CSCI 552I : Machine Learning Fundamentals. Homework 2.

[1]

- $x^i \in \mathbb{R}^d$, $X = \{x^1, x^2 \cdots x^N\}$, N iid samples.
- Multivariate Gaussian distribution in $\mathbb{R}^d$
- $\mu \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d \times d}$,

- $$P(x|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \right]$$

(a) The maximum likelihood estimates for the μ, Σ, based on X.

The log likelihood for this iid samples is

$$L(\mu, \Sigma | X) = \log P(X|\mu, \Sigma) = \log P(x_1, x_2, \cdots x_N | \mu, \Sigma)$$

$$= \log \prod_{t=1}^{N} P(x^t | \mu, \Sigma) = \sum_{t=1}^{N} \log P(x^t | \mu, \Sigma)$$

$$= \sum_{t=1}^{N} \left[\underbrace{-\frac{d}{2} \log 2\pi}_{\text{constant.}} \underbrace{- \frac{1}{2} \log |\Sigma|}_{\text{Scalar variable.}} \underbrace{- \frac{1}{2} (x^t - \mu)^T \Sigma^{-1} (x^t - \mu)}_{\text{Scalar variable.}} \right]$$

To get the maximum likelihood, the two conditions must be satisfied.

① $\frac{\partial}{\partial \mu} L(\mu, \Sigma | X) = 0$ $\mu \in \mathbb{R}^d$.

② $\frac{\partial}{\partial \Sigma} L(\mu, \Sigma | X) = 0$ $\Sigma \in \mathbb{R}^{d \times d}$.

① $\frac{\partial}{\partial \mu} L(\mu, \Sigma | X) = 0.$ $\quad \mu \in \mathbb{R}^d.$

$$-\frac{1}{2}\sum_{t=1}^{N}\left[\frac{\partial \log|\Sigma|}{\partial \mu} + \frac{\partial}{\partial \mu}(x^t-\mu)^T\Sigma^{-1}(x^t-\mu)\right] = 0.$$

$$\underbrace{N \cdot \frac{\partial}{\partial \mu}\log|\Sigma|}_{i)} + \underbrace{\sum_{t=1}^{N}\frac{\partial}{\partial \mu}(x^t-\mu)^T\Sigma^{-1}(x^t-\mu)}_{ii)}$$

$\mu \in \mathbb{R}^d \quad \Sigma \in \mathbb{R}^{d \times d}.$

i) $\frac{\partial}{\partial \mu}\log|\Sigma| = 0.$

ii) From the matrix cookbook equation # (78),

$$\frac{\partial (B\mu + x^t)^T\Sigma^{-1}(D\mu + x^t)}{\partial \mu} = B^T\Sigma^{-1}(D\mu + x^t) + D^T\Sigma^{-1T}(B\mu + x^t)$$

where $B = D = -I.$

$$\frac{\partial}{\partial \mu}(x^t-\mu)^T\Sigma^{-1}(x^t-\mu) = -\Sigma^{-1}(x^t-\mu) - (\Sigma^{-1})^T(x^t-\mu)$$

Since Σ is a covariance matrix, which is a symmetric matrix,

Σ^{-1} is also symmetric, which results in $\Sigma^{-1} = (\Sigma^{-1})^T.$

Thus, $\quad -\Sigma^{-1}(x^t-\mu) - (\Sigma^{-1})^T(x^t-\mu) = -2\Sigma^{-1}(x^t-\mu)$

From i), ii), $\quad -\frac{1}{2}\sum_{t=1}^{N}\left\{-2\Sigma^{-1}(x^t-\mu)\right\} = 0.$

$$\Sigma^{-1}\underbrace{\left\{\sum_{t=1}^{N}(x^t-\mu)\right\}}_{\shortparallel\ 0.} = 0.$$

$$\sum_{t=1}^{N} x^t - \sum_{t=1}^{N} \mu = \sum_{t=1}^{N} x^t - N\mu = 0.$$

$$\hat{\mu} = \frac{1}{N} \sum_{t=1}^{N} x^t .$$

② $\frac{\partial}{\partial \Sigma} L(\mu, \Sigma | X) = 0.$

$$-\frac{1}{2} \sum_{t=1}^{N} \Big[\underbrace{\frac{\partial}{\partial \Sigma} \log|\Sigma|}_{i)} + \underbrace{\frac{\partial}{\partial \Sigma} \{ (x^t - \mu)^T \Sigma^{-1} (x^t - \mu) \}}_{ii)} \Big] = 0.$$

