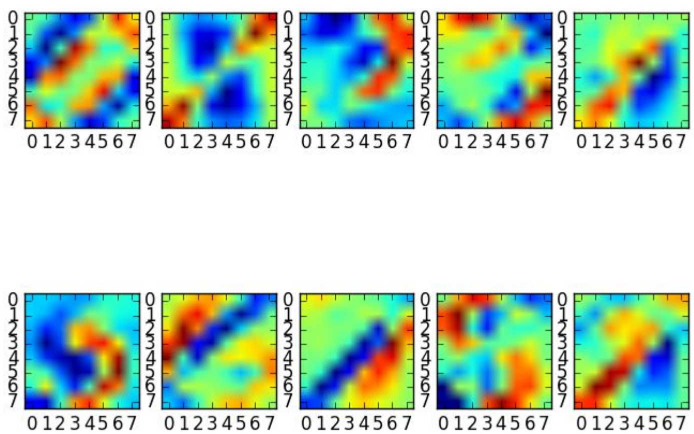
Question 1

a) It can be shown that the gaussian parameters μ_k, Σ_k that maximize the likelihood function for a *given class* k is the mean and sample variance of the observed points of the class. As per the specifications of the question statement, we assume a prior for each class that is uniform. Ordinarily, we might instead assume a multinomial distribution and take the prior as the number of occurrences of a class divided by total number of occurrences of all classes.

Average log-likelihood for training set:
-0.124624436669
Average log likelihood for test:
-0.196673203255

b) Accuracy on training set: 0.981429
Accuracy on test set: 0.972750

c)



Question 2

2 a) $P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \sim P(D|\theta)P(\theta)$

As given, $P(\theta) := \theta_1^{a_1-1} \cdots \theta_K^{a_K-1} = \prod_{i=1}^{K} \theta_i^{a_i-1}$

$P(x_1, \dots, x_N|\theta) = P(x^1|\theta) \cdots P(x^N|\theta)$ assuming independence

$$= \left(\prod_{k=1}^{K} \theta_k^{x_k^1}\right) \cdots \left(\prod_{k=1}^{K} \theta_k^{x_k^N}\right)$$

$$= \prod_{n=1}^{N} \prod_{k=1}^{K} \theta_k^{x_k^n} = \theta_1^{N_1} \cdots \theta_K^{N_K}$$

$$\boxed{\therefore P(\theta|D) = \left(\prod_{n=1}^{N}\prod_{k=1}^{K} \theta_k^{x_k^n}\right)\left(\prod_{i=1}^{K} \theta_i^{a_i-1}\right) = P(\theta_1, \dots, \theta_K|D)}$$

$P(D'|D) = \int P(\theta|D)P(D'|\theta)\,d\theta$

$P(D'|\theta) = P(X_b^{N+1} = 1|\theta) = \theta_b$

$\rightarrow \theta_i \in (0,1)$. But $\sum_{i=1}^{K} \theta_i = 1$. How to do $\int d\theta$?

$\therefore P(D'|D) = \int P(\theta|D)\theta_b\,d\theta =$

$$\int \cdots \int \left(\prod_{n=1}^{N}\prod_{k=1}^{K} \theta_k^{x_k^n}\right)\left(\prod_{i=1}^{K} \theta_i^{a_i-1}\right) \theta_b \, d\theta_1 \cdots d\theta_K$$

$$= \int \cdots \int \theta_1^{N_1} \cdots \theta_K^{N_K} \cdot \theta_1^{a_1-1} \cdots \theta_K^{a_K-1} \cdot \theta_b \, d\theta_1 \cdots d\theta_K$$

$$= \int \cdots \int \underbrace{\theta_1^{N_1+a_1-1} \cdots \theta_K^{N_K+a_K-1}}_{\text{Dirichlet} \sim (N_1+a_1, \dots, N_K+a_K)} \theta_b \, d\theta_1 \cdots d\theta_K$$

$E[\theta_b]$ given $(\theta_1, \dots, \theta_K) \sim \text{Dirichlet}(N_1+a_1, \dots, N_K+a_K)$

$$\boxed{P(X_b^{N+1}=1|\theta) = P(D'|D) = E[\theta_b] = \frac{N_b + a_b}{\sum_{i=1}^{K}(N_i + a_i)} \quad \text{where } N_i = \sum_{n=1}^{N} x_i^{(n)}}$$

2b) $P(\theta|D) \sim P(D|\theta)\,P(\theta) \sim \log P(D|\theta) + \log P(\theta)$

Recall: $P(D|\theta) = \theta_1^{N_1} \cdots \theta_K^{N_K}$

$P(\theta) = \theta_1^{a_1-1} \cdots \theta_K^{a_K-1}$

$\therefore P(\theta|D) \sim N_1 \log\theta_1 + \ldots + N_K \log\theta_K + (a_1-1)\log\theta_1 + \ldots + (a_K-1)\log\theta_K$

$(N_1+a_1-1)\log\theta_1 + \ldots + (N_K+a_K-1)\log\theta_K$

$\cancel{\dfrac{dP(\theta|D)}{d\theta_i} = \dfrac{N_i+a_i-1}{\theta_i} = 0}$ X. Need lagrange again.

