Project 2

Writeup due on Tuesday, December 18 at 9PM, via CoCalc

Overview

Welcome to your second project! Like Project 1, this project assignment is to be done in groups of two or three students.

You will be analyzing a database of movies. Your ultimate goal in this project is to build a classifier that predict the genre of a movie—action or romance—based on the frequency in which certain words appear in the movie script and to analyze the performance of these classifiers.

Lab 9, Homework 9, and Lab 10 provide some guidance on getting started with this dataset.

Main Project Components

1. **Data exploration and feature selection**

   In order to get a sense of what features can help us distinguish an action movie from a romance movie, you would have to do some initial exploration of your training dataset. Your project should include at least three data visualizations that help justify how you design your classifier.

2. **Development of your own classifier**

   In class, we have developed a few "simple classifiers". Your task here is to build your own "simple" classifier. Your classifier need not be as complicated as fancy classifiers that are used by professional machine learning experts, but it should be less simple than the ones we created in class: your classifier should use at least three features, and you should use more than just one pair of if-else statements. This classifier should be informed by your initial data exploration.

3. **Assessment of your classifier**

   Assess how your classifier does in predicting the genre of the movies in your test dataset. You should measure the accuracy of your classifier. Then, you should use one other metric (this could be one that we have discussed in class or one that you construct yourself!). If you use a metric that you come up with yourself, include a brief explanation of your idea for this metric.

4. **Comparison of the performance of your classifier to that of a k-Nearest Neighbor (kNN) Classifier**

   Use a kNN classifier (with a particular choice of k) to predict the genre of the movies in your test dataset and assess the quality of the predictions. Compare the performance of your classifier to that of the kNN classifier; comment on the result.

Accessing The Project Materials

The datasets can be found in your CoCalc project, under the `Project2` folder, which will also contain a "template" for your write-up: `Project2.ipynb`; your project write-up will be written in this Jupyter Notebook. You may add additional files (images, etc.) into this folder if you would like to include them in your write-up All content of this folder will be collected when the project is due.

Source

The original dataset is obtained from the UCI Machine Learning Repository: `https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)`

Due Dates

Project 2 Presentation (Progress Report): **Thursday, December 14, in class**
Project 2 Write-up: **Tuesday, December 18 at 9PM, via CoCalc**
Project 2 Self- and Peer-Assessments: **Wednesday, December 19 at 9PM, via Google Forms**

Deliverables

1. Presentation (Progress Report)

You will be asked to prepare a set of Google Presentation slides (3-4 slides) which your team will present on Dec 14. Your team will have 3-5 minutes to present. The slides should contain:

1. (1 slide) Results of your initial data exploration and feature selection, including at least two data visualizations. For example, you can include scatterplots like those you created in Lab 9.
2. (1-2 slides) An outline of your "simple" classifier. If you already have preliminary results on how well this classifier performs, you should include them in your presentation as well.
3. (1 slide) A summary of your plan for the other components of your project. For example, what other metric(s) do you plan to use to evaluate your classifier? Do you have ideas for modifying your current "simple" classifier to try to improve its performance? If you came across challenges that your team is still working on, you can discuss them in this last slide as well.

2. Project Write-up

In your Project2 folder on CoCalc, you must have the original datasets and your project write-up, written in a Jupyter Notebook named `Project2.ipynb`. Your write-up must include:

1. An introduction paragraph that gives readers an overview of what questions you explored, what you did in this project, and how your classifier performs.
2. A thorough write-up of each of the four main components of the project (see the grading rubric on page 3 for the details).
3. Any code that you use is included in an appendix at the end of your write-up.

3. Self- and Peer-Assessments

The link to the self- and peer-assessments form will be posted as an announcement on NYU Classes on Tuesday, December 18 at 9PM. You will have until Wednesday, December 19 at 9PM to complete this form.

Expectations

- We want to see an evidence of your **mastery of the data analysis skills in R**. You will also be evaluated in **how clearly you interpret and communicate** the outcomes of your data analysis. (See the grading rubric on the next page.)
- Your write-up needs to be a solid course project. That is, it has to be a **complete** and **well-written** paper.
- You will work together as a team, where each team member makes a meaningful contribution to the project. See pages 4-5 for guidelines for working in a team.

