Project Proposal

Vladimir Preda, Maxime Gevers, Ana Chaloska

1. Dataset:

- **Name of dataset: Breast Cancer Wisconsin (Original) Data Set**
  https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28original%29

Data Set Characteristics:	Multivariate	**Number of Instances:**	699	**Area:**	Life
Attribute Characteristics:	Integer	**Number of Attributes:**	10	**Date Donated**	1992-07-15
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	315454

- **Number of examples:** 699
- **Number of attributes:** 10 plus the class attribute
- **Data type of attributes:** Integers
  - The attributes are categorical, so we are going to use the function OneHotEncoder() from sklearn.preprocessing to get binary features.
  - The attributes are:

```
#  Attribute                     Domain
-- -----------------------------------------
1. Sample code number            id number
2. Clump Thickness               1 - 10
3. Uniformity of Cell Size       1 - 10
4. Uniformity of Cell Shape      1 - 10
5. Marginal Adhesion             1 - 10
6. Single Epithelial Cell Size   1 - 10
7. Bare Nuclei                   1 - 10
8. Bland Chromatin               1 - 10
9. Normal Nucleoli               1 - 10
10. Mitoses                      1 - 10
11. Class:                       (2 for benign, 4 for malignant)
```

Attribute 1 is not relevant because it is just an id number of a patient. Thus, there are 9 relevant attributes. Since all of them are integers varying between 1 and 10, After applying OneHotEncoder()[1]to the dataset we are going to get 90 binary features.

2. Problem:

We are predicting the probability of a patient having a malignant breast tumor. If the probability is greater or equal to 0.5, the tumor will be classified as malignant (class name: 4). If the probability is lower than 0.5, the tumor will be classified as benign (class name: 2).

[1] http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html

3. Algorithms:

We are going to train 3 algorithms and compare their performance. The algorithms we are going to use are (this is still a subject to potentially change, but this is our plan for now):

- Logistic Regression
- SVM
- Random Forrest
- Neural Network

After training the classifiers, we are also going to compare them to already developed classifiers in previous research.

4. Comparison & Evaluation:

- We are going to evaluate the classifiers by using confusion matrices.
- We are going to compare the classifiers by using ROC and AUC.
    - Is it better to use precision-recall curve? We thought it would be better to use ROC and AUC because we want to find how well the classifiers can perform in general, and we are not looking at how meaningful are the positive results.