# Classification of Breast Cancer Subtypes via Gene Expression

**Kevin O'Connor**
Department of Statistics and Operations Research
University of North Carolina - Chapel Hill
Chapel Hill, NC
koconn@live.unc.edu

## Abstract

Breast cancer subtype identification is an important problem with high clinical relevance. In this paper we use various machine learning methods to predict patients' breast cancer subtype from the measured expression levels of a subset of their genes. It is shown that multinomial logistic regression can achieve a moderate level of accuracy whereas modern classification methods like svm and a fully-connected neural network can achieve far greater results. The performance of each method is then evaluated and compared.

## 1 Introduction

For many years now, successful treatments for various types of cancers have eluded re- searchers. Breast cancer in particular continues to affect millions of women and men. An important characteristic in the study of breast cancer is the subtype of disease, identified as Luminal A, Luminal B, Basal, HER2, or Normal. For a single tumor, the subtype determines both how the disease may develop in the patient as well as the proper course of treatment. Thus, there is significant incentive for researchers to find distinctive characteristics of each subtype to more clearly identify them in future patients.

In the past decade, the cost of obtaining genetic information from an individual has de- creased substantially. This has led to a boom in the amount of genetic data available in many areas of medicine. In 2005, a database called The Cancer Genome Atlas (TCGA) was established in order to organize this data and make it widely available to researchers around the world for analysis. At this point, the database contains genetic expression levels for over 20,000 genes from patients with tumors in each of the 5 breast cancer subtypes. It offers an excellent opportunity to look for patterns in the genetic data that distinguish the subtypes.

### 1.1 Code

The code used to perform the analyses was written in R and Python/Keras and can be found at

`https://github.com/oconnor-kevin`

### 1.2 Data

The data can be downloaded from the following website,

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

## 2 Related Work

## 3 Experiments

### 3.1 Multinomial Logistic Regression

In our first experiment, we will fit a multinomial logistic regression model to the data. Given a sample $X_i \in \mathbb{R}^p$ with label $y_i \in \{1, ..., K\}$ and $K$ coefficient vectors, $\{\beta_1, ..., \beta_K\}$, we can write our model as

$$\mathbb{P}(y_i = k) = \frac{\exp\{-X_i\beta_k\}}{\sum_{k'} \exp\{-X_i\beta_{k'}\}}$$

For $n$ independent observations, $X_1, ..., X_n$, this gives us a joint likelihood of the correct classes,

$$\mathcal{L}(\beta_1, ..., \beta_K; \{X_i, y_i\}_{i=1}^n) = \prod_{i=1}^{n} \frac{\exp\{-X_i\beta_{y_i}\}}{\sum_{k'} \exp\{-X_i\beta_{k'}\}}$$

and log-likelihood,

$$\mathcal{LL}(\beta_1, ..., \beta_K; \{X_i, y_i\}_{i=1}^n) = \sum_{i=1}^{n} \left[ -X_i\beta_{y_i} - \log\left( \sum_{k'} \exp\{-X_i\beta_{k'}\} \right) \right]$$

Then in order to fit the model to the data, we maximize the log-likelihood via gradient ascent to find the maximum likelihood estimator of the parameters $\{\beta_1, ..., \beta_K\}$. Note that care has to be taken when computing the second term in the log-likelihood to avoid numerical overflow.

### 3.2 K-means

Next we consider an unsupervised learning approach to identify subgroups. We will apply the K-means algorithm to our data to produce 5 clusters which we hope will correspond to breast cancer subtypes.

### 3.3 SVM

Moving on to more modern classification methods, we will apply an SVM. This will attempt to find a boundary in the data space which maximizes the margin between the support vectors from different classes. Unfortunately for this case, SVM's do not generalize well to multi-class classification problems. Some adaptations have been developed which tackle the multi-class case using a *one vs the rest* approach but we will simplify our problem by just learning to distinguish between a variety of 2-class subtypes of our data.

### 3.4 Fully-connected Neural Network

### 3.5 Citations within the text

## 4 Discussion

## 5 Conclusion

The `natbib` package will be loaded for you by default. Citations may be author/year or numeric, as long as you maintain internal consistency. As to the format of the references themselves, any style is acceptable as long as it is used consistently.

The documentation for `natbib` may be found at

`http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf`

Of note is the command `\citet`, which produces citations appropriate for use in inline text. For example,

```
\citet{hasselmo} investigated\dots
```

produces

> Hasselmo, et al. (1995) investigated…

If you wish to load the `natbib` package with options, you may add the following before loading the `nips_2017` package:

```
\PassOptionsToPackage{options}{natbib}
```

If `natbib` clashes with another package you load, you can add the optional argument `nonatbib` when loading the style file:

```
\usepackage[nonatbib]{nips_2017}
```

### 5.1 Figures

All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction. The figure number and caption always appear after the figure. Place one line space before the figure caption and one line space after the figure. The figure caption should be lower case (except for first word and proper nouns); figures are numbered consecutively.

You may use color figures. However, it is best for the figure captions and the paper body to be legible if the paper is printed in either black/white or in color.

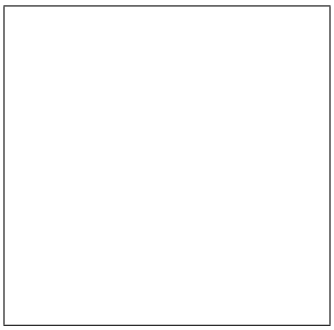

Figure 1: Sample figure caption.

### 5.2 Tables

All tables must be centered, neat, clean and legible. The table number and title always appear before the table. See Table 1.

Place one line space before the table title, one line space after the table title, and one line space after the table. The table title must be lower case (except for first word and proper nouns); tables are numbered consecutively.

Note that publication-quality tables *do not contain vertical rules.* We strongly suggest the use of the `booktabs` package, which allows for typesetting high-quality, professional tables:

`https://www.ctan.org/pkg/booktabs`

This package was used to typeset Table 1.

**Acknowledgments**

## References

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press.

Table 1: Sample table title

| | Part | |
|---|---|---|
| Name | Description | Size ($\mu$m) |
| Dendrite | Input terminal | $\sim$100 |
| Axon | Output terminal | $\sim$10 |
| Soma | Cell body | up to $10^6$ |

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer–Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.

# Appendix

## Data

`https://cmt.research.microsoft.com/NIPS2017/`