(1) For two states \( y \in \{-1, 1\} \), the decision rule depends on the discriminant function:

\[
g(x) = \log \frac{P(x|y=1)}{P(x|y=-1)} + \log \frac{P(y=1)}{P(y=-1)}
\]

Decision rule is of form:

\[
\alpha(x) = 1 \quad \text{if} \quad g(x) > T
\]

\[
\alpha(x) = -1 \quad \text{if} \quad g(x) < T
\]

\( T \) is a threshold determined by the loss function.

Call \( y = 1 \) targets

\( y = -1 \) non-targets

Two types of errors:

- False Positives - non-targets incorrectly classified as targets
- False Negatives - targets incorrectly classified as non-targets.
(2) \textbf{Eq. Two Gaussians}

Adjust the threshold $T$ to trade-off the different costs of false positives and false negatives. (As given by the loss function).

\textit{ROC Receiver Operating Curve} plots the proportion of correct detection of targets as a function of the false positives. As the threshold $T$ varies,

\[
p(\hat{y}=1|y=1)\]

\[
\hat{y} \text{ is the decision made by the decision rule } \hat{\pi}_T(x).
\]

for threshold $T$,

\[
\hat{y}(x)=1, \text{ if } \log \frac{p(y=1|x)}{p(y=-1|x)} \geq T.
\]
\[
\hat{y} | y = \begin{cases} 
1, & \log \frac{P(x|y=1)}{P(x|y=-1)} > T \\
0, & \text{otherwise}
\end{cases}
\]

\[
\hat{y} = -1| x = 1, \quad \begin{cases} 
\log \frac{P(x|y=1)}{P(x|y=-1)} < T \\
0, & \text{otherwise}
\end{cases}
\]

ROC curves are useful if you want to vary the threshold.

They were developed in the 2nd World War to characterize the effectiveness of radar for detecting enemy aircraft.

ROC curves have many interesting properties, but most are irrelevant for Machine Learning.
Bayes Decision Theory can be applied directly to classification into multiple classes: e.g. $y \in \{1, 2, \ldots, N\}$.

**E.g.**

It can also be applied to estimating continuous variables $y$. This is less relevant for this course.
The Curse of Dimensionality

The examples of Bayes Decision theory are misleading because they are given in low-dimensional spaces (1-dim, or 2-dim).

Many pattern classification tasks occur in high-dimensional spaces. In these spaces our geometric intuitions are often wrong.

Example: Consider the volume of a sphere of radius $r=1$ in $D$ dimensions.

What fraction of its volume lies in the region between $1-\epsilon < r < 1$?

$$V_D(r) = \frac{1}{2} \pi^D r^D$$

$$V_D(1) - V_D(1-\epsilon) = 1 - (1-\epsilon)^D$$

$$\frac{V_D(1) - V_D(1-\epsilon)}{V_D(1)}$$

For large $D$, the volume fraction tends to 1 even for small $\epsilon$.

Most of the volume is at the boundary!
(6) e.6. Behaviour of a Gaussian distribution.

$$p(x) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$$

$$p(x_1, x_2) = \frac{1}{2\pi \sigma^2} e^{-\frac{(x_1^2 + x_2^2)}{2\sigma^2}}$$

Let $$r = \sqrt{x_1^2 + x_2^2}$$.

Then $$p(r) = \frac{r}{2\pi \sigma^2} e^{-\frac{r^2}{2\sigma^2}}$$

In higher dimensions

$$p(r) = \frac{r^{D-1}}{\kappa} e^{-\frac{r^2}{2\sigma^2}}$$

So in high dimensions most of the probability mass of the Gaussian is concentrated on a thin shell away from the center of the Gaussian.
Learning probability distributions in high dimensions can require a lot of data. E.g., Gaussian distribution in D dimensions.

\[ \text{mean} = \mu \text{ in } D \text{ dimensions}. \]

\[ \text{covariance} = \Sigma = D(D+1) \text{ dimension}. \]

This is \( O(D^2) \), not too bad.

But suppose we represent the data by a histogram with \( B \) bins per dimension.

\( B \text{ bins in } D = 1 \)

\( B^2 \text{ bins in } D = 2 \)

\( B^D \text{ bins in } D \text{ dimensions.} \)

Exponential growth! Requires exponential amount of data to learn the distribution.
(8) How to deal with the curse of dimensionality?

In practice, data typically lies on some low-dimensional surface in the high-dimensional space.

So the effective dimension of the data may be a lot smaller than the dimension of the space.

(+) Dimension Reduction Methods attempt to reduce the dimension by seeking the low-dimensional surface. (Not always easy).

(+) Modeling, if we can guess distribution for the data (e.g. Gaussian) then dependence on the dimension is not too bad.

(+) Concentrate on the Decision Boundary - there may be enough data to learn the decision boundary even if we cannot learn the distributions.