Grading Rubric

PRESENTATION [30 points]	
Link to your team's Google Presentation slides is included in your submission of Lab 10.	4 points
- One slide containing at least two data visualizations	8 points
- One-two slides containing an outline of a "simple" classifier.	4 points
- One slide containing a summary of your plan for the remaining components of your project	4 points
Individual presentation grade: Team member participates in presenting the slides, communicates ideas clearly, and displays firm knowledge in what they are presenting.	10 points
PROJECT WRITE-UP [150 points]	
Clear Introduction	
The write-up has an introduction paragraph that outlines the rest of the paper, including brief summaries of	
- the main ideas behind your classifier;	5 points
- the performance of your classifier, as measured in terms of accuracy and at least one other metric; and	5 points
- how your classifier compares to a kNN classifier.	5 points
1. Data exploration and feature selection	
Your write-up about this step includes	
- at least three data visualizations/summary tables and	15 points
- a clear and thorough discussion/explanation/interpretation of what is illustrated by the data visualization and analysis, including what the data visualizations tell you about which features are important to incorporate in your classifier.	10 points
Your data analysis work should include splitting the dataset into training and test datasets; the data exploration step should be done on the training data only.	10 points
2. Development of your own classifier	
Your write-up about this step includes	
- a description of what your classifier does;	5 points
- an intuitive explanation of why what your classifier does seem to make sense; and	5 points
- an implementation of your classifier. That is, there should be a code cell where you write a new R function to implement this classifier.	10 points
- your classifier should use at least 3 features, and you should use more than just one pair of if-else statements	10 points
3. Assessment of your classifier	
Your write-up about this step includes	
- an R function for computing accuracy; the accuracy of your classifier;	5 points
- an R function for computing one other performance metric; the value of this metric when you use it to assess your classifier;	5 points
- a clear description of your second metric and why you chose it; and	5 points
- a thorough discussions of the results of your assessment in terms of the two metrics.	10 points
4. Comparing performance to a kNN classifier	
Your write-up about this step includes	
- an explanation of what value of k you use and why;	5 points
- the accuracy of this kNN classifier;	5 points
- the performance of this kNN clssifier in terms of your second performance metric;	5 points
- thorough discussions of the results of your assessment, including how your classifier compares to this kNN classifier in terms of the two metrics you used.	10 points
Codes and Correctness	
Any code that you use is included in your write-up. Data analysis methods are used correctly.	10 points
Grammar, Punctuation, and Spelling	
The write-up uses correct grammar, punctuation, and spelling. Sentences and paragraph structure make sense.	10 points
SELF- AND PEER-EVALUATION [20 points] (see page 5)	20 points
TOTAL	200 points

Advice for working as a team

1. **Team work.** Being able to work well in a team is an important skill. Below are some tips for working well as a team on this project.

   - **Exchange contact information and schedule at least one in-person meeting** outside class or lab. This could be a low-key meal-time or coffee meeting where you can get a bit of work started.
   - **Make sure everyone is on the same page regarding what questions you will work on as a team.** Listen to everyone's thoughts. Everybody has strengths and weaknesses; everyone can contribute something to the project and can learn something from a teammate.
   - **Make sure that the division of labor is clear and that everyone in the team is happy with it.** It's also useful to agree on a schedule/timeline.
   - **Divide the work, but make sure that everyone has some contribution to each aspect of the project.**
     For example, if Alex is very good at writing but not very confident in producing bar charts, then Alex might take on the role of your team's "lead writer", but Alex should still be involved in producing the charts, and other team members should still contribute to the writing. This project should help Alex improve their chart-making skills and help the other team members improve their writing skills.
   - **Do your part, promptly.** Every student will be asked to assess their and their teammates' contribution. **See page 5.**

2. **Working collaboratively in CoCalc.** It might be useful to be able to collaboratively edit a Jupyter Notebook with your group members, just like you can work collaboratively on other cloud-based services like Google Docs.

   If you are interested in doing this, you can create a new project on CoCalc as follows:

   - Go to `cocalc.com`. Click "Create a new project".
   - Enter the name of your new CoCalc project (for example, call it "FDTD Project 2") and click "Create Project"
   - Click the "Settings" tab. Add your group members as collaborators
   - Your group members can now access the same project together.

   Next, you need to copy the dataset and the Project2.ipynb files from our course CoCalc project to your new CoCalc project. To do this:

   - Go to our course CoCalc project
   - Click the checkbox next to the Project2 folder. On the top menu, click "Copy".
   - Select "Copy to a folder or a different project".
   - From the dropdown menu, click the project you just created. Click "Copy 1 Item"
   - Done. You can go back to the new CoCalc project you created to check that the new folder is there.

   We have included screenshots to illustrate the above steps. **See pages 6-7.**

   Once you are done working on this project collaboratively, one person in your group needs to **copy the completed Project2.ipynb file back to the course project** so that we are able to see your finished write-up and grade it.

Appendix A: Group Assessment

After the project is collected, you will be asked to complete a Google Form to assess yourself and your teammate(s). The link will be made available later, but the table below should give you an idea what the form will look like.

