# Machine Learning Engineer Nanodegree

## Capstone Proposal

Cesar Trevisan
April 12st, 2018

## Proposal

### Domain Background

Cancer occurs as a result of mutations, or abnormal changes, in the genes responsible for regulating the growth of cells and keeping them healthy. The genes are in each cell's nucleus, which acts as the "control room" of each cell. Normally, the cells in our bodies replace themselves through an orderly process of cell growth: healthy new cells take over as old ones die out. But over time, mutations can "turn on" certain genes and "turn off" others in a cell. That changed cell gains the ability to keep dividing without control or order, producing more cells just like it and forming a tumor.

The term "breast cancer" refers to a malignant tumor that has developed from cells in the breast. Usually breast cancer either begins in the cells of the lobules, which are the milk-producing glands, or the ducts, the passages that drain milk from the lobules to the nipple. Less commonly, breast cancer can begin in the stromal tissues, which include the fatty and fibrous connective tissues of the breast.

Identifying correctly whether a tumor is benign or malignant is vital in deciding what is the best treatment, saving and improving the quality of life. In this project, we used Breast Cancer Wisconsin (Diagnostic) Data Set to create a model able to predict if a tumor is or not dangerous based in characteristics that were computed from a digitized image of a fine needle aspirate (FNA) of a breast mass.

Source: BreastCancer.org, Inside Radiology

### Problem Statement

A tumor can be benign (not dangerous to health) or malignant (has the potential to be dangerous). Benign tumors are not considered cancerous: their cells are close to normal in appearance, they grow slowly, and they do not invade nearby tissues or spread to other parts of the body. Malignant tumors are cancerous. Left unchecked, malignant cells eventually can spread beyond the original tumor to other parts of the body. As the physical aspects of the malignant tumor differ from the benign tumor cells, we can measure the physical characteristics such as radius (mean of distances from center to points on the perimeter), texture (standard deviation of gray-scale values), perimeter, area, smoothness, compactness, concavity, concave points, symmetry or fractal dimension to understand and create two classes of tumor and identify which class each tumor belongs for new samples.

### Datasets and Inputs

To create a model capable to predict whether a tumor cell is or not malignant we'll use a labeled dataset [The Wisconsin Diagnostic Breast Cancer (WDBC)] (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic) was created in 1995 by: Dr. William H. Wolberg (General Surgery Dept., University of Wisconsin Clinical Sciences Center), W. Nick Street (Computer Sciences Dept., University of Wisconsin) and Olvi L. Mangasarian (Computer Sciences Dept., University of Wisconsin). Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

The dataset have the following structure:

- Number of instances: 569
- Number of attributes: 32 (ID, diagnosis, 30 real-valued input features)
- Diagnosis (M = malignant, B = benign)
- Missing attribute values: none
- Class distribution: 357 benign, 212 malignant
- All feature values are recoded with four significant digits.
- Ten real-valued features are computed for each cell nucleus:

  a) radius (mean of distances from center to points on the perimeter)
  b) texture (standard deviation of gray-scale values)
  c) perimeter
  d) area
  e) smoothness (local variation in radius lengths)
  f) compactness (perimeter^2 / area - 1.0)
  g) concavity (severity of concave portions of the contour)
  h) concave points (number of concave portions of the contour)
  i) symmetry
  j) fractal dimension ("coastline approximation" - 1)

