1 Exercise 2.1

Suppose each of K classes has an associated target t_k, which is a vector of all zeros, except a one in the k-th position. Show that classifying to the largest element of $\hat{y}$ amounts to choosing the closest target, $argmin_k||t_k - \hat{y}||$, if the elements of $\hat{y}$ sum to one.

1.1 Interpretation

This problem can be interpreted as a "target coding scheme". This specific target coding is can be found in the following papers:

`http://ieeexplore.ieee.org/document/88570/`
`https://www.researchgate.net/publication/3191927_Optimized_feature_extraction_and_the_Bayes_decision_in_feed-forwardclassifier_networks`

> This may be viewed as a gain matrix in which the gain of assigning to class j a pattern which belongs to class i is zero $(i \neq j)$, but is unity for correct classification.

`http://personal.ie.cuhk.edu.hk/~ccloy/files/aaai_2015_target_coding.pdf`

> The 1-of-K coding, containing vectors of length K, with the k-th element as one and the remaining zeros, is typically used along with a softmax function for classification. Each element in a 1-of-K code represents a probability of a specific class.

This paper also contains a more general definition of target coding, but it is not necessary for this exercise.

Following, we can say that:

$$t = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_K \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

In the problem, $\hat{y}$ can be considered as the prediction of a vector x such that y_k is the probability of x being of the class k. That is why the sum of y_k is equal to 1.This also demands that y_k are greater than 0, although this does not make difference for this particular exercise.

And now we arrive to the proof part: the k in which the $\hat{y}_k$ is the greatest, the class that we would choose in our prediction is the same k of the t_k that is nearer to the $\hat{y}$. In mathematical symbols:

$$\arg\min_k ||\hat{y} - t_k||$$

1.2 Solution

$$K > 0$$

$$x \in \Re^N$$
$$y = f(x)$$
$$\sum_{i=1}^{K} y_i = 1$$
$$y_i \geq 0, i \in 1, ..., K$$

$$1 \geq k \geq K$$
$$t_k = (t_{k1}, ..., t_{ki}, ..., t_{kK}), t_{ki} = \begin{cases} 0 & \text{if } i \neq k \\ 1 & \text{if } i = k \end{cases}$$
$$\begin{aligned} \arg\min_k ||t_k - \hat{y}|| &= \arg\min_k ||y - t_k||^2 \\ &= \arg\min_k \sum_{i=1}^{K} (y_i - t_{ki})^2 \\ &= \arg\min_k \sum_{i=1}^{K} (y_i^2 - 2y_i t_{ki} + t_{ki}^2) \\ &= \arg\min_k \Big[\sum_{i=1}^{K} y_i^2 + \sum_{i=1}^{K} (-2y_i t_{ki} + t_{ki}^2) \Big] \end{aligned}$$

Since $\sum_{i=1}^{K} y_i^2$ is a constant and does not depend on k

$$\begin{aligned} &= \arg\min_k \sum_{i=1}^{K} (-2y_i t_{ki} + t_{ki}^2) \\ &= \arg\min_k \Big[\sum_{i=1}^{K} (-2y_i t_{ki}) + \sum_{i=1}^{K} t_{ki}^2 \Big] \end{aligned}$$

Since $\sum_{i=1}^{K} t_{ki}^2 = 1$

$$= \arg\min_k \Big[\sum_{i=1}^{K} (-2y_i t_{ki}) + 1 \Big]$$

Since $\sum_{i=1}^{K} y_i t_{ki} = y_k$, because t_k is zero in all but $i = k$ position.

$$
\begin{aligned}
&= \arg\min_k \big[-2 * \sum_{i=1}^{K} y_i t_{ki} + 1 \big] \\
&= \arg\min_k \big[-2y_k + 1 \big] && \text{(argmax plus constant)} \\
&= \arg\min_k \big[-2y_k \big] && \text{(argmax times constant)} \\
&= \arg\min_k \big[-y_k \big] && \text{(argmax inverse argmin)} \\
&= \arg\max_k \big[y_k \big] && \text{(argmax inverse argmin)}
\end{aligned}
$$

Which give us that:

$$\arg\min_k ||\hat{y} - t_k|| = \arg\max_k \big[y_k \big]$$

This problem can be interpreted as the proof of why in book the author says:

> With the 0-1 loss function [...] the solution is known as the Bayes classifier, and says that we classify to the most probable class, using the conditional (discrete) distribution $Pr(G|X)$.

