Bayes misclassification for many-component GMM
Charles Zheng, STAT 312

Problem

Let $z_1, \ldots z_n \sim N(0, I_d)$ iid, and let $x \sim N(Z_i, I)$ with $i \sim Cat(n)$.

The problem is to predict the unobserved label i given the location x. Assume for now that z_i are known, so that we are computing the Bayes misclassification risk.

That is, we want to know

$$\mathbf{E}[\mathbf{E}[\min_{\hat{i}} \Pr[\hat{i}(x) \neq i]|z_1, \ldots, z_n]]$$

where $\hat{i}$ is the classifier $\hat{i}(x) : \mathbb{R}^d \to \{1, \ldots, n\}$, and the outer expectation is taken over the joint distribution of z_i.

Facts

We know that the Bayes classifier $\hat{i}$ takes the form

$$\hat{i}(x) = \text{argmin}_i ||x - z_i||^2$$

and henceforth $\hat{i}$ refers to the Bayes classifier.

By exchangeability,

$$\Pr[\hat{i}(x) \neq i] = \Pr[\hat{i}(x) \neq 1|i = 1]$$

Now condition on the distribution of z_1 and define $\eta = ||z_1 - x||$

$$\begin{aligned}\Pr[\hat{i}(x) \neq 1|i = 1] &= \int \phi(z)\phi(x - z) \Pr[\hat{i}(x) \neq 1|i = 1, z_1 = z]dzdx \\ &= \int \phi(z)\phi(x - z)p(z, x)dzdx\end{aligned}$$

where

$$\begin{aligned}p(z, x) &= \Pr[\hat{i}(x) \neq 1|i = 1, z_1 = z, x] \\ &= \Pr[\text{argmin}_i ||x - z_i||^2 \neq 1] \\ &= \Pr[||x - z_1||^2 < \min_{i>1} ||x - z_i||^2] \\ &= \Pr[z_i \in B_\eta(x) \text{ for } i > 1] \\ &= 1 - \Pr[z_i \notin B_\eta(x) \text{ for all } i > 1] \\ &= 1 - \Pr[z_2 \notin B_\eta(x)]^{n-1} \text{ (due to independence of } z_i,\ i > 1) \\ &= 1 - \left(1 - \int_{B_\eta(x)} \phi(z)dz\right)^{n-1}\end{aligned}$$

Asymptotics

We derive an asymptotic approximation of $p(z,x)$ for small η and large n. Let $C_d\eta^d$ denote the volume of a spherical ball ηD^d.

$$
\begin{aligned}
p(z,x) &\approx 1 - \left(1 - \int_{B_\eta(x)} \phi(z)dt\right)^{n-1} \\
&= 1 - (1 - \phi(z)C_d\eta^d)^{n-1} \\
&\approx 1 - e^{-(n-1)\phi(z)C_d\eta^d} \\
&\approx 1 - e^{-n\phi(z)C_d\eta^d}
\end{aligned}
$$

Seeing that in the small η limit, $p(z,x)$ only depends on z and η, let $q(z,\eta) = 1 - e^{-n\phi(z)C_d\eta^d}$, so that

$$
\begin{aligned}
\Pr[\hat{i}(x) \neq i] &\approx \int \phi(z)\chi_d(\eta)q(z,\eta)dzd\eta \\
&= \int \phi(z)dz \left[\int \chi_d(\eta)q(z,\eta)d\eta\right] \\
&= \int \phi(z)Q(z)dz
\end{aligned}
$$

where $\chi_d(\eta)$ is the density function for η (a chi-distribution with d degrees of freedom) and

$$Q(z) = \int \chi_d(\eta)q(z,\eta)d\eta$$

Recall that

$$\chi_d(\eta) = 2\Gamma(d/2)^{-1}2^{-d/2}\eta^{d-1}\exp[-\eta^2/2] = G_d\eta^{d-1}\exp[-\eta^2/2]$$

Thus

$$Q(z) = 1 - G_d \int \eta^{d-1}\exp[-\eta^2/2 - n\phi(z)C_d\eta^d]d\eta$$