Maximize $P(\theta|D)$ $:= f$

subject to $1-(\theta_1+\ldots+\theta_K)=0$ $:= g$

Lagrangian Method: $\nabla_\theta f = \lambda \nabla_\theta g$ — K equations

$g = 0$ — 1 equation

$\theta_1, \ldots, \theta_K, \lambda \rightarrow K+1$ unknowns

$\nabla_{\theta_i} f = \lambda \nabla_\theta g \Rightarrow \dfrac{N_i+a_i-1}{\theta_i} = \lambda(-1)$

$\theta_i = \dfrac{1-a_i-N_i}{\lambda}$

$\therefore 1 - \sum_{i=1}^{K} \theta_i = 0$

$1 - \sum_{i=1}^{K} \dfrac{1-a_i-N_i}{\lambda} = 0$

$1 = \frac{1}{\lambda}\left(\Sigma 1 - \Sigma a_i - \Sigma N_i\right)$

$\lambda = K - \Sigma a_i - N$

$$\therefore \boxed{\theta_i = \frac{1-a_i-N_i}{K-(\Sigma a_i)-N}} = \boxed{\frac{N_i+a_i-1}{N+\Sigma a_i - K}}$$

Question 3

scalar

3. a) $z \sim \mathcal{N}(0,1)$, $x|z \sim \mathcal{N}(z\mu, \Sigma)$, $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_D^2)$

Find $P(z|x;\theta)$

From the appendix: if $P(z) = \mathcal{N}(z|\mu, \Lambda^{-1})$, $P(x|z) \sim \mathcal{N}(x|Az+b$

Then $p(z|x) \sim \mathcal{N}(z \mid C(A^T L(x-b) + \Lambda\mu), C)$

$C = (\Lambda + A^T L A)^{-1}$

$A = \mu \quad b = 0 \quad \Lambda = 1 = \Lambda^{-1} \quad L^{-1} = \Sigma \quad , \mu = 0$

$C = (1 + \mu^T \Sigma^{-1} \mu)^{-1}$

$C(A^T L(x-b) + \Lambda\mu) = (1 + \mu^T \Sigma^{-1} \mu)^{-1} (\mu^T \Sigma^{-1}(x) + 0)$

$= (1 + \mu^T \Sigma^{-1} \mu)^{-1} (\mu^T \Sigma^{-1} x)$

$\therefore P(z|x;\theta) \sim \mathcal{N}\left(z \,\middle|\, (1+\mu^T \Sigma^{-1}\mu)^{-1} \mu^T \Sigma^{-1} x \,,\, (1+\mu^T\Sigma^{-1}\mu)^{-1}\right)$

scalars

$m = E[z|x] = (1 + \mu^T \Sigma^{-1} \mu)^{-1} \mu^T \Sigma^{-1} x$

$s = E[z^2|x] = \text{Var}(z|x) + (E[z|x])^2$

$= (1 + \mu^T \Sigma^{-1} \mu)^{-1} + [(1 + \mu^T \Sigma^{-1} \mu)^{-1} \mu^T \Sigma^{-1} x]^2$

$x \in \mathbb{R}^D$, $z \in \mathbb{R}^1$, $\mu \in \mathbb{R}^{D \times 1}$

$(1 + \underset{1\times D}{\mu^T} \underset{D\times D}{\Sigma^{-1}} \underset{D\times 1}{\mu})^{-1} \underset{1\times D}{\mu^T} \underset{D\times D}{\Sigma^{-1}} \underset{D\times 1}{x}$ ✓ univariate

$$q = P(z|x) \sim \mathcal{N}\Big(\overbrace{(1+\mu^T\Sigma^{-1}\mu)^{-1}\mu^T\Sigma^{-1}x}^{\mu_1},\ \overbrace{(1+\mu^T\Sigma^{-1}\mu)^{-1}}^{\Sigma_1}\Big)$$

$$p(x|z) \sim \mathcal{N}(z\mu, \Sigma)$$

$$p(z) \sim \mathcal{N}(0, I)$$

b) $\log p(z^n, x^n|\theta) = \log p(x^n|z^n, \theta)\, p(z^n|\theta) = \log p(x^n|z^n, \theta) + \log p(z^n|\theta)$

