Homework 1

EEE 4774 & 6777 Data Analytics

1 Gaussian parameter estimation

Assume independent and identically distributed (iid) samples $\mathcal{D} = \{x_1, \ldots, x_N\}$ from a Gaussian distribution, i.e., $x_n \sim \mathcal{N}(\mu, \sigma^2), n = 1, \ldots, N$.

a) Derive the maximum likelihood (ML) estimates of μ and σ^2. [15 pts]

b) Show if the ML estimate $\hat{\mu}_{ML}$ is biased/unbiased. [10 pts]

c) Assuming a prior $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$ find the maximum a posteriori (MAP) estimate of μ in terms of the ML estimate $\hat{\mu}_{ML}$.
(Hint: $\hat{\mu}_{MAP} = a\hat{\mu}_{ML} + b\mu_0$, find a and b.) [15 pts]

2 Experiment

Download "`baseball_data_2005.csv`". This dataset contains statistics about a number of baseball players from the 2005 season. "Season AB" denotes the total number of batting attempts, say N, by the player in the season. The last 6 columns gives the number of successful hits at different bases in these attempts, so the sum of the last 6 columns "HB(4)+...+HB(9-10)" gives the total number of hits, N_H. The probability of hit θ for a player is a principal performance measure, which can be simply estimated using the "batting average=N_H/N" where N_H is the number of hits and N is the number of attempts. However, small number of attempts does not typically provide reliable estimation, e.g., the probability for Bronson Arroyo would yield $\hat{\theta} = 0/1 = 0$. This is nothing but the overfitting problem.

a) By step by step derivation show that batting average, N_H/N, is the maximum likelihood (ML) estimate $\hat{\theta}_{ML}$ for the probability parameter θ in the Binomial model $Binom(N_H, N, \theta)$. [15 pts]

b) Using the $Beta(100, 300)$ distribution as the prior distribution for the Binomial probability parameter (i.e., $p(\theta) = Beta(100, 300)$) compute the *maximum a posteriori (MAP)* estimates for the successful hit probability of all players. Comment on the comparison MAP estimates vs. ML estimates (see part a). [30 pts]

3 Principal Component Analysis (PCA)

For D-dimensional data points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, write the PCA procedure for M principal components step by step. [15 pts]