Classification

Damien Benveniste

June 17, 2017

1 Introduction to Logistic Regression

The idea is to find an algorithm that learn to classify labels. Here we focus on binary classification (2 labels). For example classify "yes" or "no", "true" or "false", "spam" or "not spam", "will survive" or "will not survived", ...

1.1 The logistic function

The logistic function is defined as:

$$f(z) = \frac{1}{1 + \exp(-z)} \tag{1}$$

We have:

$$\begin{aligned} \lim_{z \to \infty} f(z) &= 1 \\ \lim_{z \to -\infty} f(z) &= 0 \\ f(0) &= \frac{1}{2} \end{aligned} \tag{2}$$

We use this function to estimate the probability $p(y = 1|X; \beta)$:

$$p(y = 1|X; \beta) = \frac{1}{1 + \exp\left(-\sum_{j=0}^{m} \beta_j X_j\right)} \tag{3}$$

We now need to learn the β_j that makes the best estimation.

1.2 Maximum Likelihood Estimation

Let's say we observe a sample pair $(y = 1, X)$. Since we know $y = 1$, we expect the estimation $p(y = 1|X; \beta) \simeq 1$ therefore we want β that maximizes $p(y = 1|X; \beta)$. If we

observe another sample pair $(y = 0, X')$ we should expect $p(y = 1|X'; \beta) \simeq 0$ therefore we want β that minimizes $p(y = 1|X'; \beta)$. We know that

$$\begin{aligned} p(y = 0|X'; \beta) + p(y = 1|X'; \beta) = 1 \\ \Rightarrow p(y = 0|X'; \beta) = 1 - p(y = 1|X'; \beta) \end{aligned} \tag{4}$$

thus equivalently, we want β that maximizes $p(y = 0|X'; \beta)$.

To summarize, if $y = 1$ we want to maximize $p(y = 1|X; \beta)$ but if $y = 0$, we want to maximize $1 - p(y = 1|X; \beta)$. We want to maximize:

$$p(y|X; \beta) = p(y = 1|X; \beta)^y (1 - p(y = 1|X; \beta))^{1-y} \tag{5}$$

We need to perform this optimization process for all the samples available. We define the likelihood function

$$L(\beta) = \prod_{i=1}^{n} p(y^{(i)}|X^{(i)}; \beta) \tag{6}$$

It is usually easier to deal with the log-likelihood function

$$l(\beta) = \log(L(\beta)) = \sum_{i=1}^{n} \log(p(y^{(i)}|X^{(i)}; \beta)) \tag{7}$$

For simplicity, let's redefine $p(y = 1|X; \beta) = h_\beta(X)$. We have

$$l(\beta) = \sum_{i=1}^{n} y^{(i)} \log(h_\beta(X^{(i)})) + (1 - y^{(i)}) \log(1 - h_\beta(X^{(i)})) \tag{8}$$

We can solve this optimization problem by using gradient ascent (because we maximize)

$$\beta_j \leftarrow \beta_j + \alpha \frac{\partial l(\beta)}{\partial \beta_j}. \tag{9}$$

We have

$$\begin{aligned} \frac{\partial \log(h_\beta(X^{(i)}))}{\partial \beta_j} &= \frac{1}{h_\beta(X^{(i)})} \frac{\partial h_\beta(X^{(i)})}{\partial \beta_j} \\ &= \frac{1}{h_\beta(X^{(i)})} h_\beta(X^{(i)})(1 - h_\beta(X^{(i)})) X_j^{(i)} \end{aligned} \tag{10}$$

and

$$\begin{aligned} \frac{\partial \log(1 - h_\beta(X^{(i)}))}{\partial \beta_j} &= -\frac{1}{1 - h_\beta(X^{(i)})} \frac{\partial h_\beta(X^{(i)})}{\partial \beta_j} \\ &= \frac{1}{h_\beta(X^{(i)}) - 1} h_\beta(X^{(i)})(1 - h_\beta(X^{(i)})) X_j^{(i)} \end{aligned} \tag{11}$$