$p(x|z;\theta) \sim \mathcal{N}(z\mu, \Sigma)$ $\quad p(z) \sim \mathcal{N}(0, I)$

$$① = E_{p(z^n|x^n;\mu_{old},\Sigma_{old})}\left[\log p(x^n|z^n;\mu,\Sigma) + \log p(z^n)\right]$$

$$= \int_{-\infty}^{\infty} dz^n \frac{\exp(\frac{-1}{2}(z^n-\overset{old}{\mu_1})^T\overset{old}{\Sigma_1^{-1}}(z^n-\mu_1))}{\sqrt{(2\pi)|\Sigma_1|}} \left(\overbrace{\frac{-1}{2}(x^n - z^n\mu)^T\Sigma^{-1}(x^n - z^n\mu) - \frac{1}{2}\log((2\pi)^D|\Sigma|)}^{\log P(x^n|z^n;\mu,\Sigma)} \underbrace{-\frac{1}{2}(z^n)^T z^n - \frac{1}{2}\log(2\pi)}_{\log p(z^n)} \right)$$

Recall:

$\int_x f(x)dx = 1$

if $f(x)$ is PDF

$-\frac{1}{2}(z^n)^2$

$$((x^n)^T - \mu^T(z^n)^T)\Sigma^{-1}(x^n - z^n\mu)$$

$$= (x^n)^T\Sigma^{-1}x^n - (x^n)^T\Sigma^{-1}z^n\mu - \mu^T(z^n)^T\Sigma^{-1}x^n + \mu^T(z^n)^T\Sigma^{-1}z^n\mu$$

$$= \underbrace{(x^n)^T\Sigma^{-1}x^n} - \underbrace{z^n(x^n)^T\Sigma^{-1}\mu} - \underbrace{z^n\mu^T\Sigma^{-1}x^n} + \underbrace{(z^n)^2\mu^T\Sigma^{-1}\mu}$$

$$① \quad = (x^n)^T\Sigma^{-1}x^n - (x^n)^T\Sigma^{-1}\mu m^n - \mu^T\Sigma^{-1}x^n m^n + \mu^T\Sigma^{-1}\mu s^n - \frac{1}{2}\log((2\pi)^D|\Sigma|)$$
$$-\frac{1}{2}s^n - \frac{1}{2}\log(2\pi)$$

$$\mathcal{L}(q,\theta) = \frac{1}{N}\sum_{n=1}^{N} ①$$

$$\frac{d\mathcal{L}}{d\mu} = \frac{1}{N}\sum_{n=1}^{N} -(x^n)^T\Sigma^{-1}m^n - (\Sigma^{-1}x^n)^T m^n + 2s^n\mu^T\Sigma^{-1}$$

$$= \frac{1}{N}\sum_{n=1}^{N} -(x^n)^T\Sigma^{-1}m^n - (x^n)^T\Sigma^{-1}m^n + 2s^n\mu^T\Sigma^{-1}$$

Note: $(\Sigma^{-1})^T = \Sigma^{-1}$

$\Sigma^T = \Sigma$

$$0 = \frac{1}{N}\sum_{n=1}^{N} -2(x^n)^T\Sigma^{-1}m^n + 2s^n\mu^T\Sigma^{-1}$$

$$\Sigma(x^n)^T\Sigma^{-1}m^n = \Sigma s^n\mu^T\Sigma^{-1}$$

$$\Sigma m^n\Sigma^{-1}x^n = \Sigma s^n\Sigma^{-1}\mu$$

mult. both sides by Σ

$$\Sigma m^n x^n = \Sigma s^n\mu$$

$$\boxed{\mu = \frac{\Sigma m^n x^n}{\Sigma s^n}}$$

Note: it's a sum at the end, not the covariance. I dropped the bounds for convenience.

Proof of some derivatives used above:

What is $\frac{d}{d\mu}\left(\overset{1\times D}{\mu^T}\,\overset{D\times D}{\Sigma^{-1}}\,\overset{D\times 1}{x^n}\,\overset{1\times 1}{m^n}\right)$?

$\frac{d}{d\mu}(\mu^T b)$, $b \in D\times 1$

$$[\mu_1 \cdots \mu_D]\begin{bmatrix}b_1\\ \vdots\\ b_D\end{bmatrix} = \mu_1 b_1 + \cdots + \mu_D b_D$$

$$\frac{d}{d\mu}\left(\mu_1 b_1 + \dots + \mu_D b_D\right) = [b_1 \cdots b_D] = b^T$$

Previously we showed: $\frac{d(B^TAB)}{dB} = 2B^TA$ if A is diag matrix

What is $\frac{d}{d\mu}\left(\overset{1\times D}{(x^n)^T}\,\overset{D\times D}{\Sigma^{-1}}\,\mu\right)$?

$\frac{d}{d\mu}(b^T\mu)$

$$\hookrightarrow = [b_1 \cdots b_D]\begin{bmatrix}\mu_1\\ \vdots\\ \mu_D\end{bmatrix} = [b_1\mu_1 \cdots b_D\mu_D]$$

$$\therefore \frac{d}{d\mu}(b^T\mu) = b^T$$