EE 445: Final Project

April 3, 2018

Please work individually. You can, of course, consult each other and the Internet.

1 Deliverables

1. Working code for competition due in class on 5/2
2. 2-4 page report explaining methodology (Due 5/4). Ask me for template if you want to write it in a conference paper format.

2 Problem

Rather than working with tech company data, we will work with medical data (clinical and protein data) of Leukemia. You need to design a binary classifier that can help to identify if a patient will (complete remission) or will not (complete resistant) respond to treatment. **This dataset has missing data.**

You will be given training data for 181 patients (**AMLtrain.csv**). After patient ID, the first 40 features are clinical data (see Figure 1 for description). The field **respsimple** is what we want to predict: CR stands for complete remission vs. RESISTANT. The rest of the features are proteomic measurements from 231 antibodies.

3 Project/Report Requirements

Do not forget to cite the source/ package if you use other people's functions and methods. Your report needs to contain the following information:

1. What did you do about the missing data? Why? How might your approach affect learning
2. You have to explore 2 different data ”extraction” techniques. This may be some sort of unsupervised structural learning, feature extraction algorithms, linear combination of features, etc. Clearly explain what methods you used, a basic description of how the method work, and why you chose it.

3. You have to explore 3 different binary classifier. You may use classifiers not covered in class. However, you need to clearly explain what methods you used and a basic description of how the method work (i.e., don't just use google's sophisticated learning tools as a black box).

4. Basically, you have 6 different learning methods. You have to pick the best method for the competition. Explain your selection process of how you picked the best classifier.

