note on: Classification Probability

Amin Ahmadi

August 2020

1 Introduction

Logistic Regression is one way to evaluate the posteriori probability in terms of a *linear* function of estimators. In *binary classification*, the probability that an event falls into class 0 is expressed by

$$p(y=0|x) = \frac{1}{1+e^{\beta x}}, \tag{1}$$

and since the sum of probabilities must add up to one; $p(y=0|x) + p(y=1|x) = 1$; the probability to find the event in class 1 is

$$p(y=1|x) = 1 - p(y=0|x) \tag{2}$$

$$= \frac{e^{\beta x}}{1+e^{\beta x}} \tag{3}$$

What we notice is that the ratio of probabilities is

$$\frac{p(y=1|x)}{p(y=0|x)} = e^{\beta x} \tag{4}$$

The same idea can be extended to multi-class classification with ratio probability of given two classes i and j

$$\frac{p(y=k_i|x)}{p(y=k_j|x)} = e^{(\beta_i - \beta_j)x} \tag{5}$$

Again, using the fact that the sum of probabilities is one $\sum_{i=1}^{K} p(y=k_i|x) = 1$, we can realize

$$\begin{aligned}
p(y=k_1|x) &= \frac{e^{\beta_1 x}}{1+\sum_{i=1}^{K} e^{\beta_i x}} \\
p(y=k_2|x) &= \frac{e^{\beta_2 x}}{1+\sum_{i=1}^{K} e^{\beta_i x}} \\
&\vdots \\
p(y=k_{K-1}|x) &= \frac{e^{\beta_{K-1} x}}{1+\sum_{i=1}^{K} e^{\beta_i x}} \\
p(y=k_K|x) &= \frac{1}{1+\sum_{i=1}^{K} e^{\beta_i x}}
\end{aligned}$$

2 Why Boltzmann Distribution

The question that comes to mind is that what is the rationale of the assumption of Boltzmann probability for being in one of the classes. With that we means that the ratio of probabilities to be in two different classes is

$$\frac{p(y = k_i|x)}{p(y = k_j|x)} = e^{(\beta_j - \beta_i)x} \tag{6}$$

Let assume that an event falls into one of K possible classes by a process that is not known to us and can be considered random. We can picture such problem as a ball that is randomly casted into one of K available boxes. The number of ways that these N balls can be distributed between K boxes is

$$\Omega = \frac{N!}{n_1!n_2!\dots n_K!} \tag{7}$$

where n_j is the number of ball in box j (i.e. number of datapoints in class j) with constraint of the total number of balls in all classes should be N, or

$$\sum_{j=1}^{K} n_j = N \tag{8}$$

The other constraint that comes to the picture and plays an important role in the notion of learning is that the underlying process will remain the same. This assumption is justified just because we apply the same trained model to the future dataset. If the process has been changed, the trained model does not perform as it was expected. That means the model has been expired and must be retrained on a new dataset generated by the new process. If this assumption holds, we have the second constraint that the average of the *discrimination function* among classes is constant.

$$\bar{\delta} = \frac{1}{N}\sum_{i=1}^{K} n_i\delta_i(x) \tag{9}$$

This constraint comes from the fact that each box is different from the others and it is specified by a function so called discrimination function. If the discrimination functions are the same for two classes then there is no way to distinguish between these two classes. For now, we take a general form for the discrimination function not necessarily linear.
It is easier to work with the logarithm of big numbers such as Ω

$$\ln(\Omega) = \ln N! - \sum_{i=1}^{K} \ln n_i! \tag{10}$$

We assume that the population of each class is large enough to use Stirling's approximation $\ln n! \sim n(\ln n - 1)$ and write

$$\ln(\Omega) = N(\ln N - 1) - \sum_{i=1}^{K} n_i(\ln n_i - 1) \tag{11}$$

We can work it out further and use $p_i = \frac{n_i}{N}$ to get

$$\begin{aligned}
\ln \Omega &= N \ln N - N - \sum_{i=1}^{K} n_i \ln n_i + \sum_{i=1}^{K} n_i \\
&= N \ln N - \sum n_i \ln n_i \\
&= N \ln N - \sum N p_i \ln(N p_i) \\
&= N \ln N (1 - \sum_i p_i) + N \sum p_i \ln p_i \\
&= N \sum_i p_i \ln p_i
\end{aligned} \tag{12}$$

we recognize the last line is the entropy of the system with an extra N. If we make series of measurements on randomly generated data by the same underlying process, we expect to observe the most probable configuration most of the time. This idea is very close to the notion of equilibrium in thermodynamics.

To find the most probable configuration between others we can maximize Ω or equivalently the entropy subjecting to two constraints that was introduced earlier.

Problem: Maximize

$$\ln \Omega = \sum_i p_i \ln p_i \tag{13}$$

subjecting to two constraints

$$\sum_i p_i = 1, \quad \text{and} \quad \sum_i p_i \delta_i = \bar{\delta} \tag{14}$$

Here we express the expression in terms of class probability p_i instead of class polpulation n_i. One standard approach to a such optimization problem is method of *Lagrange multipliers*. We add the constraints as two zero terms with two undetermined multipliers that we figure out the values later

$$\max \left[\sum_i p_i \ln p_i - \alpha(\sum_i p_i - 1) - \beta(\sum_i p_i \delta_i - \bar{\delta}) \right] \tag{15}$$

that is equivalent to maximizing every term of

$$\max \sum_i \left[p_i \ln p_i - \alpha\, p_i - \beta\, p_i \delta_i \right] = \max \sum_i \Lambda_i \tag{16}$$

Taking derivative respect to p_i, we get

$$\frac{\partial \Lambda_i}{\partial p_i} = \ln p_i + 1 - \alpha - \beta\, \delta_i \tag{17}$$

That gives us

$$p_i = e^{(\alpha - 1)} e^{\beta \delta_i} \tag{18}$$

The value of coefficient α can be determined by application of the first constraint

$$\sum_i p_i = e^{(\alpha-1)} \sum_i e^{\beta\delta_i} = 1 \qquad \Rightarrow \qquad e^{(\alpha-1)} = \frac{1}{\sum_i e^{\beta\delta_i}} = \mathcal{Z} \tag{19}$$

where $\mathcal{Z}$ is called *partition function*. In thermodynamics all properties of a thermal system can be derived just by knowing $\mathcal{Z}$. With that we found that the ratio of probability to be into two different classes is indeed

$$\frac{p_i}{p_j} = e^{\beta(\delta_i - \delta_j)} \tag{20}$$

And this assumption is valid for many classification problems regardless of the nature of underlying process.