There are six categories that we would like you to use to assess your peers and yourself. For each item, you will be asked to rate each teammate and yourself using the scale given. Please think hard and honestly about each of the categories and how you and each group member performed.

It is not necessary that everyone get the highest score on each item. Different people will have different strengths and different contributions. What we want to distill from this form is individual accountability: Each member of the team contributes to the group project and no one receives an undeserved grade/credit based on the work of their teammates.

Team work skills are not easy to master but (as with anything) you will get better with thoughtful practice. By making this form available to you now, we hope to give you some direction for being a better team member.

Rater's Name: _____

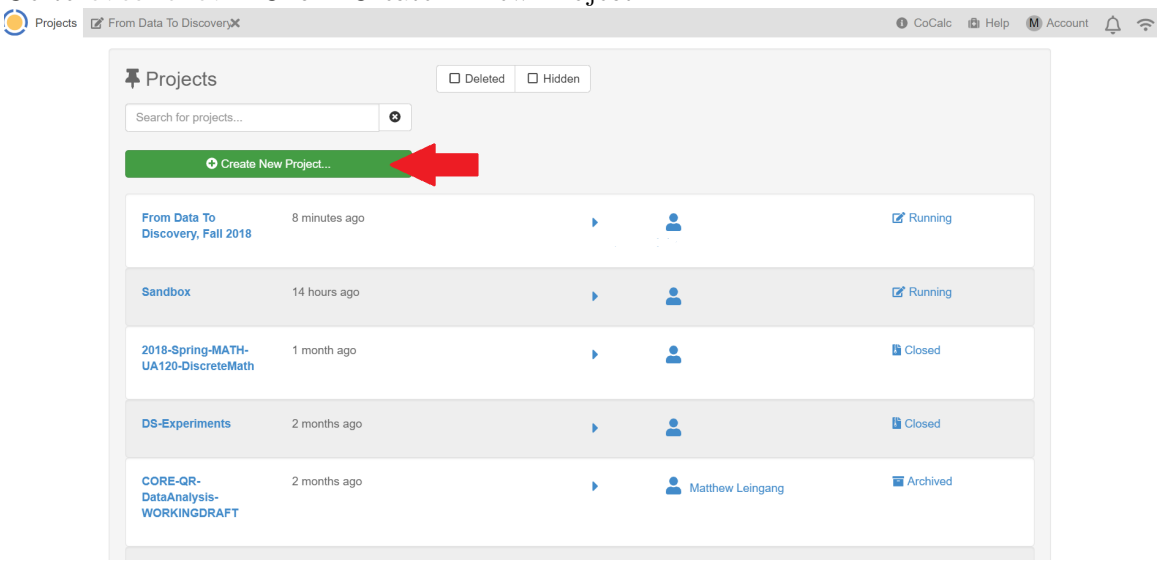
Assessment of: _____

Rating		Comments, Examples, Explanations, etc.
Group Participation Attends meetings regularly and on time.		
Time Management & Responsibility Accepts fair share of work and reliably completes it by the required time.		
Adaptability Displays or tries to develop a wide range of skills in service of the project, readily accepts changed approach or constructive criticism.		
Creativity/Originality Problem-solves when faced with impasses or challenges, originates new ideas, initiates team decisions.		
Communication Skills Effective in discussions, good listener, capable presenter, proficient at diagramming, representing, and documenting work.		
General Team Skills Positive attitude, encourages and motivates team, supports team decisions, helps team reach consensus, helps resolve conflicts in the group.		
Technical Skills Ability to create and develop materials on own initiative, provides technical solutions to problems.		
Scoring For each category, award yourself and each member of your team a score using this scale.	**3** – Better than most of the group in this respect **2** – About average for the group in this respect **1** – Not as good as most of the group in this respect **0** – No help at all to the group in this respect	

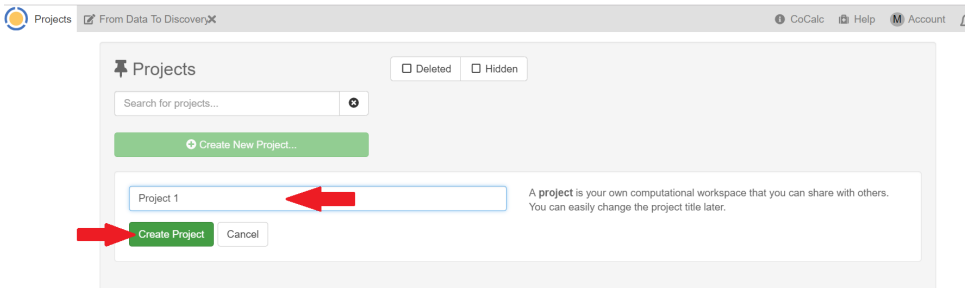
(adapted from Goldfinch, 1994; Lejk & Wyvill, 2001)

