Machine learning applications in cancer prognosis and prediction

Andrii Rozumnyi
25.11.2016
Content

- Introduction
- ML structure
- Process steps in ML
- ML methods for cancer prediction
- Scientific researches
- Data science challenges related to cancer problem
- Homework
Introduction

Cancer is a group of diseases involving abnormal cell growth with the potential to invade or spread to other parts of the body (Malignant tumors).

The most common treatments for cancer are

- surgery
- chemotherapy
- radiation
### Estimated New Cases

<table>
<thead>
<tr>
<th></th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prostate</td>
<td>180,850</td>
<td>246,650</td>
</tr>
<tr>
<td>Lung &amp; bronchus</td>
<td>117,920</td>
<td>106,470</td>
</tr>
<tr>
<td>Colon &amp; rectum</td>
<td>70,520</td>
<td>63,570</td>
</tr>
<tr>
<td>Urinary bladder</td>
<td>59,950</td>
<td>60,050</td>
</tr>
<tr>
<td>Melanoma of the skin</td>
<td>46,870</td>
<td>49,350</td>
</tr>
<tr>
<td>Non-Hodgkin lymphoma</td>
<td>40,170</td>
<td>32,410</td>
</tr>
<tr>
<td>Kidney &amp; renal pelvis</td>
<td>39,650</td>
<td>29,510</td>
</tr>
<tr>
<td>Oral cavity &amp; pharynx</td>
<td>34,780</td>
<td>26,050</td>
</tr>
<tr>
<td>Leukemia</td>
<td>34,090</td>
<td>25,400</td>
</tr>
<tr>
<td>Liver &amp; intrahepatic bile duct</td>
<td>28,410</td>
<td>23,050</td>
</tr>
<tr>
<td>All Sites</td>
<td>841,350</td>
<td>843,820</td>
</tr>
</tbody>
</table>

### Estimated Deaths

<table>
<thead>
<tr>
<th></th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung &amp; bronchus</td>
<td>85,920</td>
<td>72,180</td>
</tr>
<tr>
<td>Prostate</td>
<td>26,120</td>
<td>40,450</td>
</tr>
<tr>
<td>Colon &amp; rectum</td>
<td>26,020</td>
<td>23,170</td>
</tr>
<tr>
<td>Pancreas</td>
<td>21,450</td>
<td>20,330</td>
</tr>
<tr>
<td>Liver &amp; intrahepatic bile duct</td>
<td>18,280</td>
<td>14,240</td>
</tr>
<tr>
<td>Leukemia</td>
<td>14,130</td>
<td>10,470</td>
</tr>
<tr>
<td>Esophagus</td>
<td>12,720</td>
<td>10,270</td>
</tr>
<tr>
<td>Urinary bladder</td>
<td>11,820</td>
<td>8,990</td>
</tr>
<tr>
<td>Non-Hodgkin lymphoma</td>
<td>11,520</td>
<td>8,630</td>
</tr>
<tr>
<td>Brain &amp; other nervous system</td>
<td>9,440</td>
<td>6,510</td>
</tr>
<tr>
<td>All Sites</td>
<td>314,230</td>
<td>281,400</td>
</tr>
</tbody>
</table>

**FIGURE 1. Ten Leading Cancer Types for the Estimated New Cancer Cases and Deaths by Sex, United States, 2016.**

Estimates are rounded to the nearest 10 and cases exclude basal cell and squamous cell skin cancers and in situ carcinoma except urinary bladder.
How can we use Machine Learning in cancer prognosis and prediction?
Introduction

ML can help in:

• prediction of cancer **susceptibility** – risk assessment prior to occurrence.
• prediction of cancer **recurrence** – likelihood of redeveloping.
• prediction of cancer **survivability** – life expectancy, survival, progression, tumor-drug sensitivity.
Machine Learning structure

