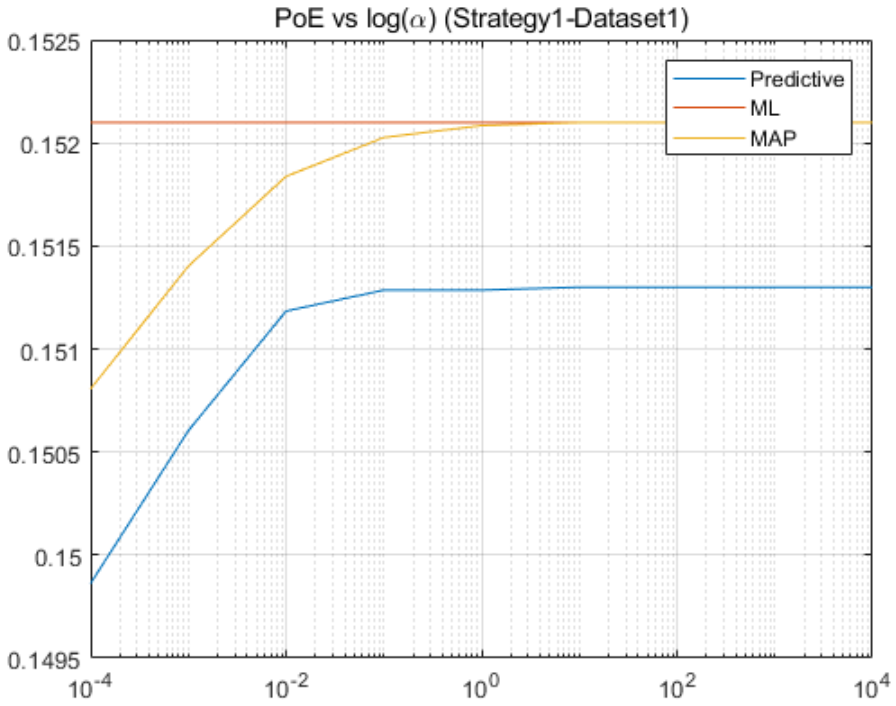
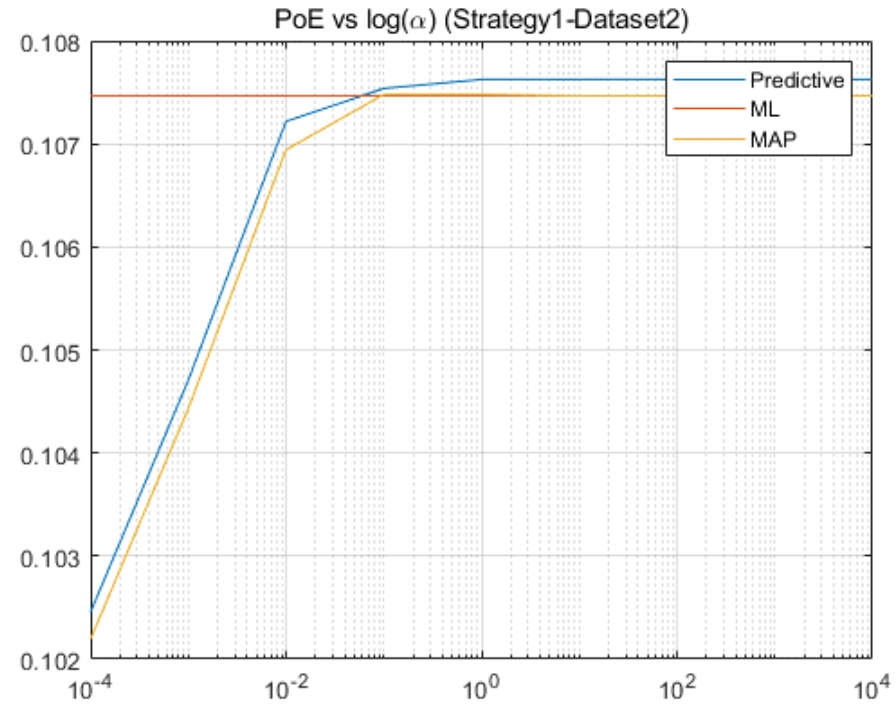
[ECE271A] Homework 3 & 4 Solution

