Cairo University

Faculty of Engineering

Systems And Biomedical Engineering

THORACIC SURGERY SURVIVAL PROPOSAL

Group Number : 2

Group Members :

Asmaa Mahmoud Mahmoud

Alaa Gamal Abdelaziz

Salma Mohamed Zakaria

Marwa Adel Youssef

1 Project background and motivation

We have chosen Thoracic surgery as its evolving and surgeons have major collaborative roles in management of lung cancer, respiratory infections, chest trauma, pediatric respiratory disorders and end-stage respiratory. Today, lung cancer is the most frequent indication for thoracic surgery. Thoracic Surgeries focuses on the chest organs, including the esophegus , trachea , pleura , chest wall, diaphragm, heart, and lungs. Technological advances have increased the safety and availability of these complex surgical procedures. Lung cancer surgeries and anti-reflex surgeries save and improve lives around the world. The most common diseases requiring thoracic surgery include lung cancer (its rate of survival is very low among women and men) , chest trauma (require urgent thoracic surgery), esophageal cancer,(its rate is rising slowly recently) emphysema, and lung transplantation. So in conclusion we chose this project because we are hopeful that our model will be useful to detect the patients who are at risk after surgery by importing their symptoms in our model which we will be using the data from former patients' information in.[1]

2 Objectives of project

1. Learning statistics which is mathematical study of data as you cannot do statistics without having data and it will be done by using statistical model which is a model for the data that is used either to infer something about the relationships within the data or to create a model that is able to predict future values. [2]

2. Making statistical models which are designed for inference about the relationships between variables.

3. Learning Machine learning which has a purpose of obtaining a model that can make repeatable predictions.

4. Learning the three methods of the machine learning which are:

    - Naive Bayes (NB) Classier (or Gaussian NB Classifier).
    - K-nearest neighbors (KNN) model.
    - Logistic regression.

5. Learning how to make Data Visualization.

3 Data Set Information:

The data was collected retrospectively at Wroclaw Thoracic Surgery Centre for patients who underwent major lung resections for primary lung cancer in the years 2007 & 2011. The Centre is associated with the Department of Thoracic Surgery of the Medical University of Wroclaw and Lower-Silesian Centre for Pulmonary Diseases, Poland, while the research database constitutes a part of the National Lung Cancer Registry, administered by the Institute of Tuberculosis and Pulmonary Diseases in Warsaw, Poland.[3]

3.1 Attribute Information:

1. DGN: Diagnosis -specific combination of ICD-10 codes for primary and secondary as well multiple tumours if any (DGN3, DGN2, DGN4, DGN6, DGN5, DGN8, DGN1) ICD-10 codes are alphanumeric codes used by doctors, health insurance companies, and public health agencies across the world to represent diagnoses. Every disease, disorder, injury, infection, and symptom has its own ICD-10 code.

2. ForcedVitalCapacity : is the total amount of air exhaled during the Forced Expiatory Volume test. FVC (numeric)

3. ForcedExpiratoryVolume1: Volume that has been exhaled at the end of the first second of forced expiration - FEV1 (numeric)
   Diagnose obstructive lung diseases such as asthma and chronic obstructive pulmonary disease (COPD). A person who has asthma or Chronic Obstructive Pulmonary Disease has a lower FEV1 result than a healthy person.

   - See how well medicines used to improve breathing are working.
   - Check if lung disease is getting worse. Decreases in the FEV1 value may mean the lung disease is getting worse.

4. PerformanceStatus: - Zubrod scale (PRZ2,PRZ1,PRZ0)
   Zubrod score runs from 0 to 5, with 0 denoting perfect health and 5 death: Its advantage lies in its simplicity.

   - 0 – Asymptomatic (Fully active, able to carry on all predisease activities without restriction)
   - 1 – Symptomatic but completely ambulatory (Restricted in physically strenuous activity but ambulatory and able to carry out work

of a light or sedentary nature. For example, light housework, office work)

- 2– Symptomatic, <50% in bed during the day (Ambulatory and capable of all self care but unable to carry out any work activities. Up and about more than 50% of waking hours)
- 3 – Symptomatic, >50% in bed, but not bedbound (Capable of only limited self-care, confined to bed or chair 50% or more of waking hours)
- 4– Bedbound (Completely disabled. Cannot carry on any self-care. Totally confined to bed or chair)
- 5 – Death

5. PainBS: Pain before surgery (T,F)
6. HaemoptysisBS: Haemoptysis before surgery (T,F)
   - Hemoptysis: is the coughing up of blood or blood-stained mucus from the bronchi, larynx, trachea, or lungs
7. DyspnoeaBS: Dyspnoea before surgery (T,F)
   - dyspnea : is the feeling that one cannot breathe well enough. (Shortness of breath)
8. CoughBS: Cough before surgery (T,F)
9. WeaknessBS: Weakness before surgery (T,F)
10. SizeOfTumer: T(Tumer) in clinical TNM - size of the original tumour, from OC11(smallest) to OC14(largest) (OC11,OC14,OC12,OC13)
    - The TNM Classification of Malignant Tumors (TNM): is a globally recognised standard for classifying the extent of spread of cancer.
11. Type2Diabetes: Type 2 DM - diabetes mellitus (T,F)

12. HeartAttack6M: MI up to 6 months (T,F)

    - Myocardial infarction (MI): also known as a heart attack has happened in the last 6 months

13. PeripheralArterialDiseases: PAD - peripheral arterial diseases (T,F)

    - Peripheral arterial disease is a common circulatory problem in which narrowed arteries reduce blood flow to your limbs.