2 Exercise 2.2

From the previous exercise, we know that the Bayes Classifier, will classify the point to the most probable class. So, the boundary is located exactly where there is no clear most probable class. To simplify let us imagine that we have just 2 classes.

$$
\begin{aligned}
P(K_1|X) &= P(K_2|X) \\
\frac{P(K_1, X)}{P(X)} &= \frac{P(K_2, X)}{P(X)} \\
\frac{P(X|K_1)P(K_1)}{P(X)} &= \frac{P(X|K_2)P(K_2)}{P(X)}
\end{aligned}
$$

$P(X|K_i)$ is the same as the likelihood of seeing X using a generative model from the i-th class. If we suppose that all classes come from a Gaussian Distribution, this simplify to:

$$\frac{\mathcal{L}(\theta_1|X)P(K_1)}{P(X)} = \frac{\mathcal{L}(\theta_2|X)P(K_2)}{P(X)}$$

To calculate $P(X)$ we can marginalize X over all possibilities, in this case two, so:

$$
\begin{aligned}
P(X) &= \sum_{k \in K} P(X|K=k) \\
&= \sum_{k \in K} \mathcal{L}(\theta_k|X)
\end{aligned}
$$

So now we have:

$$
\frac{P(X|K_1)P(K_1)}{\sum_{k \in K} \mathcal{L}(\theta_k|X)} = \frac{P(X|K_2)P(K_2)}{\sum_{k \in K} \mathcal{L}(\theta_k|X)}
$$

But we can simplify it.

$$
P(X|K_1)P(K_1) = P(X|K_2)P(K_2)
$$

We can do the same for $P(K_1)$

$$
\begin{aligned}
P(K_1) &= \int_{x \in X} P(K_1|X)P(dx) \\
&= \int_{x \in X} P(K_1|X)P(dx) \\
&= \int_{x \in X} \mathcal{L}(X|\theta_1)P(dx)
\end{aligned}
$$

Which give us the final formula:

$$
P(X|K_1) \int_{x \in X} \mathcal{L}(X|\theta_1)P(dx) = P(X|K_2) \int_{x \in X} \mathcal{L}(X|\theta_2)P(dx)
$$

The first possible simplification is to estimate $P(K_i)$ and two possible ways: - First, we can say that they both have the same probability; - Second, if we have a dataset we can calculate them by

$$
\begin{aligned}
P(K_1) &= \frac{\sum_{y_i \in Y} \mathbb{1}(y_i = k_1)}{N} \\
P(K_2) &= \frac{\sum_{y_i \in Y} \mathbb{1}(y_i = k_2)}{N}
\end{aligned}
$$

Equal Prior Distributions:
In this case the equality can be simplified to

$$
P(X|K_1) = P(X|K_2)
$$

For futher simplification we can imagine a two dimensional X:

$$P(x_1, x_2|K_1) = P(x_1, x_2|K_2)$$

if we choose another simplification, that x1 and x2 and independent, we arrive at a Naive Bayer Classifier, and the equation become:

$$P(x_1|K_1)P(x_2|K_1) = P(x_1|K_2)P(x_2|K_2)$$

with two classes and Gaussian distribution: Different mean, same variance = line/plane Same mean, different variance = circle/elipse general case = parabolic curve

with more than two cases will be a piecewise combination of the above three cases.

$$(\frac{1}{\sqrt{2\pi\sigma_1^2}}e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}})(\frac{1}{\sqrt{2\pi\sigma_1^2}}e^{-\frac{(x_2-\mu_1)^2}{2\sigma_1^2}}) = (\frac{1}{\sqrt{2\pi\sigma_2^2}}e^{-\frac{(x_1-\mu_2)^2}{2\sigma_2^2}})(\frac{1}{\sqrt{2\pi\sigma_2^2}}e^{-\frac{(x_2-\mu_2)^2}{2\sigma_2^2}})$$

$$(\frac{1}{\sqrt{2\pi\sigma_1^2}})^2(e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}})(e^{-\frac{(x_2-\mu_1)^2}{2\sigma_1^2}}) = (\frac{1}{\sqrt{2\pi\sigma_2^2}})^2(e^{-\frac{(x_1-\mu_2)^2}{2\sigma_2^2}})(e^{-\frac{(x_2-\mu_2)^2}{2\sigma_2^2}})$$

$$A = (\frac{1}{\sqrt{2\pi\sigma_1^2}})^2$$

$$B = (\frac{1}{\sqrt{2\pi\sigma_2^2}})^2$$

$$C = -\frac{1}{2\sigma^2}$$

$$A(e^{\frac{(x_1-\mu_1)^2}{C}})(e^{\frac{(x_2-\mu_1)^2}{C}}) = B(e^{\frac{(x_1-\mu_2)^2}{C}})(e^{\frac{(x_2-\mu_2)^2}{C}})$$

$$A\frac{(e^{\frac{(x_1-\mu_1)^2}{C}})}{(e^{\frac{(x_1-\mu_2)^2}{C}})} = B\frac{(e^{\frac{(x_2-\mu_2)^2}{C}})}{(e^{\frac{(x_2-\mu_1)^2}{C}})}$$

