Homework 0 (50 points)

Kathleen McKeown, Fall 2019
COMS W4705: Natural Language Processing

Due 9/7/19 at 10:00am

This is a preliminary assignment that must be completed in order to be considered for enrollment in the class. If you are already enrolled, failure to complete this assignment will result in removal from the course. The goals are twofold: first, to ensure that your environment is set up correctly for future assignments; and second, to check that you remember the basics from other fields that you will need in this class (or, if not, to refresh you on them). We expect this assignment to be straightforward; if you find any section particularly difficult at this time, you may find the course content much more difficult. **This assignment will not count towards your final grade, but you must score at least 40 out of 50 points to remain in the class.**

Since this is a preliminary assignment, you may not use any late days on this assignment and no late submissions will be accepted.

Cite any external sources used. No collaboration is allowed on this assignment.

For written problems, make sure to show all work. Points will be deducted if only the answer is given without explanation.

Piazza: You may post clarification questions about this homework on the class Piazza under the "hw0" folder. Since this homework is a prerequisite to remaining in the class, please treat it as an exam and do not include partial solutions or hints or ask "how-to" questions. You may post your question privately to the instructors if you are unsure.

1 Environment Setup and Programming (10 points)

1.1 Environment Setup (5 points)

Environment setup instructions are provided in a separate document, which we link here and on the course webpage. You will need to create your Google Cloud account and create a virtual machine on which you will test your homework assignments before submitting. Use this virtual machine to complete problem 1.2.

Submit a screenshot of your Google Cloud "VM instances" page (found under the Compute Engine) showing your virtual machine running (i.e., with the green checkmark next to your VM instance). Include this in your typeset submission.

start VM instance,
find external IP
mac$ ssh -Y jq2282@35.231.5.177

1.2 Programming (5 points)

You are given some starter code for sentiment classification adapted from scikit-learn's tutorials. The objective of this code is to train a model that can decide whether a given piece of text is expressing a positive or negative attitude—don't worry if you don't understand the details at this stage; the machine learning is done for you, as this is a programming problem and not a machine learning one.

Unzip the `hw0.zip` file on your virtual machine (you may use `scp` or `sftp` to transfer the files onto your VM, or download them directly from the website on your VM with `wget`, or use the file browser on MobaXTerm...). Remember to activate your Anaconda environment with `conda activate coms4705`.

You should be able to run `python hw0.py` from inside `hw0/` with no errors, and it should take seconds to run. Spend some time understanding the code, the task, and the output. Finally, make the following modifications:

- Plot the precision-recall curve for the existing classifier. **Hint: check the docs and examples.**
- Plot the precision-recall curve when the number of neighbors is 30 and 50 instead. **Hint: docs.**

Submit the three precision-recall plots for this part (one each for n_neighbors $\in \{10, 30, 50\}$). Include them as images in your typeset submission and be sure to indicate which is which.

2 Calculus (10 points)

2.1 Chain rule and multivariate derivatives (5 points)

i) Consider the following three functions:

$$f(x, y) = xg(x, y) + 2y$$

$$g(x, y) = x^2 y - xh(x^2, y)$$

$$h(x, y) = xy^2 + 5$$

Compute $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$.

ii) Let $f(x, y, z) = x/y^2 + z \exp(x^2)$. Compute $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$, $\frac{\partial f}{\partial z}$.

2.2 Maxima and minima (5 points)

Let $f(x) = x \log_2(x) + (1-x) \log_2(1-x)$. Find the maxima and minima of $f(x)$ for $x \in [0, 1]$.

3 Probability and Statistics (10 points)

3.1 Conditional probability (5 points)

Let's say that there are 10 strips of paper in a hat, each with a single word written on it. Five of those strips of paper have the word 'buffalo'. The other five have the word 'police'. You pull out three strips of paper at random. (You pull one strip of paper out of the hat at a time – you don't put the strip of paper back when you reach for the next one.) What is the probability that you pull out the following words:

buffalo buffalo buffalo

Show your work.

3.2 Bayes' rule (5 points)

Recall that Bayes' rule gives the conditional probability of event Y given event X in terms of $P(X|Y)$ and their individual probabilities:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}.$$

Let's say you have two friends called Maria, one (Maria A) who talks about cats 90% of the time, and dogs 10% of the time, and another friend (Maria B) who talks about dogs 90% of the time, and cats 10% of the time. They both text you equally. You get a text message from Maria about dogs, but you don't know which friend it is. What's the probability this message is from Maria B? Show your work.

4 Linear Algebra (20 points)

4.1 Basic matrix operations (10 points)

i) Multiply the below matrices (by hand).

$$\begin{bmatrix} 0 & 2 \\ 1 & 3 \\ 2 & 0 \end{bmatrix} \begin{bmatrix} 1 & 5 & 2 \\ 4 & 3 & 1 \end{bmatrix}$$

ii) Recall that for random variables X and Y with respective standard deviations σ_X and σ_Y, their Pearson's correlation coefficient is given by

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}.$$

Also recall that the cosine similarity between two vectors $\mathbf{u}$, $\mathbf{v}$ is the cosine of the angle θ between them, and is related to the inner product as follows:

$$\langle \mathbf{u}, \mathbf{v} \rangle = |\mathbf{u}||\mathbf{v}| \cos\theta.$$

Show that for two vectors $\mathbf{u}, \mathbf{v} \in \mathbf{R}^n$ with zero elementwise mean, the correlation between their elements is equal to their cosine similarity.

4.2 Singular Value Decomposition (10 points)

Recall that, for a real $m \times n$ matrix M, its singular value decomposition is given by

$$M = U\Sigma V^T,$$

where U and V are orthogonal matrices and Σ is diagonal.

i) What must the dimensions of U, Σ, and V be, respectively? Why?

ii) Given U, Σ, and V, show how to compute the inverse of M.

5 Submission instructions

This assignment must be submitted via Gradescope as a **typeset PDF**. You are encouraged to use LaTeX, but solutions can be written up using Word, etc., as long as they are typed up in a legible font. Handwritten submissions will not be accepted. Include your name and UNI on your submission.

If you are on the waitlist and would like to be considered for admission, submit your assignment by emailing it to columbianlpfall19@gmail.com by the same deadline, using the subject "[HW0] YOUR_UNI". Make sure to use your Lionmail account. This is a **requirement** for being admitted to the class from the waitlist.

6 Academic integrity

Copying or paraphrasing someone's work (code included), or permitting your own work to be copied or paraphrased, even if only in part, is not allowed, and will result in an automatic grade of 0 for the entire assignment or exam in which the copying or paraphrasing was done. Your grade should reflect your own work. If you believe you are going to have trouble completing an assignment, please talk to the instructor or a TA in advance of the due date.