Case Study: Data Analysis of Cervical Cancer

Bob Gotwals
Computational Science Educator

September 25, 2019

This case study is based on data from `https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29#`([1]).

---

1 Introductory Reading

Cervical cancer in women is a serious condition, and results in a significant number of cancer-related deaths every year. The passage below (from Wikipedia, `https://en.wikipedia.org/wiki/Cervical_cancer` [2]), describes the condition and some of the causative factors.

> Cervical cancer is a cancer arising from the cervix. It is due to the abnormal growth of cells that have the ability to invade or spread to other parts of the body. Early on, typically no symptoms are seen. Later symptoms may include abnormal vaginal bleeding, pelvic pain or pain during sexual intercourse. While bleeding after sex may not be serious, it may also indicate the presence of cervical cancer.
>
> *Human papillomavirus infection* (HPV) causes more than 90% of cases; most people who have had HPV infections, however, do not develop cervical cancer. Other risk factors include smoking, a weak immune system, birth control pills, starting sex at a young age and having many sexual partners, but these are less important. Cervical cancer typically develops from precancerous changes over 10 to 20 years. About 90% of cervical cancer cases are squamous cell carcinomas, 10% are adenocarcinoma and a small number are other types. Diagnosis is typically by cervical screening followed by a biopsy. Medical imaging is then done to determine whether or not the cancer has spread.

HPV vaccines protect against two-to-seven high-risk strains of this family of viruses and may prevent up to 90% of cervical cancers. As a risk of cancer still exists, guidelines recommend continuing regular Pap tests. Other methods of prevention include having few or no sexual partners and the use of condoms. Cervical cancer screening using the Pap test or acetic acid can identify precancerous changes which when treated can prevent the development of cancer. Treatment of cervical cancer may consist of some combination of surgery, chemotherapy and radiation therapy. Five-year survival rates in the United States are 68%. Outcomes, however, depend very much on how early the cancer is detected.

Worldwide, cervical cancer is both the fourth most common cause of cancer and the fourth most common cause of death from cancer in women. In 2012, an estimated 528,000 cases of cervical cancer occurred, with 266,000 deaths. This is about 8% of the total cases and total deaths from cancer. About 70% of cervical cancers occur in developing countries and 90% of deaths. In low-income countries, it is one of the most common causes of cancer death. In developed countries, the widespread use of cervical screening programs has dramatically reduced rates of cervical cancer. In medical research, the most famous immortalized cell line, known as HeLa, was developed from cervical cancer cells of a woman named Henrietta Lacks.

In this case study, you are provided with a dataset presented at the 8th Iberian Conference on Pattern Recognition and Image Analysis ([3]). The front cover of the conference proceedings is shown in Figure 1.

Figure 1: 8th Iberian Pattern Recognition and Image Analysis Conference ([3])

Figure 2 shows the title, authors, abstract, and keywords for the article containing the data. The article is available upon request, but does not contain information of particular interest for this case study.

Transfer Learning with Partial Observability Applied to Cervical Cancer Screening

Kelwin Fernandes[1,2(✉)], Jaime S. Cardoso[1,2], and Jessica Fernandes[3]

[1] INESC TEC, Porto, Portugal
{kafc,jaime.cardoso}@inesctec.pt
[2] Universidade do Porto, Porto, Portugal
[3] Universidad Central de Venezuela, Caracas, Venezuela

Abstract. Cervical cancer remains a significant cause of mortality in low-income countries. As in many other diseases, the existence of several screening/diagnosis methods and subjective physician preferences creates a complex ecosystem for automated methods. In order to diminish the amount of labeled data from each modality/expert we propose a regularization-based transfer learning strategy that encourages source and target models to share the same coefficient signs. We instantiated the proposed framework to predict cross-modality individual risk and cross-expert subjective quality assessment of colposcopic images for different modalities. Thus, we are able to transfer knowledge gained from one expert/modality to another.