14. Smoking (T,F)

15. Asthma (T,F)

    - Asthma is a condition in which your airways narrow and swell and produce extra mucus. This can make breathing difficult and trigger coughing, wheezing and shortness of breath.

16. Age: Age at surgery (numeric)

17. Risk1Y: 1 year survival period - (T)rue value if died (T,F)

3.2 Class Distribution:

the class value (Risk1Y) is binary valued.

- Risk1Y Value: Number of Instances:

    - T ← 70

    - F ← 400

3.3 Summary Statistics:

1. Binary Attributes Distribution:

   - PainBS Value: Number of Instances:
     - T ← 31
     - F ← 439
   - HaemoptysisBS Value: Number of Instances:
     - T ← 68
     - F←402
   - DyspnoeaBS Value: Number of Instances:
     - T←31
     - F ←439
   - CoughBS Value: Number of Instances:
     - T ←323
     - F ←147
   - WeaknessBS Value: Number of Instances:
     - T ← 78
     - F ← 392
   - Type2Diabetes Value: Number of Instances:
     - T←35
     - F ← 435
   - HeartAttack6M Value: Number of Instances:
     - T ← 2
     - F ←468
   - PeripheralArterialDiseases Value: Number of Instances:
     - T ← 8

    - F ← 462

  - Smoking Value: Number of Instances:
    - T ←386
    - F ←84
  - Asthma Value: Number of Instances:
    - T ← 368
    - F ← 2

2. Nominal Attributes Distribution:
   - DGN Value: Number of Instances:
     - DGN3 ← 349
     - DGN2 ← 52
     - DGN4 ← 47
     - DGN6 ← 4
     - DGN5 ← 15
     - DGN8 ← 2
     - DGN1←1
   - PerformanceStatus Value: Number of Instances:
     - PRZ2 ← 27
     - PRZ1 ← 313
     - PRZ0 ← 130
   - SizeOfTumer Value: Number of Instances:
     - OC11 ← 177
     - OC14 ← 17
     - OC12 ← 257
     - OC13 ← 19

3. Numeric Attributes Statistics:

Numeric Attributes	Min	Max	Mean	STD
ForcedVitalCapacity	1.4	6.3	3.3	0.9
ForcedExpiratoryVolume1	0.96	86.3	4.6	11.8
Age	21	87	52.5	8.7

4 Pre-processing of Data

In recent year, the difficulty for creating model increasing explosively as the data to be analyzed growing. When the attempt is classifying or creating a model for a specified problem, the various technique could be used, but dealing with problems contain a high number of features is a very challengeable task. So we will be focusing on the analyzing and selecting the data using statistical methods before creating the model.

In our data we have three scale variables and thirteen nominal variables, So we need to determine the correlation between the scale and nominal variables and the one-year status.[4]

We will use statistical tests to analyze and visualize the Thoracic Surgery Data. We will use two types of relations. The ANOVA test [5] between the one-year status and scale variables while the chi-square[6] between one-year status and the nominal variables. From the 16 variables, the related variables with the one year-status will be the variables which we will use to create our model. Therefore, we could create a model with less number of variables by understanding and visualizing the dataset.

However we won't do any data imputation in our statistical model because our dataset is complete and doesn't have any missing values in them.

5 Exploratory data analysis (EDA)

Exploratory data analysis depends on visualization of data by using **R-Language** libraries as **ggplot**.

From point 4 after using correlation, we mentioned that we have two tables one between the one-year status and scale variables while the other between one-year status and the nominal variables so we will use statistical tests to analyze and visualize the one-year status that depend on variables known from point 4.[7]

6 Methodology

6.1 Naive Bayes

Naive Bayes is a classification algorithm for binary (two-class) and multi-class classification problems. The technique is easiest to understand when described using binary or categorical input values

. It is called naive Bayes or idiot Bayes because the calculation of the probabilities for each hypothesis are simplified to make their calculation tractable. Rather than attempting to calculate the values of each attribute value $P(d1, d2, d3|h)$

, they are assumed to be conditionally independent given the target value and calculated as $P(d1|h)P(d2|H)$ and so on.

This is a very strong assumption that is most unlikely in real data, i.e. that the attributes do not interact. Nevertheless, the approach performs surprisingly well on data where this assumption does not hold.[8]

Bayes' Theorem is stated as:

$P(h|d) = \frac{(P(d|h)P(h)}{P(d)}$

Where

- $P(h|d)$ is the probability of hypothesis h given the data d. This is called the posterior probability.
- $P(d|h)$ is the probability of data d given that the hypothesis h was true.
- P(h) is the probability of hypothesis h being true (regardless of the data). This is called the prior probability of h.
- P(d) is the probability of the data (regardless of the hypothesis).

