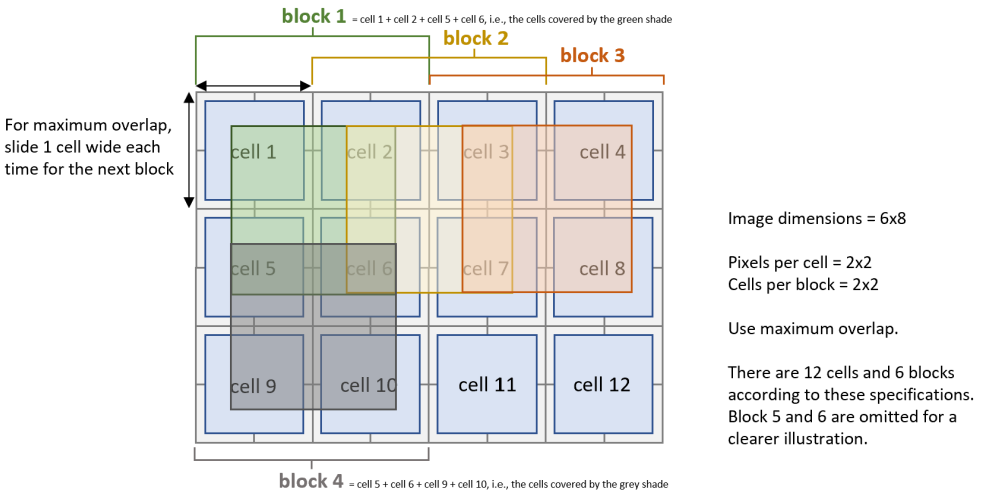
```
        # Extra page if you need it
```

Only Q4 should be on this page

Q5: [10 points] In this homework, we will use a feature descriptor called HOG—'Histogram of Oriented Gradients'. As its name implies, it works similarly to SIFT. In classification, we might extract many HOG features across the entire image (not just at corners) to create visual words.

HOG creates one feature descriptor per image 'block'. Each block is split into 'cells' covering pixels. HOG outputs a 9-bin histogram of oriented gradients per cell. We append these together to obtain the feature descriptor for each block. As a result, if we have (z, z) cells per block, the feature descriptor for each block will be of size $z \times z \times 9$. In other words, computing HOG over the whole image produces a matrix where each row is a descriptor.

Blocks can overlap as displayed in the diagram below.



(a) **[6 points]**

Given a 72×72 image, calculate the number of cells, blocks, and feature descriptor size that will occur when we extract one HOG feature with the following parameters using maximum overlap between blocks.

(i) **[3 points]** Scenario 1: Pixels per cell = 4×4, cells per block = 4×4

1. **[1 points]** Number of cells:

TODO: Your answer for (a) (i) (1) here

2. **[1 points]** Number of blocks:

TODO: Your answer for (a) (i) (2) here

3. **[1 points]** Dimensions of resulting single feature descriptor for the whole image:

   TODO: Your answer for (a) (i) (3) here

(ii) **[3 points]** Scenario 2: Pixels per cell = 8×8, cells per block = 2×2.

1. **[1 points]** Number of cells:

   TODO: Your answer for (a) (ii) (1) here

2. **[1 points]** Number of blocks:

   TODO: Your answer for (a) (ii) (2) here

3. **[1 points]** Dimensions of resulting single feature descriptor for the whole image:

   TODO: Your answer for (a) (ii) (3) here

(b) **[4 points]** When using HOG, the parameters such as pixels per cell and cells per block impact the resulting feature descriptor and so our performance on a classification task.

What are the pros and cons of the two parameter combinations? Which might you expect to have better performance? **[3–6 sentences]**

TODO: Your answer for (b) here

Feedback? (Optional)

Please help us make the course better. If you have any feedback for this assignment, we'd love to hear it!