  The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

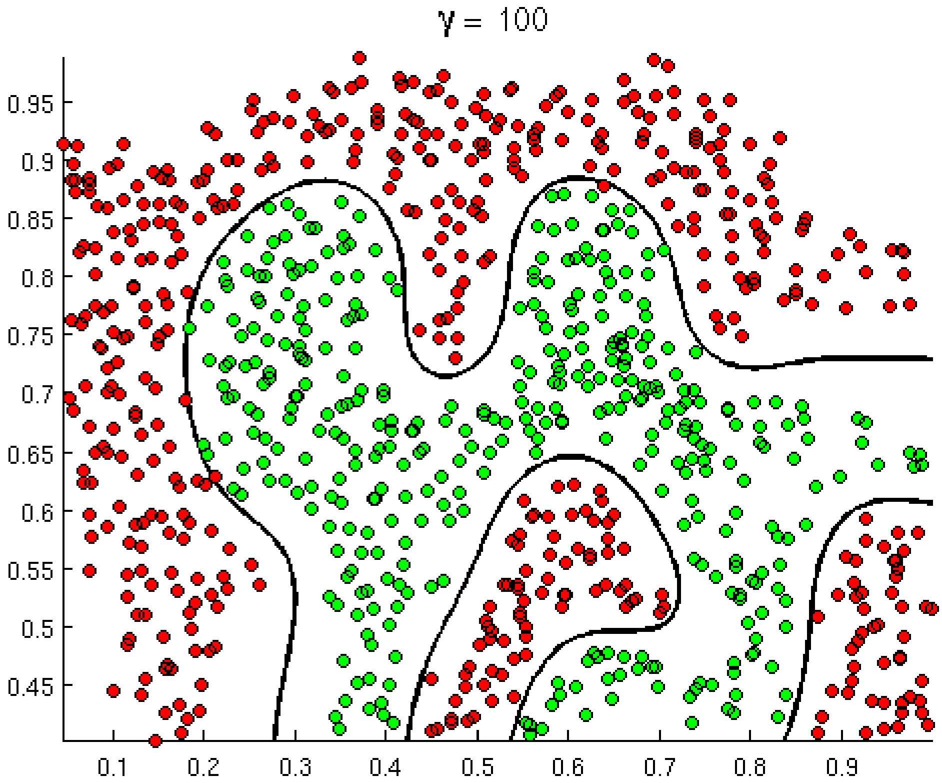
We will used this physical characteristics as features to train a machine learning algorithm to create and evaluate a model that classifies if a cell is benign or malignant based in this characteristics.

## Solution Statement

In this section, clearly describe a solution to the problem. The solution should be applicable to the project domain and appropriate for the dataset(s) or input(s) given. Additionally, describe the solution thoroughly such that it is clear that the solution is quantifiable (the solution can be expressed in mathematical or logical terms) , measurable (the solution can be measured by some metric and clearly observed), and replicable (the solution can be reproduced and occurs more than once).

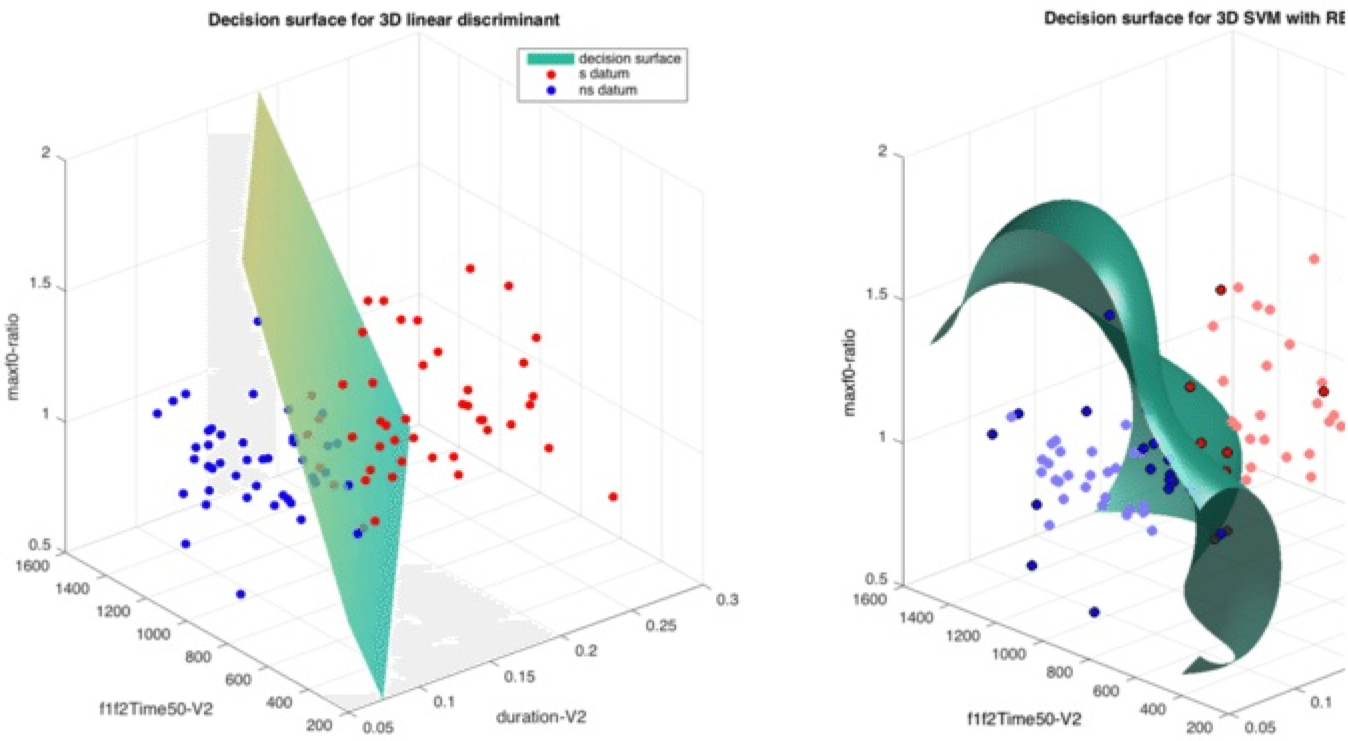
Using a dataset with 30 features going to predict if class a sample belongs, it means that we face a multidimensional classification problem. Our job is understand and prepare data to feed a algorithm wich should find the best decision boundary or in our case, the best decision hypperplane, see:

** Decision Boundary 2-dimensional **



When we have 2-dimensional data we can separate data points using a decision boundary.

** Decision Surface - 3 dimensional **



Decision surface, a surface that separates data points in a 3-dimensional space.

We going to performe Feature Selection using tools like feature_selection, SelectKBest or SelectPercentile, modules in sklearn. And Dimension Reduction to get the optiomal decision bondary/surface mesuring accuracy of different models, combination of features and dimensionality.

## Benchmark Model

To realize how much effective the method is, after optimize our model I'll compare results with results extracted of the following Paper:

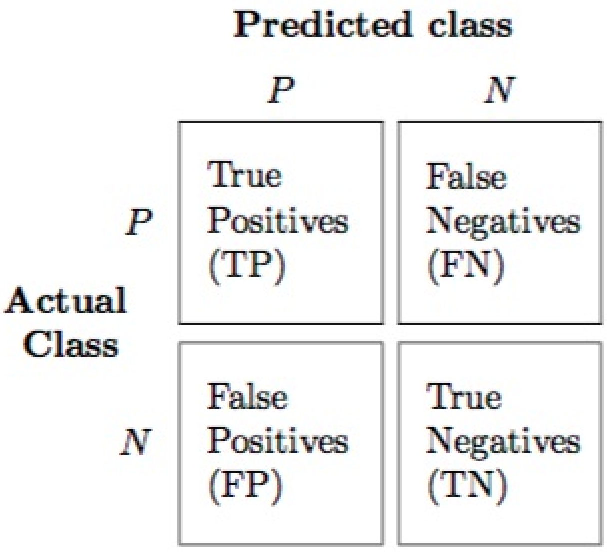
Approximate Distance Classification
Adam H. Cannon, Lenore J. Cowen, Carey E. Priebe
Department of Mathematical Sciences
The Johns Hopkins University
Baltimore, MD, 21218

They used k-nearest neighbor to predict cancer in same dataset to diagnosis Breast Cancer and their effectiveness, in this way we can measure and compare results in the same domain.

## Evaluation Metrics

For final model I will analyze the model performance ploting a confusion matrix and score model using F1 Score that is a methot to measure performance of binary classification, the F1 score is a measure of a test's accuracy, is the harmonic average of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.

** Confusion Matrix **



| | | Predicted class | |
|---|---|---|---|
| | | *P* | *N* |
| Actual Class | *P* | True Positives (TP) | False Negatives (FN) |
| | *N* | False Positives (FP) | True Negatives (TN) |

## Project Design

The work-flow will follow these steps:

1. Import Data
2. Analyze data structure and statistics a. Clean Data (handle with outiliers and missing values)
3. Create Visualizations to better understand data
4. Create new features
5. Identify the best features or combination of features
6. Test dimension reduction using PCA
7. Test, measure and tune different methods a. test classifiers (e. g. XGBoot, Decision Trees, Support Vector Machine, Logistic Regression, Bayes, etc)
   b. test results and select some of the bests algoritms
   c. tune algorithms parameters
8. Analyze results and make conclusion