i) From the matrix cookbook #(57),

$$\frac{\partial \log|\Sigma|}{\partial \Sigma} = (\Sigma^{-1})^T = \Sigma^{-1} \quad \text{(symmetric)}$$

ii) From the matrix cookbook #(61),

$$\frac{\partial}{\partial X} \{ a^T X^{-1} b \} = -X^{-T} a b^T X^{-T}$$

$$\frac{\partial}{\partial \Sigma} \{ (x^t - \mu)^T \Sigma^{-1} (x^t - \mu) \} = -\Sigma^{-T} (x^t - \mu)(x^t - \mu)^T \Sigma^{-T}$$

$$= -\Sigma^{-1} (x^t - \mu)(x^t - \mu)^T \Sigma^{-1} \quad (\text{since } \Sigma^{-1} \text{ is symmetric})$$

Thus, $-\frac{1}{2} \sum_{t=1}^{N} \{ \Sigma^{-1} - \Sigma^{-1} (x^t - \mu)(x^t - \mu)^T \Sigma^{-1} \} = 0$

$$\sum_{t=1}^{N} (\Sigma^{-1}) - \sum_{t=1}^{N} \left\{ \Sigma^{-1} (x^t - \mu)(x^t - \mu)^T \Sigma^{-1} \right\} = 0.$$

$$N \cdot \Sigma^{-1} - (\Sigma^{-1}) \sum_{t=1}^{N} \left\{ (x^t - \mu)(x^t - \mu)^T \Sigma^{-1} \right\} = 0$$

$$= \Sigma^{-1} \Big[\underbrace{N - \sum_{t=1}^{N} \left\{ (x^t - \mu)(x^t - \mu)^T \Sigma^{-1} \right\}}_{= 0} \Big] = 0$$

$$N - \sum_{t=1}^{N} \left\{ (x^t - \mu)(x^t - \mu)^T \Sigma^{-1} \right\} = 0.$$

$$\sum_{t=1}^{N} \left\{ (x^t - \mu)(x^t - \mu)^T \Sigma^{-1} \right\} = N.$$

$$\sum_{t=1}^{N} (x^t - \mu)(x^t - \mu)^T = N \cdot \Sigma$$

$$\therefore \hat{\Sigma} = \frac{1}{N} \sum_{t=1}^{N} (x^t - \mu)(x^t - \mu)^T.$$

II (b) $\cdot\ \hat{\mu}$: maximum likelihood estimate of the mean.

$\cdot\ \hat{\mu}$ is biased estimate of the true mean μ?

$$E[\hat{\mu}] = E\left[\frac{1}{N}\sum_{t=1}^{N} x^t\right] = \frac{1}{N} E[x^1 + x^2 + \cdots + x^N]$$

$$= \frac{1}{N}(\mu + \mu + \cdots \mu) = \frac{1}{N} \cdot N\mu = \mu.$$

Since the estimate is same as the true mean which means $\hat{\mu}$ is unbiased

II (c) $\hat{\Sigma}$ is unbiased?

$$E[\hat{\Sigma}] = E\left[\frac{1}{N}\sum_{t=1}^{N}(x^t - \hat{\mu})(x^t - \hat{\mu})^T\right]$$

$$= \frac{1}{N}\sum_{t=1}^{N} E[(x^t - \hat{\mu})(x^t - \hat{\mu})^T]$$

$$= \frac{1}{N}\sum_{t=1}^{N} E[(x^t - \hat{\mu})(x^{tT} - \hat{\mu}^T)]$$

$$= \frac{1}{N}\sum_{t=1}^{N} E[x^t x^{tT} + \hat{\mu}\hat{\mu}^T - \hat{\mu}x^{tT} - x^t\hat{\mu}^T]$$

$$= \frac{1}{N}\sum_{t=1}^{N} E[x^t x^{tT}] + \frac{1}{N}\sum_{t=1}^{N} E[\hat{\mu}\hat{\mu}^T] - E\Big[\hat{\mu}\underbrace{\Big(\frac{1}{N}\sum_{t=1}^{N} x^{tT}\Big)}_{= \hat{\mu}^T}\Big] - E\Big[\underbrace{\Big(\frac{1}{N}\sum_{t=1}^{N} x^t\Big)}_{= \hat{\mu}} \cdot \hat{\mu}^T\Big]$$

$$= \frac{1}{N}\sum_{t=1}^{N} E[x^t \cdot x^{tT}] + \frac{1}{N}\sum_{t=1}^{N} E[\hat{\mu}\hat{\mu}^T] - 2E[\hat{\mu}\hat{\mu}^T]$$

$$= \frac{1}{N}\sum_{t=1}^{N}\left\{E[x^t x^{tT}] - E[\hat{\mu}\hat{\mu}^T]\right\}$$

