Proof that the Bayes Decision Rule is Optimal

Theorem *For any decision function* $g : \mathbb{R}^d \xrightarrow{g} \{0, 1\}$,

$$Pr\{g(\mathbf{X})! = Y\} \geq Pr\{g^*(\mathbf{X})! = Y\}$$

We'll prove it in the 2-class problem.

Proof
First we concentrate the attention on the error rate (probability of classification error) of the generic decision function $g(\cdot)$. Look at a SPECIFIC feature vector (namely, condition on $\mathbf{X} = \mathbf{x}$), and recall that uppercase letters denote a random variable, while a lowercase letter denotes a value.

$\Pr\{g(\mathbf{X}) \neq Y \mid \mathbf{X} = \mathbf{x}\} = 1 - \Pr\{g(\mathbf{X}) = Y \mid \mathbf{X} = \mathbf{x}\}$

so, when $\mathbf{X} = \mathbf{x}$ the probability of error is 1 minus the probability of correct decision. We make a correct decision if $g(\mathbf{X}) = 1$ and $Y = 1$ OR if $g(\mathbf{X}) = 0$ and $Y = 0$. Note that the events are disjoint, so the probability of the union (OR) is the sum of the probabilities

$1 - \Pr\{g(\mathbf{X}) = Y \mid \mathbf{X}\} = 1 - \Pr\{g(\mathbf{X}) = 1, Y = 1 \mid \mathbf{X} = \mathbf{x}\} - \Pr\{g(\mathbf{X}) = 0, Y = 0 \mid \mathbf{X} = \mathbf{x}\}.$

We now show that **conditional on $\mathbf{X} = \mathbf{x}$, the events $\{g(\mathbf{X}) = k\}$ and $\{Y = k\}$ are independent** (surprising, isn't it?).

First, note that conditional on $\mathbf{X} = \mathbf{x}$, $g(\mathbf{X}) = g(\mathbf{x})$, and that, therefore, $g(\mathbf{x})$ is just the value of g evaluated at $\mathbf{x}$. This is either 0 or 1.

Assume WLOG that $g(\mathbf{x}) = 1$. Then $\Pr\{g(\mathbf{x}) = 0, Y = 0 \mid \mathbf{X} = \mathbf{x})\}$ is equal to zero, because $g(\mathbf{x})$ is equal to 1. Note, therefore, that the event $\{g(\mathbf{X}) = 1\}$ has probability 0, and is conditionally independent of the event $Y = 0$ given $\mathbf{X} = \mathbf{x}$. therefore:

$\Pr\{g(\mathbf{X}) = 0, Y = 0 \mid \mathbf{X} = \mathbf{x})\} = \Pr\{g(\mathbf{X}) = 0 \mid \mathbf{X} = \mathbf{x}\} \Pr\{Y = 0 \mid \mathbf{X} = \mathbf{x})\}.$

Similarly, $\Pr\{g(\mathbf{x}) = 1, Y = 1 \mid \mathbf{X} = \mathbf{x})\} = \Pr\{Y = 1 \mid \mathbf{X} = \mathbf{x}\}$ because, by assumption, $\Pr\{g(\mathbf{X} = 1) \mid \mathbf{X} = \mathbf{x}\} = 1$:
BUT an event having probability 1 is independent of any other event (can you prove it ?), then

$$\Pr\{g(\mathbf{X}) = 1, Y = 1 \mid \mathbf{X} = \mathbf{x}\} = \Pr\{g(\mathbf{X}) = 1 \mid \mathbf{X} = \mathbf{x}\} \Pr\{Y = 1 \mid \mathbf{X} = \mathbf{x}\}$$

by definition of independence.

Thus, for each $\mathbf{x}$ where $g(\mathbf{x}) = 1$,

$$\Pr\{g(\mathbf{X}) = k, Y = k \mid \mathbf{X} = \mathbf{x})\} = \Pr\{g(\mathbf{X}) = k \mid \mathbf{X} = \mathbf{x}\} \Pr\{Y = k \mid \mathbf{X} = \mathbf{x})\},$$

for $k = 0, 1$, and independence for this case is proved.

The same argument applies for each $\mathbf{x}$ where $g(\mathbf{x}) = 0$: thus we can always write

$$\Pr\left\{g(\mathbf{X}) = k, Y = k \mid \mathbf{X} = \mathbf{x})\right\} = \Pr\left\{g(\mathbf{X}) = k \mid \mathbf{X} = \mathbf{x}\right\} \Pr\left\{Y = k \mid \mathbf{X} = \mathbf{x})\right\},$$

for $k = 0, 1$, which concludes the independence proof.