ML

- Supervised
  - Classification
    - Define a class from finite number of classes
  - Regression
    - Define a relation between variables
- Unsupervised
  - Clustering
    - Discover groups of data
  - Association
    - Rules generation
Process steps in ML

Data processing
- Filling missing data
- Identify/remove outliers
- Data transformation (normalization, aggregation)
- Resolve inconsistencies

Feature engineering
- Feature importance
- Dimensionality reduction (PCA, SVD)
- Feature construction

Model building (+ evaluation)
- Choose method
- Pick parameters
- Results evaluation (Cross-validation)

Ensembling
- Combine different models
- Results evaluation (Cross-validation)
Logistic regression

Breast Cancer Logistic Regression Example

$$P\{y = 1 \mid x\} = f(z) = \frac{1}{1 + e^{-z}}$$

$$z = \theta^T x$$

$x$ – values vector

$\theta$ – parameters vector
Breast cancer classification by DT

Decision tree learned from the Wisconsin Breast Cancer dataset
Random Forest
SVM

\[ K(x_i, x_j) = \phi(x_i)\phi(x_j) \]

- Misclassified point
- Support Vector
- Margin = \( \frac{2}{\sqrt{\mathbf{w}^T\mathbf{w}} \)}
- \( \xi > 1 \)
- \( \xi < 1 \)
- \( \xi = 0 \)
- \( w^T\phi(x) + b = -1 \)
- \( w^T\phi(x) + b = 0 \)
- \( w^T\phi(x) + b = +1 \)
Kernel Trick
Neural Networks
Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties

Experiments

Cell line features

Neural network

Prediction

Microsatellites
Sequence variation
Copy number variation

Physicochemical: lipophilicity
weight, rule of five, etc.
Fingerprints: CDK, PubChem,
Klekota-Roth, Estate, etc.

feat_1
feat_2
feat_3

\( IC_{50} \)

\( \text{drug conc.} \)

\( \text{cell survival} \)

Link: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0061318
Help prevent cervical cancer by identifying at-risk populations

This is a Masters competition. You must be a Kaggle Master to participate.

Cervical cancer is the third most common cancer in women worldwide, affecting over 500,000 women and resulting in approximately 275,000 deaths every year. After reading these statistics, you may be surprised to hear that cervical cancer is potentially preventable and curable.

Cervical cancer can be prevented through early administration of the HPV vaccine and regular pap smear screenings, which indicate the presence of precancerous cells. It is also sometimes curable.
Multiple Myeloma DREAM Challenge

Opening January 2017

The Multiple Myeloma DREAM Challenge provides an opportunity to combine integration of large scale molecular and clinical data and state of the art analytical approaches to facilitate risk stratification in over 25,000 patients in the US alone. Additionally, it provides the ability to benchmark novel methods with the greatest potential to yield patient care benefits in the future.

The Digital Mammography DREAM Challenge.

June 29, 2016 - Feb. 20, 2017 (open)

With generous support from the Laura and John Arnold Foundation this $1.2 million Challenge, one of two large prize Coding4Cancer Challenges, seeks to improve the accuracy of breast cancer detection and reduce the current rate of patient callbacks.

Calling RNA Challenge (SMC-RNA)

Summer 2016 (open)

Leaders from the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA) have come together to develop a Challenge to assess the accuracy of methods to work with Cancer RNA Sequencing data.
Risk reducing factors

- Balanced Diet
- Healthy Weight
- Active Lifestyle
- No Smoking
- Reduce Alcohol
Conclusion

• Cancer is a serious problem which leads to a lot of deaths each year
• ML is actively involved in cancer related problems
• Your lifestyle influences on cancer related risks
Please send your homework to

andrey.rozumnuy@gmail.com

by December 2
10:00 am
Find either data science challenge or scientific paper (that has not been mentioned in the presentation) where cancer related problem has been stated. Answer the following questions:

- What is the goal of the challenge/article?
- Describe the input data.
- Which Machine Learning techniques have been/could be used for problem solving?
- Which difficulties have been/could be faced there?