Variance mean

From the matrix cookbook #(321), $^* E(xx^T) = M + mm^T$

$$= \frac{1}{N}\sum_{t=1}^{N}\left[\Sigma + E[x^t]E[x^t]^T - \left\{ Var(\hat{\mu}) + E[\hat{\mu}]E[\hat{\mu}]^T \right\} \right]$$

we already know $E[x^t] = E[\hat{\mu}] = \mu$.

$$= \frac{1}{N}\sum_{t=1}^{N}\left[\Sigma - Var(\hat{\mu}) \right]$$

$$* \ Var(\hat{\mu}) = Var\left(\frac{1}{N}\sum_{t=1}^{N} x^t\right) = \frac{1}{N^2} Var\left(\sum_{t=1}^{N} x^t\right)$$

$$= \frac{1}{N^2}\sum_{t=1}^{N} Var(x^t) = \frac{1}{N^2}\cdot N \cdot \Sigma = \frac{1}{N}\Sigma.$$

$$\therefore \ \frac{1}{N}\sum_{t=1}^{N}\left[\Sigma - \frac{1}{N}\Sigma \right] = \frac{1}{N}\cdot N \cdot \Sigma \left(1 - \frac{1}{N}\right)$$

$$= \underline{\frac{N-1}{N}\Sigma} \quad \neq \Sigma \ (\text{true variance matrix})$$

→ biased estimate of the true covariance.

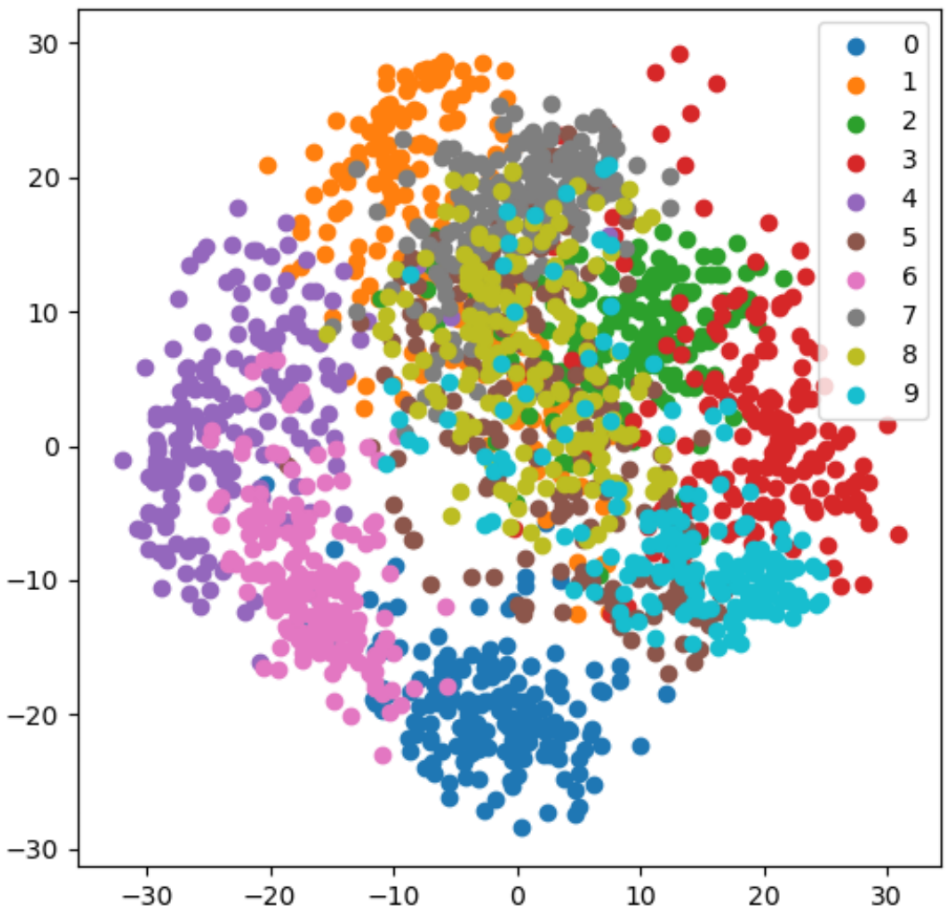
2 Table of MGC and Logistic regression

==== Error rates for MultiGaussClassify Full Cov with Boston50====						
F1	F2	F3	F4	F5	Mean	SD
0.3235	0.297	0.099	0.3465	0.1485	0.2429	0.0998

