MATH 574M: Homework 3


2. Denote \( \hat{y} = (y_1, \ldots, y_K) \) with \( \sum_{k=1}^{K} y_k = 1 \). Denote the center for class \( k \) by \( t_k = (t_{k1}, \ldots, t_{kK}) \) with \( t_{kj} = 0 \) for any \( j \neq k \) and \( t_{kk} = 1 \) for \( k = 1, \ldots, K \). Then

\[
\|t_k - \hat{y}\|^2 = \sum_{j=1}^{K} (t_{kj} - y_j)^2 = \sum_{j \neq k} (y_j - 0)^2 + (y_k - 1)^2
\]

\[
= \sum_{j=1}^{K} y_j^2 + 1 - 2y_k, \quad k = 1, \ldots, K,
\]

where the term \( \sum_{j=1}^{K} y_j^2 + 1 \) in the above is common for all \( k \). Therefore

\[
\arg \min_{k=1,\ldots,K} \|t_j - \hat{y}\|^2 = \arg \max_{k=1,\ldots,K} y_k.
\]

3. (a) Denote the distance from the origin to an arbitrary point in the unit ball by \( R \).

Assume the distance of the \( i \)-th point to the origin is \( R_i \), for \( i = 1, \ldots, N \). Let \( R_{(1)} \) be the closest distance among \( N \) data points.

Denote the cdf of \( R_{(1)} \) by \( F_{R_{(1)}} \). Then we have

\[
1 - F_{R_{(1)}}(r) = \prod_{i=1}^{N} P(R_i \geq r) = \left[ \frac{V_p(1) - V_p(r)}{V_p(1)} \right]^N = (1 - r^p)^N,
\]

where \( V_p(r) \) is the volume of an \( p \)-dimensional ball with radius \( r \), which is equal to

\[
V_p(r) = \frac{\pi^{p/2}r^p}{\Gamma(p/2 + 1)}.
\]

The median, which is equal to \( F_{R_{(1)}}^{-1}(\frac{1}{2}) \), is computed as

\[
F_{R_{(1)}}(m) = 1 - (1 - m^p)^N = \frac{1}{2}, \quad \Rightarrow m = \left( 1 - \frac{1}{2} \right)^{1/p}.
\]

(b) For \( n = 500 \) and \( p = 10 \), \( m \approx 0.52 \), and for \( n = 500, p = 100 \), we have \( m \approx 0.94 \).

This implies that when the data dimension is high, all the data points are far away from the origin and close to the boundary of the ball. For example, when \( n = 500 \) and \( p = 10 \), the closest point is more than halfway to the boundary, and therefore most data points are closer to the boundary of the sample space than to any other data point. When \( n = 500 \) and \( p = 10 \), the point which is the closest to the origin among the samples is actually far away from the origin with the median distance 0.936, and the rest of points are even farther away.

4. For Questions 4-6, please see the code.