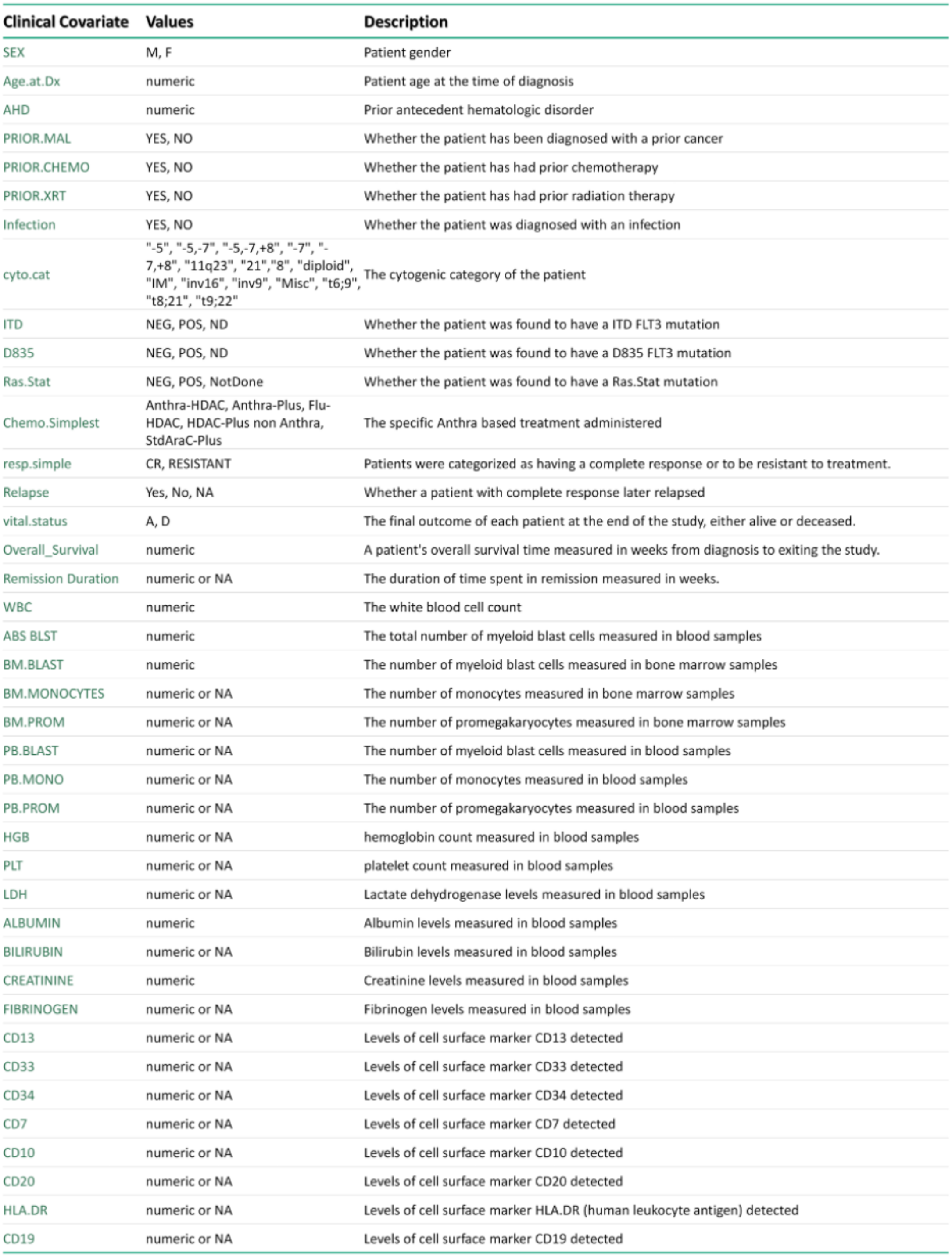
4 Project Competition

10 patients are withheld from the training data to be used for testing. We will have the competition live in class on 4/2 (last day of class).

You code should be able to take information on 10 patients (no missing information) and predict **respsimple** (complete remission (CR) or resistant). The fields you should expect are in **testheader.csv**. You are automatically disqualified if your code does not run. You can run it on your own computer or give it to me to run it.

The winner (most accurate classifier) has to give a short summarization of their approach (assuming we have time in class).

Winner gets a $50 gift card or cold hard cash (your preference). If there are multiple winners, we'll split it because I'm poor and cheap.



Clinical Covariate	Values	Description
SEX	M, F	Patient gender
Age.at.Dx	numeric	Patient age at the time of diagnosis
AHD	numeric	Prior antecedent hematologic disorder
PRIOR.MAL	YES, NO	Whether the patient has been diagnosed with a prior cancer
PRIOR.CHEMO	YES, NO	Whether the patient has had prior chemotherapy
PRIOR.XRT	YES, NO	Whether the patient has had prior radiation therapy
Infection	YES, NO	Whether the patient was diagnosed with an infection
cyto.cat	"-5", "-5,-7", "-5,-7,+8", "-7", "-7,+8", "11q23", "21","8", "diploid", "IM", "inv16", "inv9", "Misc", "t6;9", "t8;21", "t9;22"	The cytogenic category of the patient
ITD	NEG, POS, ND	Whether the patient was found to have a ITD FLT3 mutation
D835	NEG, POS, ND	Whether the patient was found to have a D835 FLT3 mutation
Ras.Stat	NEG, POS, NotDone	Whether the patient was found to have a Ras.Stat mutation
Chemo.Simplest	Anthra-HDAC, Anthra-Plus, Flu-HDAC, HDAC-Plus non Anthra, StdAraC-Plus	The specific Anthra based treatment administered
resp.simple	CR, RESISTANT	Patients were categorized as having a complete response or to be resistant to treatment.
Relapse	Yes, No, NA	Whether a patient with complete response later relapsed
vital.status	A, D	The final outcome of each patient at the end of the study, either alive or deceased.
Overall_Survival	numeric	A patient's overall survival time measured in weeks from diagnosis to exiting the study.
Remission Duration	numeric or NA	The duration of time spent in remission measured in weeks.
WBC	numeric	The white blood cell count
ABS BLST	numeric	The total number of myeloid blast cells measured in blood samples
BM.BLAST	numeric	The number of myeloid blast cells measured in bone marrow samples
BM.MONOCYTES	numeric or NA	The number of monocytes measured in bone marrow samples
BM.PROM	numeric or NA	The number of promegakaryocytes measured in bone marrow samples
PB.BLAST	numeric or NA	The number of myeloid blast cells measured in blood samples
PB.MONO	numeric or NA	The number of monocytes measured in blood samples
PB.PROM	numeric or NA	The number of promegakaryocytes measured in blood samples
HGB	numeric or NA	hemoglobin count measured in blood samples
PLT	numeric or NA	platelet count measured in blood samples
LDH	numeric or NA	Lactate dehydrogenase levels measured in blood samples
ALBUMIN	numeric	Albumin levels measured in blood samples
BILIRUBIN	numeric or NA	Bilirubin levels measured in blood samples
CREATININE	numeric	Creatinine levels measured in blood samples
FIBRINOGEN	numeric or NA	Fibrinogen levels measured in blood samples
CD13	numeric or NA	Levels of cell surface marker CD13 detected
CD33	numeric or NA	Levels of cell surface marker CD33 detected
CD34	numeric or NA	Levels of cell surface marker CD34 detected
CD7	numeric or NA	Levels of cell surface marker CD7 detected
CD10	numeric or NA	Levels of cell surface marker CD10 detected
CD20	numeric or NA	Levels of cell surface marker CD20 detected
HLA.DR	numeric or NA	Levels of cell surface marker HLA.DR (human leukocyte antigen) detected
CD19	numeric or NA	Levels of cell surface marker CD19 detected

Figure 1: Clinical feature description