Therefore

$$\begin{aligned}\frac{\partial l(\beta)}{\partial \beta_j} &= \left(\frac{y^{(i)}}{h_\beta(X^{(i)})} + \frac{(1-y^{(i)})}{h_\beta(X^{(i)})-1}\right) h_\beta(X^{(i)})(1-h_\beta(X^{(i)}))X_j^{(i)} \\ &= -\left(y^{(i)}(h_\beta(X^{(i)})-1) + (1-y^{(i)})h_\beta(X^{(i)})\right) X_j^{(i)} \\ &= \left(y^{(i)} - h_\beta(X^{(i)})\right) X_j^{(i)} \end{aligned} \tag{12}$$

We finally have

$$\beta_j \leftarrow \beta_j + \alpha \left(y^{(i)} - h_\beta(X^{(i)})\right) X_j^{(i)}. \tag{13}$$

2 Classification Metrics

2.1 The confusion matrix

Here is the confusion matrix

		predicted condition	
	total population	prediction positive	prediction negative
true condition	condition positive	**True Positive (TP)**	**False Negative (FN)** (type II error)
	condition negative	**False Positive (FP)** (Type I error)	**True Negative (TN)**

There are few important quantity to know:

- The accuracy:

$$a = \frac{\sum TP + \sum TN}{\sum TP + \sum TN + \sum FP + \sum FN} \tag{14}$$

- The misclassification error:

$$\epsilon = 1 - a \tag{15}$$

- The true positive rate:

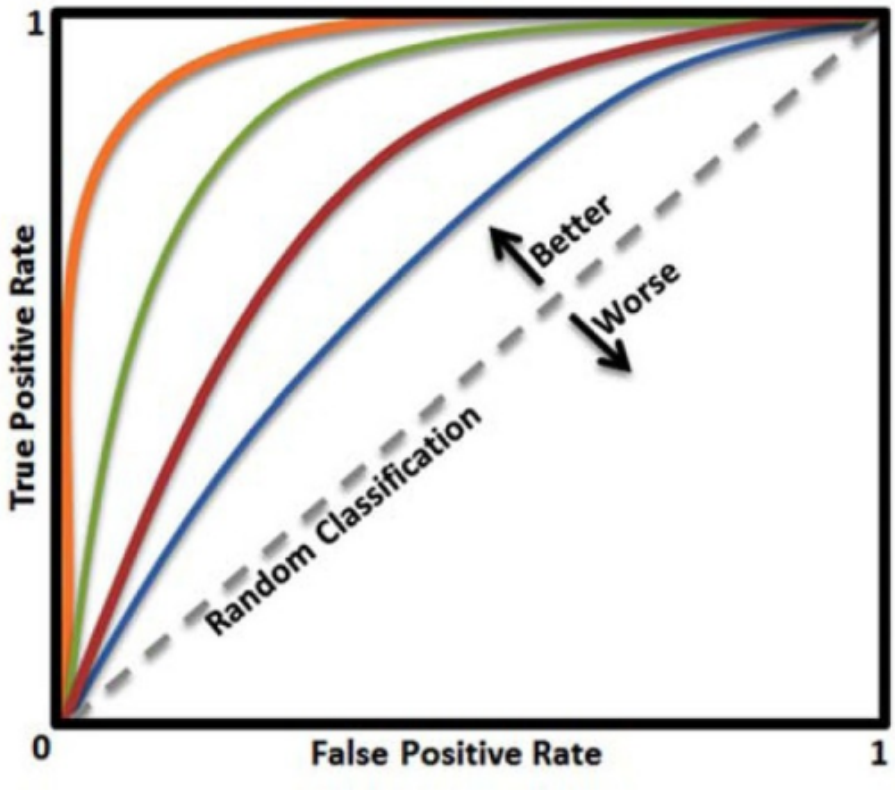
$$TPR = \frac{\sum TP}{\sum TP + \sum FN} \tag{16}$$

- The false positive rate:

$$TPR = \frac{\sum FP}{\sum FP + \sum TN} \tag{17}$$

2.2 Receiver operating characteristic curve

The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.



A particularly important metric for classification is the area under the ROC curve.