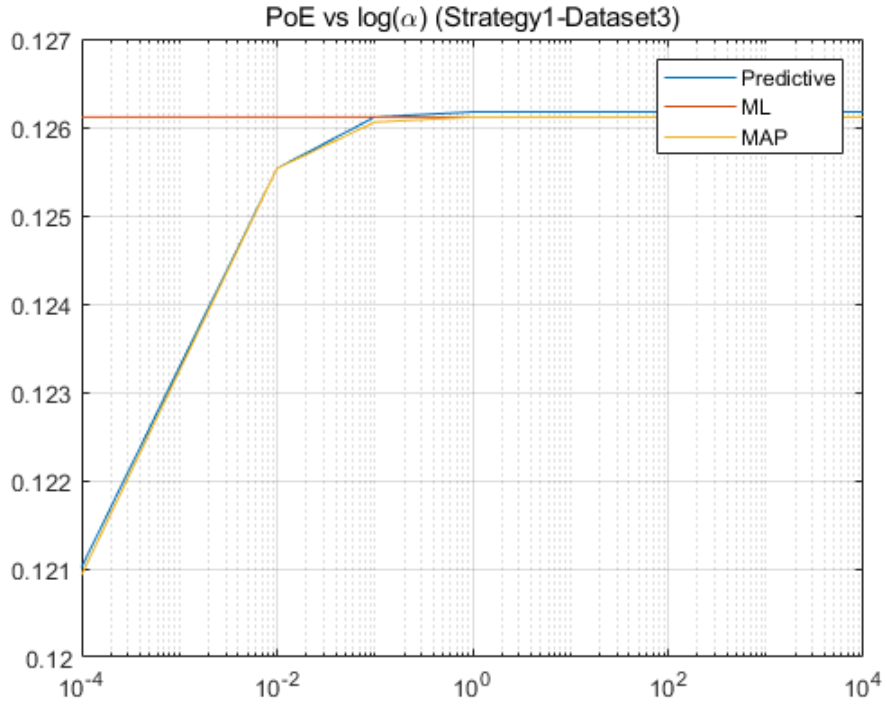
Name: Yang Yue PID: A53301503

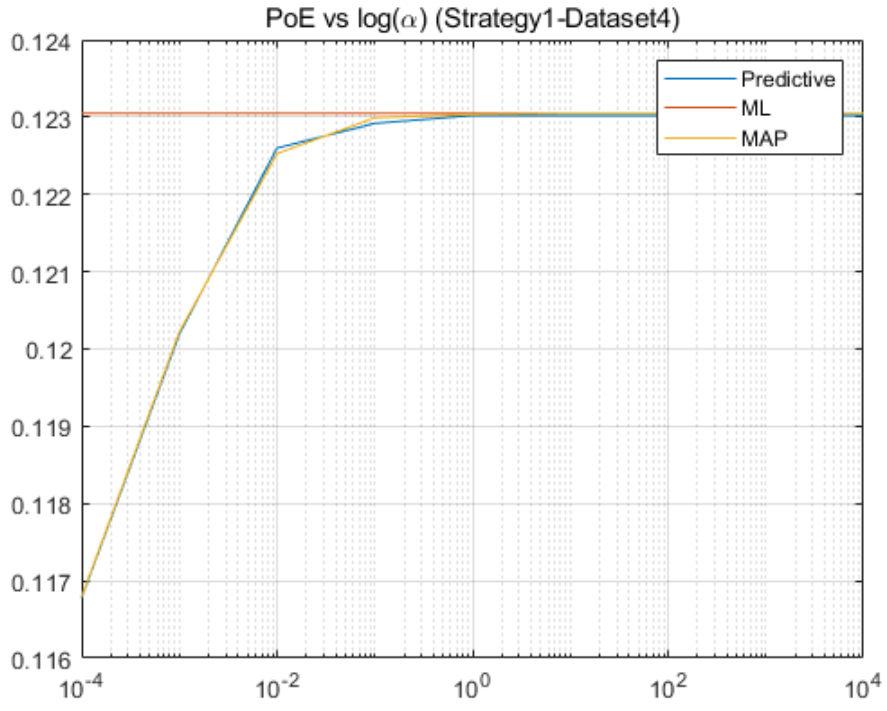
Problem 4 (computer)

PoE (Probability of Error) vs different alphas with respect to different Datasets using Strategy 1