Keywords: Transfer learning · Regularization · Cervical cancer · Digital colposcopy

Figure 2: Title, authors, abstract, keywords ([1])

The data for this case study is contained in a comma-separated values (.csv) file found on Canvas under Files → Week 5. Figure 3 shows the attributes of this dataset. As described below, this is a multivariate (many variables) dataset in the life sciences. It is a relatively recent dataset (2017) that studied 858 women and some of the characteristics of both symptoms and behaviors, such as smoking, acquisition of sexually-transmitted diseases (STDs), and the like.

Data Set Characteristics:	Multivariate	**Number of Instances:**	858	**Area:**	Life
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	36	**Date Donated**	2017-03-03
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	93456

Figure 3: Cervical Cancer data set attributes ([1])

The data contains 36 variables, listed below. Also listed is the type of data: int (integer), bool (Boolean, or true/false).

1. (int) Age
2. (int) Number of sexual partners
3. (int) First sexual intercourse (age)

4. (int) Num of pregnancies
5. (bool) Smokes
6. (real) Smokes (years)
7. (real) Smokes (packs/year)
8. (bool) Hormonal Contraceptives
9. (int) Hormonal Contraceptives (years)
10. (bool) IUD
11. (int) IUD (years)
12. (bool) STDs
13. (int) STDs (number)
14. (bool) STDs: condylomatosis
15. (bool) STDs: cervical condylomatosis
16. (bool) STDs: vaginal condylomatosis
17. (bool) STDs: vulvo-perineal condylomatosis
18. (bool) STDs: syphilis
19. (bool) STDs: pelvic inflammatory disease
20. (bool) STDs: genital herpes
21. (bool) STDs: molluscum contagiosum
22. (bool) STDs: AIDS
23. (bool) STDs: HIV
24. (bool) STDs: Hepatitis B
25. (bool) STDs: HPV
26. (int) STDs: Number of diagnosis
27. (int) STDs: Time since first diagnosis
28. (int) STDs: Time since last diagnosis

29. (bool) Dx:Cancer
30. (bool) Dx:CIN
31. (bool) Dx:HPV
32. (bool) Dx
33. (bool) Hinselmann: target variable
34. (bool) Schiller: target variable
35. (bool) Cytology: target variable
36. (bool) Biopsy: target variable

Notice that four of the variables are listed as target variables: Hinselmann, Schiller, Cytology, and Biopsy. These all represent medical tests conducted to diagnose or screen for cervical cancer. Note that this dataset was generated for purposes of image analysis and pattern recognition, primarily using machine learning, so these tests may or may not be targets for your purposes!

There are several articles from Scientific American posted on the Canvas site, under Week 5 → Readings. These can be used or not used as you see fit. You are welcome to use other resources as appropriate.

2 Deliverable

2.1 Part I: Data Cleaning

For this part, your task is to perform standard cleaning of the data set, including (but not limited to):

1. Reading in data to remove non-standard characters
2. Replacement of all NAs with zeros, unless you prefer to fill NAs with other values, such as the mean.
3. Configuring of variable names to be lowercase and separated with hyphens as appropriate
4. Replacement of all Boolean 0s and 1s with logicals (TRUE and FALSE)
5. Exploration of all numerical categories to ensure they are normally distributed. If not, apply the appropriate transformation, test with qqnorm and qqline. Alternatively, you could normalize one or more variables. Any transformed and/or normalized variables should be added as new columns to the modified dataset.

6. Other cleaning modifications as considered to be appropriate.

7. Clean dataset should be written to a csv file entitled "ccdataMod.csv" for use in Part II.

2.2 Part II: Data Analysis

For Part II, your task is to perform analyses of your modified (cleaned) dataset, with a **strong focus** on using the data to understand and explain the science of cervical cancer. The research question here is: "what does the data teach us about cervical cancer? What insights into this insidious disease can be obtained by a data science analysis?"

It is assumed that, in your Markdown report, you will have one- to two-pages of background information about cervical cancer, with an emphasis and focus on the data contained in the dataset. For example, you should describe the different types of STDs somewhere in your report. You might want to consider (but are not required) to display them as a table.