$$Ae^{\frac{(x_1-\mu_1)^2}{C}*\frac{C}{(x_1-\mu_2)^2}} = Be^{\frac{(x_2-\mu_2)^2}{C}*\frac{C}{(x_2-\mu_1)^2}}$$

$$Ae^{\frac{(x_1-\mu_1)^2}{(x_1-\mu_2)^2}} = Be^{\frac{(x_2-\mu_2)^2}{(x_2-\mu_1)^2}}$$

$$ln(Ae^{\frac{(x_1-\mu_1)^2}{(x_1-\mu_2)^2}}) = ln(Be^{\frac{(x_2-\mu_2)^2}{(x_2-\mu_1)^2}})$$

$$ln(A)ln(e^{\frac{(x_1-\mu_1)^2}{(x_1-\mu_2)^2}}) = ln(B)ln(e^{\frac{(x_2-\mu_2)^2}{(x_2-\mu_1)^2}})$$

$$A' = ln(A)$$

$$B' = ln(B)$$

$$A'\frac{(x_1-\mu_1)^2}{(x_1-\mu_2)^2} = B'\frac{(x_2-\mu_2)^2}{(x_2-\mu_1)^2}$$

$$A'(\frac{x_1-\mu_1}{x_1-\mu_2})^2 = B'(\frac{x_2-\mu_2}{x_2-\mu_1})^2$$

$$(\frac{x_1-\mu_1}{x_1-\mu_2})^2 = \frac{B'}{A'}(\frac{x_2-\mu_2}{x_2-\mu_1})^2$$

$$\frac{x_1-\mu_1}{x_1-\mu_2} = \sqrt{\frac{B'}{A'}(\frac{x_2-\mu_2}{x_2-\mu_1})^2}$$

$$\frac{x_1-\mu_1}{x_1-\mu_2} = (\frac{x_2-\mu_2}{x_2-\mu_1})\sqrt{\frac{B'}{A'}}$$

$$C' = \sqrt{\frac{B'}{A'}}$$

$$\frac{x_1-\mu_1}{x_1-\mu_2} = C'(\frac{x_2-\mu_2}{x_2-\mu_1})$$

We can simplify the left side of the equation, because:

$$\begin{aligned}\frac{x-\mu_1}{x-\mu_2} &= \frac{x-\mu_2+\mu_2+\mu_1}{x-\mu_2} \\ &= \frac{x-\mu_2}{x-\mu_2} + \frac{\mu_2+\mu_1}{x-\mu_2} \\ &= 1 + \frac{\mu_2+\mu_1}{x-\mu_2}\end{aligned}$$

and

$$\begin{aligned}\frac{x_2-\mu_2}{x_2-\mu_1} &= \frac{x_2-\mu_1+\mu_1-\mu_2}{x_2-\mu_1} \\ &= \frac{x_2-\mu_1}{x_2-\mu_1} + \frac{\mu_1-\mu_2}{x_2-\mu_1} \\ &= 1 + \frac{\mu_1-\mu_2}{x_2-\mu_1}\end{aligned}$$

$$\begin{aligned}\frac{x_1-\mu_1}{x_1-\mu_2} &= C'(\frac{x_2-\mu_2}{x_2-\mu_1}) \\ 1 + \frac{\mu_2+\mu_1}{x_1-\mu_2} &= C'(1 + \frac{\mu_1-\mu_2}{x_2-\mu_1}) \\ \frac{\mu_2+\mu_1}{x_1-\mu_2} &= C'(1 + \frac{\mu_1-\mu_2}{x_2-\mu_1}) - 1 \\ \frac{1}{x_1-\mu_2} &= \frac{C'(1 + \frac{\mu_1-\mu_2}{x_2-\mu_1}) - 1}{\mu_2+\mu_1} \\ x_1-\mu_2 &= \frac{\mu_2+\mu_1}{C'(1 + \frac{\mu_1-\mu_2}{x_2-\mu_1}) - 1} \\ x_1 &= \frac{\mu_2+\mu_1}{C'(1 + \frac{\mu_1-\mu_2}{x_2-\mu_1}) - 1} + \mu_2 \\ x_1 &= \frac{\mu_2+\mu_1}{\sqrt{\frac{ln(B)}{ln(A)}}(1 + \frac{\mu_1-\mu_2}{x_2-\mu_1}) - 1} + \mu_2 \\ x_1 &= \frac{\mu_2+\mu_1}{\sqrt{\frac{ln((\frac{1}{\sqrt{2\pi\sigma_2^2}})^2)}{ln((\frac{1}{\sqrt{2\pi\sigma_1^2}})^2)}}(1 + \frac{\mu_1-\mu_2}{x_2-\mu_1}) - 1} + \mu_2\end{aligned}$$