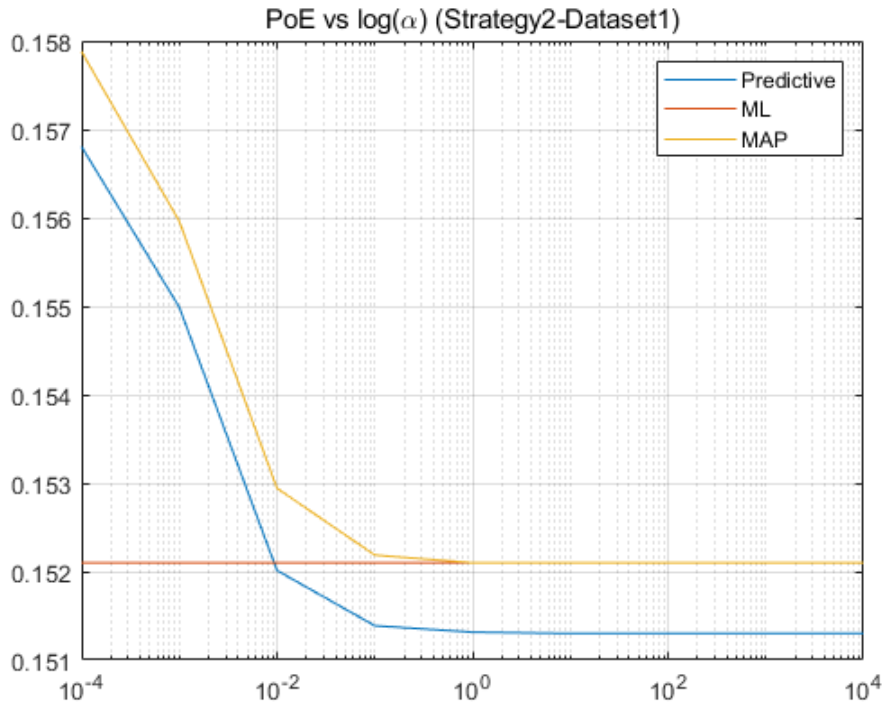


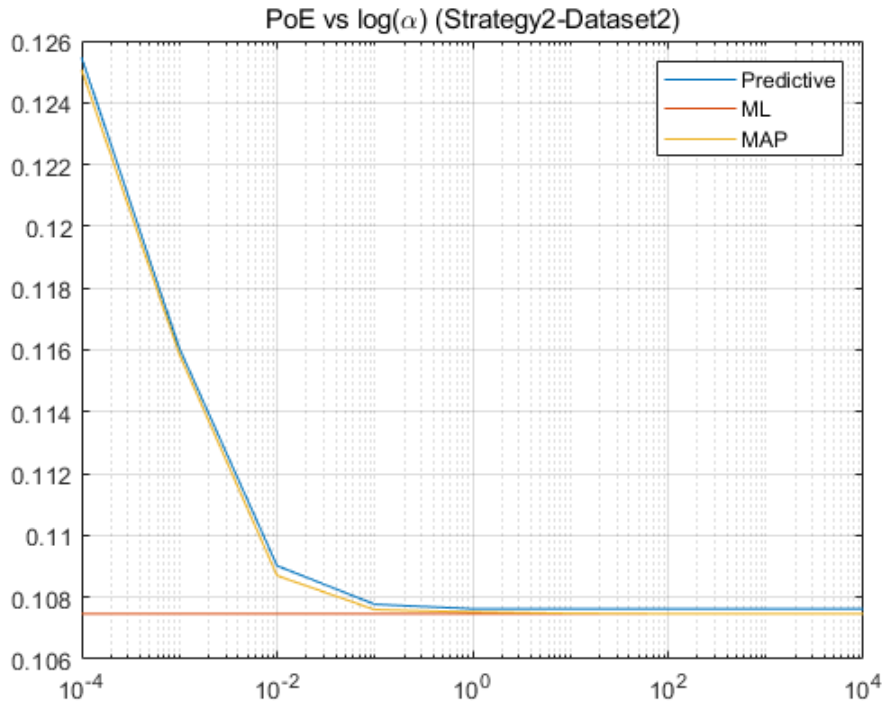


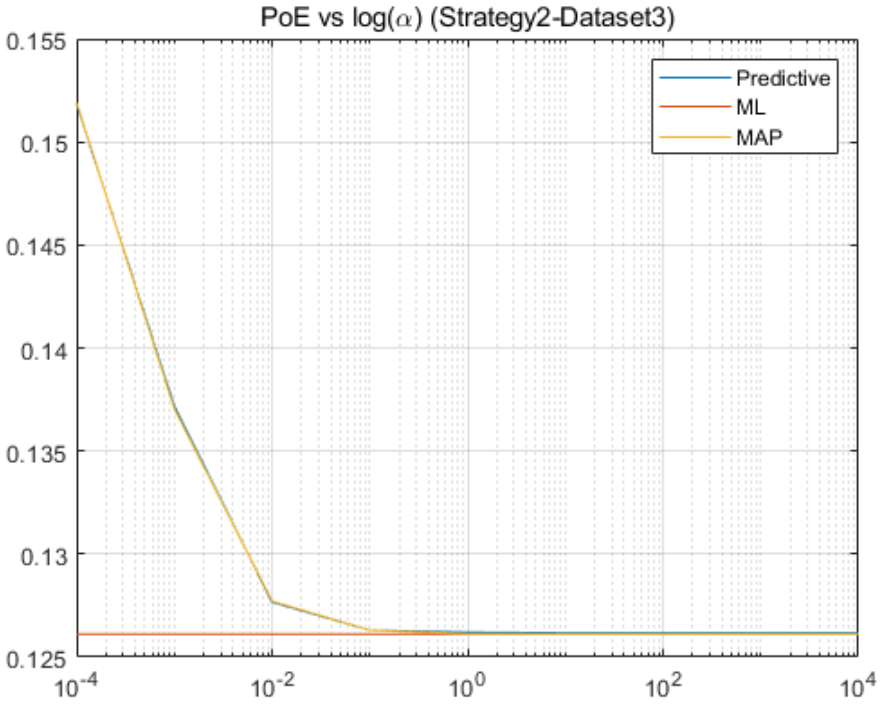

PoE vs log(α) (Strategy1-Dataset3)
Predictive
ML
MAP
0.127
0.126
0.125
0.124
0.123
0.122
0.121
0.12
10^-4
10^-2
10^0
10^2
10^4

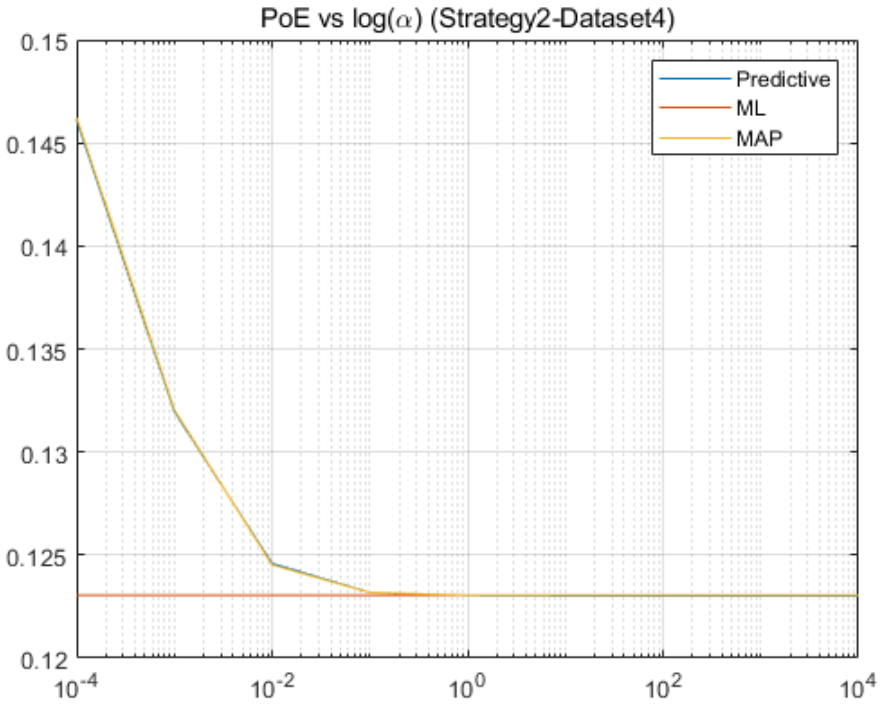

PoE vs log(α) (Strategy1-Dataset4)
Predictive
ML
MAP
0.124
0.123
0.122
0.121
0.12
0.119
0.118
0.117
0.116
10^-4
10^-2
10^0
10^2
10^4

PoE (Probability of Error) vs different alphas with respect to different Datasets using Strategy 2






PoE vs log(α) (Strategy2-Dataset3)
Predictive
ML
MAP
0.155
0.15
0.145
0.14
0.135
0.13
0.125
10^-4
10^-2
10^0
10^2
10^4


PoE vs log(α) (Strategy2-Dataset4)
Predictive
ML
MAP
0.15
0.145
0.14
0.135
0.13
0.125
0.12
10^-4
10^-2
10^0
10^2
10^4

Explanation

Given that

$$P_\mu(\mu) = \mathcal{G}(\mu, \mu_0, \Sigma_0),$$

where μ_0 is given, $\Sigma_0^2 = diag(\alpha_i \cdot \omega_0)$ is given as well,

And

$$P_{\mu|T}(\mu|D) = \mathcal{G}(\mu, \mu_n, \Sigma_n),$$
$$P_{X|\mu,\Sigma} = \mathcal{G}(x, \mu, \Sigma)$$