You can see that we are interested in calculating the posterior probability of $P(h|d)$ from the prior probability P(h) with P(d) and $P(d|h)$.

Gaussian Naive Bayes

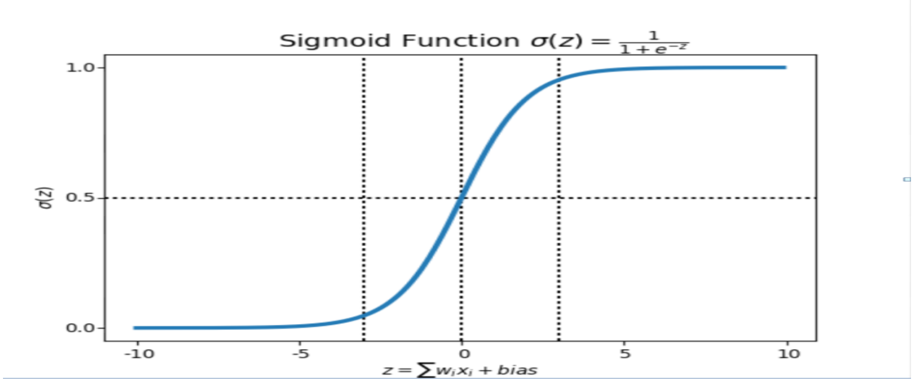
Naive Bayes can be extended to real-valued attributes, most commonly by assuming a Gaussian distribution. This extension of naive Bayes is called Gaussian Naive Bayes. Other functions can be used to estimate the distribution of the data, but the Gaussian (or Normal distribution) is the easiest to work with because you only need to estimate the mean and the standard deviation from your training data.

Advantages of Naive Bayes:

1. When assumption of independent predictors holds true, a Naive Bayes classifier performs better as compared to other models.
2. Naive Bayes requires a small amount of training data to estimate the test data. So, the training period is less.
3. Naive Bayes is also easy to implement.

6.2 Logistic regression

Logistic regression is a technique borrowed by machine learning from the field of statistics.
It is the go-to method for binary classification problems (problems with two class values) in our case it's whether the patient died after one year or not(0,1).
We can call a Logistic Regression a Linear Regression model but the Logistic Regression uses a more complex cost function, this cost function can be defined as the 'Sigmoid function' or also known as the 'logistic function' instead of a linear function.[9]



1. When using linear regression we used a formula of the hypothesis i.e.
   - $h\theta(x) = \beta_0 + \beta_1 X$
2. For logistic regression we are going to modify it a little bit i.e.
   - $\sigma(Z) = \sigma(\beta_0 + \beta_1 X)$
3. We have expected that our hypothesis will give values between 0 and 1.
   - $h\theta(x) = sigmoid(Z)$
   - $h\theta(x) = \frac{1}{1+e^{-}(\beta_0+\beta_1 X)}$

Advantages of Logistic regression:

1. Very efficient
2. Doesn't require too many resources
3. Easy to regularize
4. Easy to implement

6.3 K-nearest neighbors (KNN) mode

KNN is an algorithm that is considered both non-parametric and an example of lazy learning.[10]

- Non-parametric means that it makes no assumptions. The model is made up entirely from the data given to it rather than assuming its structure is normal.
- Lazy learning means that the algorithm makes no generalizations. This means that there is little training involved when using this method. Because of this, all of the training data is also used in testing when using KNN.

Its classification uses k which is the number of its nearest neighbors , step of implementations :

- Calculation of the distance between all points
- Finds the k points that are closest
- The class is chosen by the majority of the surrounding points.

7 Project Schedule

Target	Deadline
Data Pre-proccesing	31 October
Data visualization	7 November
Learning Naive Bayes method	14 November
Learning KNN and logistics regression methods	21 November
Submitting Prototype	25 November
Final submission	15 December

8 Personal Websites

1. Asmaa Mahmoud
2. Alaa Gamal
3. Salma Zakaria
4. Marwa Adel

9 References

[1] Thoracic Surgery | Encyclopedia.com

http://bit.ly/2BTPxzL

[2] The Actual Difference Between Statistics and Machine Learning

http://bit.ly/2PloPI2

[3] UCI Machine Learning Repository: Thoracic Surgery Data Data Set

http://bit.ly/2MNVjsW

[4] (PDF) Analyzing and visualizing Thoracic Surgery Data Set

http://bit.ly/364giit

[5] A Simple Introduction to ANOVA (with applications in Excel)

http://bit.ly/2Wdmj7Y

[6] Chi-Square Test

http://bit.ly/2MP5TzT

[7] Introduction to Data Visualization with ggplot2 | DataCamp

http://bit.ly/2qKiYlo

[8] 6 Easy Steps to Learn Naive Bayes Algorithm

http://bit.ly/2MKPqMN

[9] Logistic Regression for Machine Learning

http://bit.ly/2NahU1H

[10] Machine Learning Basics with the K-Nearest Neighbors Algorithm

http://bit.ly/367NENp