Optimal Features for Classification Models

Stephen Vardeman
Analytics Iowa LLC
ISU Statistics and IMSE

K=2 and the "likelihood ratio"

In a K-class classification model, where y takes values in $\mathcal{G} = \{0, 1, \ldots, K-1\}$, P has K conditional distributions for $\mathbf{x}|y$ specified by densities

$$p(\mathbf{x}|0), p(\mathbf{x}|1), \ldots, p(\mathbf{x}|K-1)$$

Statistical theory concerning "minimal sufficiency" promises that (regardless of the size of p) there is a $(K-1)$-dimensional feature that carries all available information about y encoded in $\mathbf{x}$.

For the $K = 2$ case, the 1-dimensional **likelihood ratio statistic**

$$\mathcal{L}(\mathbf{x}) = \frac{p(\mathbf{x}|1)}{p(\mathbf{x}|0)}$$

is "minimal sufficient." If one knew the value of $\mathcal{L}(\mathbf{x})$ one would know all $\mathbf{x}$ has to say about y. An optimal single feature is $\mathcal{L}(\mathbf{x})$ (or any strictly monotone transform of it). The closer that one can come to engineering features "like" $\mathcal{L}(\mathbf{x})$, the more parsimoniously one represents $\mathbf{x}$.

Toy example

To make clear what is meant by the ratio $p(\mathbf{x}|1)/p(\mathbf{x}|0)$, below are $K = 2$ hypothetical probability mass functions for $(p = 2)$ observations $\mathbf{x}$. The likelihood ratio gives the proper ordering of the 9 possible values of $\mathbf{x}$ for classification purposes.

$p(\mathbf{x}|1)$

$x_1 \setminus x_2$	1	2	3
3	.1	.2	.3
2	.1	0	.1
1	.05	.1	.05

$p(\mathbf{x}|0)$

$x_1 \setminus x_2$	1	2	3
3	.15	.2	.1
2	.01	.14	.05
1	.2	0	.15

$p(\mathbf{x}|1)/p(\mathbf{x}|0)$

$x_1 \setminus x_2$	1	2	3
3	2/3	1	3
2	10	0	2
1	1/4	∞	1/3

From least indicative of class 1 to most indicative, these are $(2,2)$, $(1,1)$, $(1,3)$, $(3,1)$, $(3,2)$, $(2,3)$, $(3,3)$, $(2,1)$, and $(1,2)$.

K>2 and likelihood ratios

For $K > 2$, roughly speaking $K - 1$ ratios $p(\mathbf{x}|k) / p(\mathbf{x}|0)$ form a minimal sufficient statistic for the model. This potentially isn't quite true because of possible problems where $p(\mathbf{x}|0) = 0$. But it is true that with $s(\mathbf{x}) = \sum_{k=0}^{K-1} p(\mathbf{x}|k)$ the vector

$$\left(\frac{p(\mathbf{x}|1)}{s(\mathbf{x})}, \frac{p(\mathbf{x}|2)}{s(\mathbf{x})}, \ldots, \frac{p(\mathbf{x}|K-1)}{s(\mathbf{x})} \right)$$

(and many variants of it) is minimal sufficient. To the extent that one can engineer features approximating these $K - 1$ ratios[1], one can parsimoniously represent the input vector.

[1] These are the $K - 1$ conditional probabilities $P[y = 1|\mathrm{x}], \ldots, P[y = K-1|\mathrm{x}]$ for the case where each $P[y = k] = 1/K$.