Appendix B: Creating a new project in CoCalc

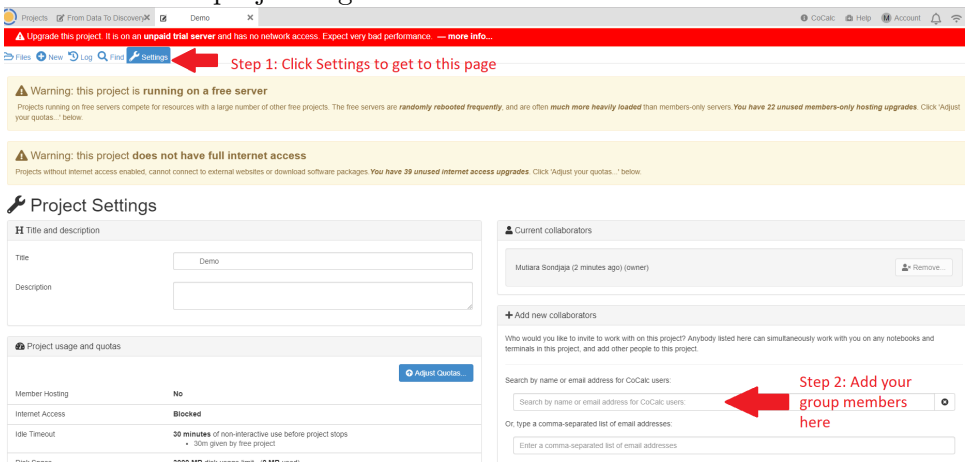
Go to `cocalc.com`. Click "Create A New Project".



Enter the name of your new CoCalc project (for example, call it "FDTD Project 2") and click "Create Project"

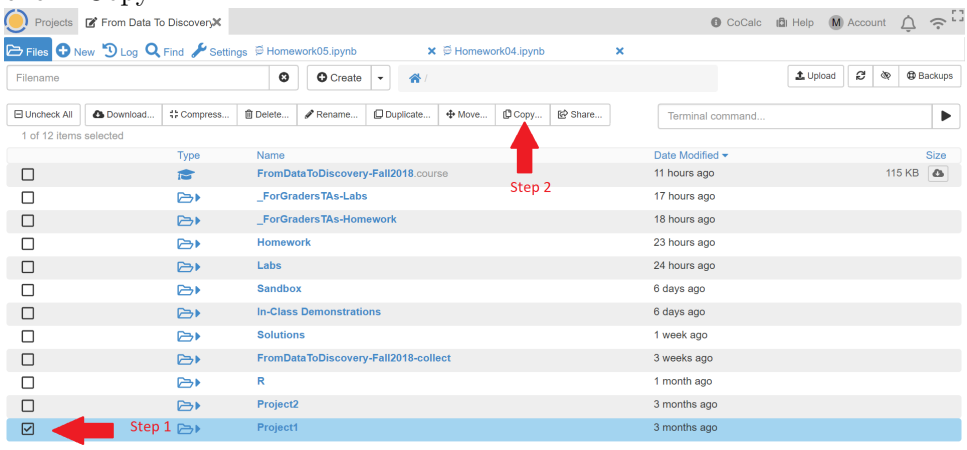


Click the "Settings" tab. Add your group members as collaborators.Your group members can now access the same project together.

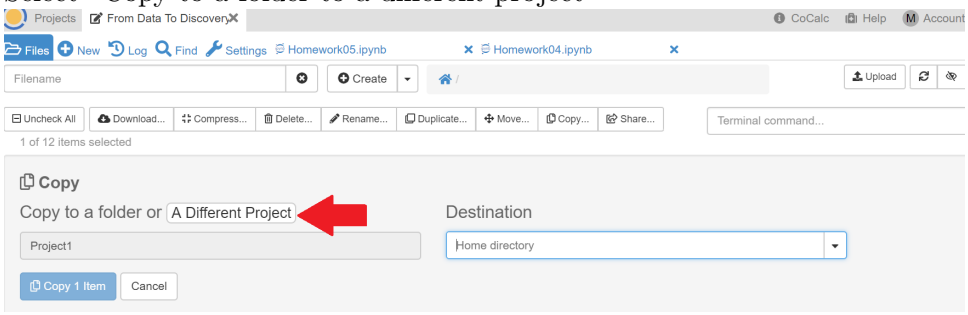


Appendix C: Copying a file from one CoCalc project to another

Go to our course CoCalc project. Click the checkbox next to the Project1 folder. On the top menu, click "Copy"



Select "Copy to a folder to a different project"



From the dropdown menu, click the project you just created. Click "Copy 1 Item"