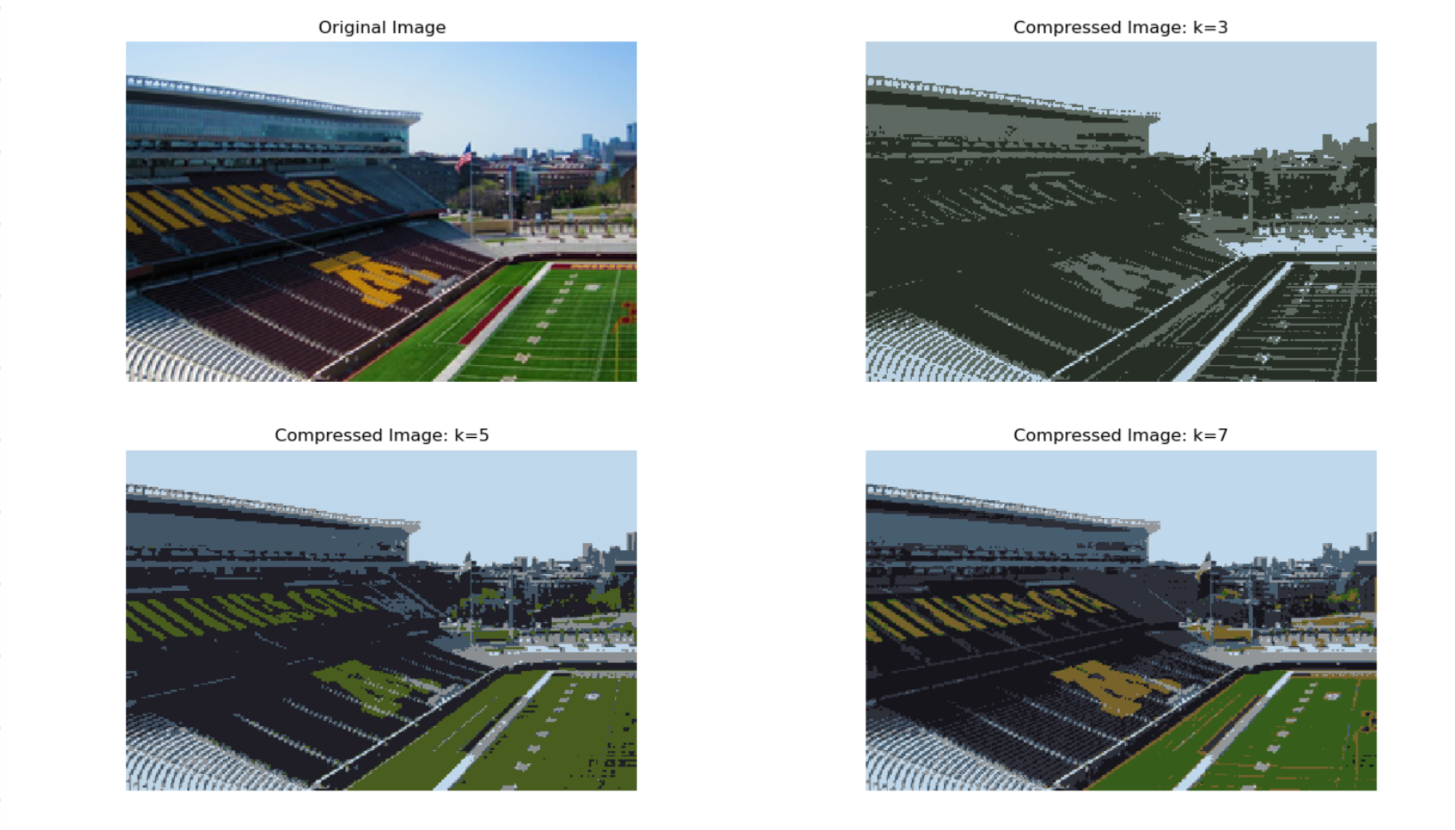
==== Error rates for MultiGaussClassify Diag Cov with Boston50====						
F1	F2	F3	F4	F5	Mean	SD
0.3431	0.1782	0.1287	0.3762	0.1386	0.233	0.1053

==== Error rates for LogisticRegression with Boston50====						
F1	F2	F3	F4	F5	Mean	SD
0.1569	0.1287	0.099	0.2772	0.1881	0.17	0.0612

3 plot of the projected feature



4 Compressed image.



Reconstruction error.