This is an open-ended case study, but, in addition to using the data to learn about cervical cancer, the goal is to demonstrate data science techniques and tools that you have learned up to this point. As such, you should try to **demonstrate** those skills in your analyses. Here is a list of some of the skills you have obtained to date:

1. Importing and describing the dataset, using commands such as names, dim, glimpse, str, and summary

2. Creating basic plots, using commands such as plot, hist, boxplot, and the built-in pairwise correlation/histogram plot

3. Using models – linear regression, multiple linear regression, and BIC models – to look at relationships between two or more variables and for use in determining causality, if any

4. Using specialized packages such as tidyverse/dplyr, stringr, and others as appropriate. Recall that dplyr is known as the "grammar of data science", with functions such as summarize, mutate, arrange. Figure 4 shows a "cheat sheet" of dplyr commands.

5. Other Exploratory Data Analysis (EDA) applications as needed and as appropriate.

BE SURE to not create an object (table, plot, model, etc.) without ALSO describing it and its significance in the report. There should be a fairly substantial amount of narrative in your report, showing evidence that you learned something about cervical cancer (NOT a book report, however), based on your analysis of the data!

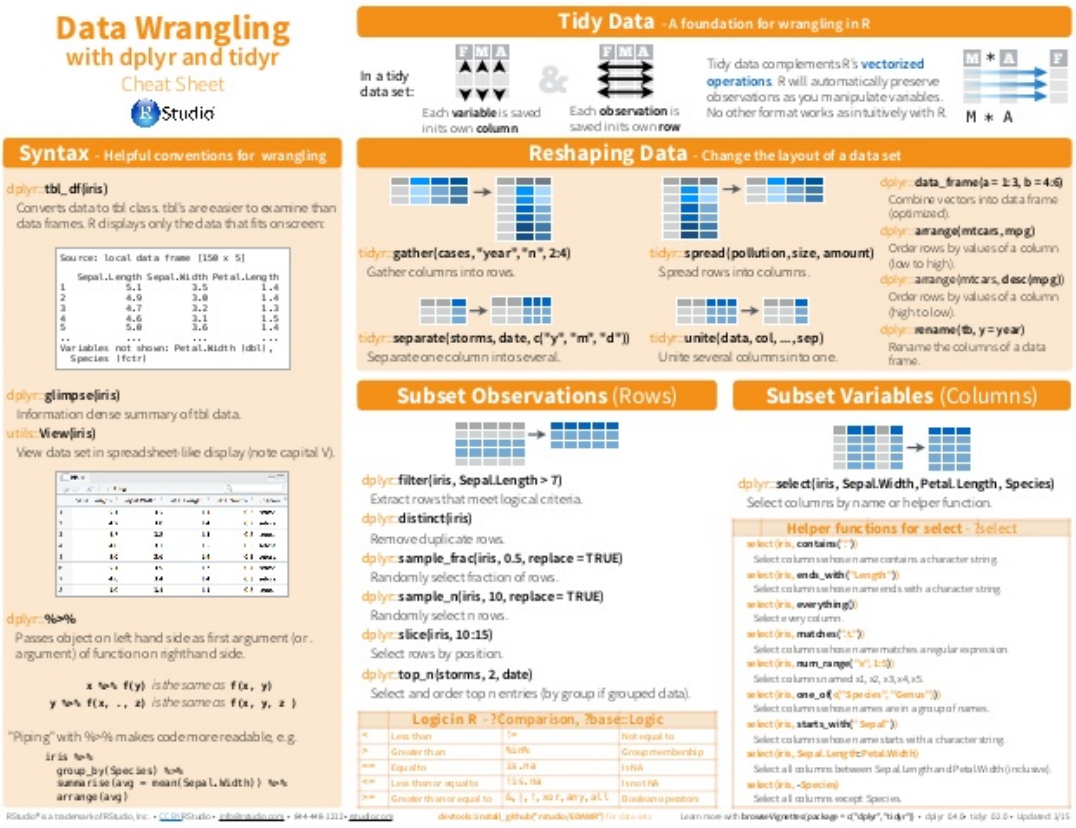

Figure 4: dplyr cheat sheet

References

[1] https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29#. Accessed September 2019.

[2] https://en.wikipedia.org/wiki/Cervical_cancer. Accessed September 2019.

[3] http://www.ibpria.org/2017

[4] https://www.waidroka.com/fiji-shark-dive/. Accessed July 2019.