We say that $X \in \{-1,1\}$ has Rademacher distribution with parameter $b \in [-1,1]$, denoted $X \sim \mathrm{R}(b)$, if

$$\frac{1}{2}+\frac{b}{2}=\mathbb{P}(X=1)=1-\mathbb{P}(X=-1).$$

All distributions on $\{-1,1\}$ are of this form.

Now consider the following process: draw $Y \sim \mathrm{R}(0)$ and conditioned on Y, draw $X_1,\ldots,X_n$ iid $\sim Y\cdot\mathrm{R}(b)$. In words, a fair coin Y is flipped followed by n biased coins, independent conditionionally on Y, biased (for $b>0$) to agree with Y. A *decision rule* $h:\{-1,1\}^n \to \{-1,1\}$ is any function that tries to "guess" Y based on the n observations X_i.

It is a classic fact that the **simple majority vote**

$$h_n^{\mathrm{SMV}}(X_1,\ldots,X_n)=\mathrm{sign}\left(\sum_{i=1}^n X_i\right) \tag{1}$$

is the optimal decision rule, in the sense of being the unique minimizer of the probability of error

$$\mathrm{opt}(n,b):=\min_{h:\{-1,1\}^n\to\{-1,1\}}\mathbb{P}(h(X_1,\ldots X_n)\neq Y)=\mathbb{P}(h_n^{\mathrm{SMV}}(X_1,\ldots X_n)\neq Y) \tag{2}$$

(we leave sign(0) undefined since in this case opt $\equiv \frac{1}{2}$ and the choice of h makes no difference). This is proved, for example, in Slides 3–11 of these lecture notes: `https://www.cs.bgu.ac.il/~inabd161/wiki.files/lecture10_handouts.pdf`

A simple and intuitive expression for $\mathrm{opt}(n,b)$ was given in [1, Equation (25)]:

Theorem 0.1

$$\mathrm{opt}(n,b)=\frac{1}{2}\left(1-\frac{1}{2}\left\|\mathrm{R}(b)^n-\mathrm{R}(-b)^n\right\|_1\right), \tag{3}$$

where $\mathrm{R}(\cdot)^n$ *is the corresponding product distribution on* $\{-1,1\}^n$ *and* $\frac{1}{2}\left\|\cdot\right\|_1$ *is the total variation norm.*

As a consistency check, note that $0\le \mathrm{opt}\le\frac{1}{2}$, that the "worst" value of b is 0 (since $\mathrm{opt}(\cdot,0)=\frac{1}{2}$), and that $\mathrm{opt}(n,b)$ decreases monotonically to 0 as $b\uparrow 1$.

Proof: Define $Z_i=YX_i$ and note that

$$h^{\mathrm{SMV}}(X_1,\ldots,X_n)\neq Y \quad\Longleftrightarrow\quad \sum_{i=1}^n Z_i\le 0.$$

Define $p = \frac{1}{2} + \frac{b}{2}$, $q = 1 - p$, $w = \log(p/q)$. Now for $\boldsymbol{z} = (z_1, \ldots, z_n)$,

$$\begin{aligned}
\sum_{i=1}^n z_i \le 0 \quad &\iff \quad \sum_{i=1}^n w z_i \le 0 \\
&\iff \quad \exp\left(\sum_{i=1} w z_i\right) \le 1 \\
&\iff \quad \prod_{i=1}^n e^{w z_i} \le 1 \\
&\iff \quad \prod_{i:z_i=1}^n \frac{p}{q} \cdot \prod_{i:z_i=-1}^n \frac{q}{p} \le 1 \\
&\iff \quad \mathbb{P}(\boldsymbol{Z} = \boldsymbol{z}) \le \mathbb{P}(\boldsymbol{Z} = -\boldsymbol{z}).
\end{aligned}$$

So:

$$\begin{aligned}
\mathbb{P}(h^{\text{SMV}}(\boldsymbol{X}) \neq Y) &= \mathbb{P}\left(\sum_{i=1}^n Z_i \le 0\right) \\
&= \sum_{\boldsymbol{z} \in \{-1,1\}^n} \mathbb{P}(\boldsymbol{Z} = \boldsymbol{z}) \mathbb{1}_{\{\mathbb{P}(\boldsymbol{Z}=\boldsymbol{z}) \le \mathbb{P}(\boldsymbol{Z}=-\boldsymbol{z})\}} \\
&= \frac{1}{2} \sum_{\boldsymbol{z} \in \{-1,1\}^n} \left[\mathbb{P}(\boldsymbol{Z} = \boldsymbol{z}) \mathbb{1}_{\{\mathbb{P}(\boldsymbol{Z}=\boldsymbol{z}) \le \mathbb{P}(\boldsymbol{Z}=-\boldsymbol{z})\}} + \mathbb{P}(\boldsymbol{Z} = -\boldsymbol{z}) \mathbb{1}_{\{\mathbb{P}(\boldsymbol{Z}=-\boldsymbol{z}) \le \mathbb{P}(\boldsymbol{Z}=\boldsymbol{z})\}}\right] \\
&= \frac{1}{2} \sum_{\boldsymbol{z} \in \{-1,1\}^n} \min\left\{\mathbb{P}(\boldsymbol{Z} = \boldsymbol{z}), \mathbb{P}(\boldsymbol{Z} = -\boldsymbol{z})\right\}.
\end{aligned}$$

Finally, recall that for $x, y \in \mathbb{R}$, we have $\min\{x, y\} = \frac{1}{2}(x+y-|x-y|)$, and hence for all distributions $P = (p_1, \ldots, p_N)$ and $Q = (q_1, \ldots, q_N)$, we have

$$\begin{aligned}
\sum_{i=1}^N \min\{p_i, q_i\} &= \frac{1}{2} \sum_{i=1}^N (p_i + q_i - |p_i - q_i|) \\
&= 1 - \frac{1}{2} \sum_{i=1}^N |p_i - q_i| \\
&= 1 - \frac{1}{2} \|P - Q\|_1.
\end{aligned}$$

■

References

[1] Daniel Berend and Aryeh Kontorovich. A finite sample analysis of the naive bayes classifier. *Journal of Machine Learning Research*, 16:1519–1545, 2015.