Also, given that,

$$P_{X|T}(x|D) = \mathcal{G}(x, \mu_n, \Sigma + \Sigma_n)$$

In hw3, we assume that

$$\Sigma = cov(X), \text{ and } \Sigma_n = n\Sigma_0(\Sigma + n\Sigma_0)^{-1}.$$

For a given n

$$\mu_n = \underbrace{n\Sigma_0(\Sigma + n\Sigma_0)^{-1}}_{\alpha_n}\mu_{ML} + \underbrace{\Sigma(\Sigma + n\Sigma_0)}_{1-\alpha_n}\mu_0$$

Which can be written as,

$$\mu_n = \alpha_n\,\hat{\mu} + (1 - \alpha_0)\mu_0 \text{ where } \alpha_n \in [0,1]$$

In MAP case,

$$P_{X|T}(x|D) = \mathcal{G}(x, \mu_{MAP}, \Sigma_n),$$

where

$$\mu_{MAP} = n\Sigma_0(\Sigma + n\Sigma_0)^{-1}\mu_{ML} + \Sigma(\Sigma + n\Sigma_0)\mu_0,$$

In ML case,

$$P_{X|T}(x|D) = \mathcal{G}(x, \mu_{ML}, \Sigma_n),$$

1) In hw3, $\sigma_0^2 = \alpha \times \omega_0$ represents how confident we are about the prior probability of μ, for a hypothetical expectation of μ, a larger σ_0^2 shows less confidence.
   Assuming that we have no confidence about the distribution of μ is correct, we can show this by giving the distribution a very large covariance of distribution Σ_0. This indicates that $\alpha_n \to 1$, and $\mu_n \to \hat{\mu} = \mu_{ML}$, and because of that, the difference between of each curve using predictive equation, MAP, and MLE is tend to be very small when $\alpha \to \infty$.
   Inversely, assuming that we are very certain about the distribution of μ, we will make Σ_0 be very small. Then μ_n is determined by i) training dataset and ii) μ_0.
   As each plot of Strategy1 has shown, PoE increases with an increasing α, which indicates that the assumption we make about the distribution of μ is reasonable. Due to the reasonable assumption, intuitively, taking it into account can reduce the PoE. What the plots show is that a reasonable prior distribution of μ help reduce the error rate, the more certainty we have about the correct assumption, the more "helpful" the assumption will be. It is obvious especially when the size of training set is small. As a result of this, PoE(predictive) is obviously smaller than PoE(MAP) and PoE(MLE) in Strategy1 with Dataset1.

2) When the size of training dataset is sufficiently large enough, let's suppose $n \to \infty$, then,

$$\Sigma_{Predictive} = \Sigma + \Sigma\Sigma_0(\Sigma + n\Sigma_0)^{-1} = \Sigma + \frac{1}{n}\Sigma\Sigma_0\left(\frac{1}{n}\Sigma + \Sigma_0\right)^{-1} \to \Sigma = \Sigma_{MAP},$$

   which implies that when α remains, the larger the size of training set is, the closer the PoE(predictive) and PoE(MAP) get.

Also, while $n \to \infty$,

$$\mu_{MAP} = n\Sigma_0(n\Sigma_0^2 + \Sigma^2)^{-1}\mu_{ML} + \Sigma^2(n\Sigma_0^2 + \Sigma^2)\mu_0$$
$$= \Sigma_0\left(\Sigma_0^2 + \frac{1}{n}\Sigma^2\right)^{-1}\mu_{ML} + \frac{1}{n}\Sigma^2\left(\Sigma_0^2 + \frac{1}{\mathrm{n}}\Sigma^2\right)\mu_0 \to \mu_{ML}$$

Which indicates that as n gets larger, the plot of MAP solution and the plot of ML solution tend to coincide.

3) When the strategy changes from to Strategy2, it is suggesting that the distribution of $P_\mu(\mu)$ of front ground and back ground are the same, thus unreasonable. Because we cannot distinguish μ_{BG} and μ_{FG}, which, with the intuition that there should be some differences of the distribution of parameter from 2 different things' images. It can be shown from the plots that as α gets larger, the weight of prior of μ gets less, the error rate decreases, which intuitively suggests that "the data is trying to fix the mistake made by wrong prior estimation".

   Other relative behaviors remain even though the estimated distribution of prior is wrong. For example, i) the curves are still tending to be together with an increasing α; ii) as n grows, the estimation using Bayesian estimate combines the prior beliefs is more alike to the estimation using MAP rule.