Now note that $\Pr\left\{g(\mathbf{X}) = k \mid \mathbf{X} = \mathbf{x}\right\} = 1$ if $g(\mathbf{x}) = k$, and $= 0$ if $g(\mathbf{x}) \neq k$. By using the notation 1_A to denote the the indicator of the set A, we can write:

$$1 - \Pr\left\{g(\mathbf{X}) = Y \mid \mathbf{X}\right\} = 1 - \left(1_{g(\mathbf{x})=1} \Pr\left\{Y = 1 \mid \mathbf{X} = \mathbf{x}\right\} + 1_{g(\mathbf{x})=0} \Pr\left\{Y = 0 \mid \mathbf{X} = \mathbf{x}\right\}\right),$$

Let's now subtract $\Pr\left\{g(\mathbf{X}) = Y \mid \mathbf{X} = \mathbf{x}\right\}$ from $\Pr\left\{g^*(\mathbf{X}) = Y \mid \mathbf{X} = \mathbf{x}\right\}$:

$$\begin{aligned}
\Pr\left\{g^*(\mathbf{X}) = Y \mid \mathbf{X} = \mathbf{x}\right\} &- \Pr\left\{g(\mathbf{X}) = Y \mid \mathbf{X} = \mathbf{x}\right\} \\
&= \Pr\left\{Y = 1 \mid \mathbf{X} = \mathbf{x}\right\} \left(1_{g^*(\mathbf{x})=1} - 1_{g(\mathbf{x})=1}\right) \\
&\quad + \Pr\left\{Y = 0 \mid \mathbf{X} = \mathbf{x}\right\} \left(1_{g^*(\mathbf{x})=0} - 1_{g(\mathbf{x})=0}\right)
\end{aligned}$$

(simple algebra). Noting that $\Pr\left\{Y = 0 \mid \mathbf{X} = \mathbf{x}\right\} = 1 - \Pr\left\{Y = 1 \mid \mathbf{X} = \mathbf{x}\right\}$, we can then write

$$\begin{aligned}
\Pr\left\{g^*(\mathbf{X}) = Y \mid \mathbf{X} = \mathbf{x}\right\} &- \Pr\left\{g(\mathbf{X}) = Y \mid \mathbf{X} = \mathbf{x}\right\} \\
&= \Pr\left\{Y = 1 \mid \mathbf{X} = \mathbf{x}\right\} \left(1_{g^*(\mathbf{x})=1} - 1_{g(\mathbf{x})=1}\right) \\
&\quad + \left(1 - \Pr\left\{Y = 1 \mid \mathbf{X} = \mathbf{x}\right\}\right) \left(1_{g^*(\mathbf{x})=0} - 1_{g(\mathbf{x})=0}\right) \qquad (1)
\end{aligned}$$

Now, note that $1_{g^*(\mathbf{X})=0} = 1 - 1_{g^*(\mathbf{X})=1}$, etc. Hence,

$$\begin{aligned}
\Pr\left\{g^*(\mathbf{X}) = Y \mid \mathbf{X} = \mathbf{x}\right\} &- \Pr\left\{g(\mathbf{X}) = Y \mid \mathbf{X} = \mathbf{x}\right\} \\
&= \Pr\left\{Y = 1 \mid \mathbf{X} = \mathbf{x}\right\} \left(1_{g^*(\mathbf{x})=1} - 1_{g(\mathbf{x})=1}\right) \\
&\quad + \left(1 - \Pr\left\{Y = 1 \mid \mathbf{X} = \mathbf{x}\right\}\right) \left(1 - 1_{g^*(\mathbf{x})=1} - 1 + 1_{g(\mathbf{x})=1}\right) \\
= &\quad \left(2\Pr\left\{Y = 1 \mid \mathbf{X} = \mathbf{x}\right\} - 1\right) \left(1_{g^*(\mathbf{x})=1} - 1_{g(\mathbf{x})=1}\right) \qquad (2)
\end{aligned}$$

Now, note that, for each $\mathbf{x}$,

- if $\Pr\left\{Y = 1 \mid \mathbf{X} = \mathbf{x}\right\} > 1/2$, then by definition of the Bayes Decision Rule, $1_{g^*(\mathbf{x})=1} = 1$, and, in general $1_{g(\mathbf{x})=1} \leq 1$; thus, Eq $2 \geq 0$.
- if $\Pr\left\{Y = 1 \mid \mathbf{X} = \mathbf{x}\right\} < 1/2$, then again by definition the Bayes Decision Rule, $1_{g^*(\mathbf{x})=1} = 0$, and, in general $1_{g(\mathbf{x})=1} \geq 0$; thus, Eq $2 \geq 0$.

This is true for $\mathbf{X} = \mathbf{x}$; Now, take the expectation with respect to $f(\mathbf{X})$. ■