The University of Connecticut **Fall 2018**
Dept. of ECE **KRP**

Problem Set # 3
(Due October 22, 2018)

1. (a) Show that if

$$p(\underline{z}) \sim N(\underline{z}; \underline{\mu}_z, \Sigma_z); \underline{z} \in R^{n_z} \text{ and } p(\underline{x} \mid \underline{z}) \sim N(\underline{x}; A\underline{z}, \Sigma); \underline{x} \in R^{n_x}$$

Then

$$p(\underline{z}, \underline{x}) = N\left(\begin{bmatrix}\underline{z}\\ \underline{x}\end{bmatrix}; \begin{bmatrix}\underline{\mu}_z\\ A\underline{\mu}_z\end{bmatrix}, \begin{bmatrix}\Sigma_z & \Sigma_z A^T\\ A\Sigma_z & A\Sigma_z A^T + \Sigma\end{bmatrix}\right)$$

$$\begin{aligned}
E(\underline{z} \mid \underline{x}) &= \underline{\mu}_z + \Sigma_z A^T (A\Sigma_z A^T + \Sigma)^{-1}(\underline{x} - A\underline{\mu}_z)\\
&= \left(I - \Sigma_z A^T (A\Sigma_z A^T + \Sigma)^{-1} A\right)\underline{\mu}_z + \Sigma_z A^T (A\Sigma_z A^T + \Sigma)^{-1}\underline{x}\\
&= \left(I - \Sigma_z A^T \Sigma^{-1}(A\Sigma_z A^T \Sigma^{-1} + I_x)^{-1} A\right)\underline{\mu}_z + \Sigma_z A^T \Sigma^{-1}(A\Sigma_z A^T \Sigma^{-1} + I)^{-1}\underline{x}\\
&= [I + \Sigma_z A^T \Sigma^{-1} A]^{-1}\underline{\mu}_z + \Sigma_z A^T \Sigma^{-1}[I - A\Sigma_z (I + A^T \Sigma^{-1} A\Sigma_z)^{-1} A^T \Sigma^{-1}]\underline{x}\\
&= \left(\Sigma_z^{-1} + A^T \Sigma^{-1} A\right)^{-1} \Sigma_z^{-1}\underline{\mu}_z + \Sigma_z[I - A^T \Sigma^{-1} A(\Sigma_z^{-1} + A^T \Sigma^{-1} A)^{-1}] A^T \Sigma^{-1}\underline{x}\\
&= \left(\Sigma_z^{-1} + A^T \Sigma^{-1} A\right)^{-1}\left(A^T \Sigma^{-1}\underline{x} + \Sigma_z^{-1}\underline{\mu}_z\right)
\end{aligned}$$

$$\text{covar}(\underline{z} \mid \underline{x}) = \Sigma_{z|x} = \left(\Sigma_z^{-1} + A^T \Sigma^{-1} A\right)^{-1} = \Sigma_z - \Sigma_z A^T (\Sigma + A\Sigma_z A^T)^{-1} A\Sigma_z$$

(b) Let $\underline{x}$ be a random p-vector following the normal distribution $N(\underline{x}; \underline{\mu}, \Sigma)$. If the prior for $\underline{\mu}$ is also normal $N(\underline{\mu}; \underline{\mu}_0, \Sigma_0)$ and $\{\underline{x}_n : n=1,2,..,N\}$ are i.i.d. observations, compute the posterior distribution $p(\underline{\mu} \mid \underline{x}_1, \underline{x}_2, ..., \underline{x}_N)$ using the result in (a). Express the result in the simplest possible form.

2. Let $(\underline{x}_k^j : j = 1, 2, .., n_k; k = 1, 2, .., C\}$ be the training data. Let $\underline{\mu}$ be the overall sample mean and $(\underline{\mu}_k : k = 1, 2, .., C\}$ be the sample means for each class k. Show that:

$$\sum_{k=1}^{C}\sum_{j=1}^{n_k}(\underline{x}_k^j - \underline{\mu})\,(\underline{x}_k^j - \underline{\mu})^T = \sum_{k=1}^{C}\sum_{j=1}^{n_k}(\underline{x}_k^j - \underline{\mu}_k)\,(\underline{x}_k^j - \underline{\mu}_k)^T + \sum_{k=1}^{K} n_k(\underline{\mu}_k - \underline{\mu})\,(\underline{\mu}_k - \underline{\mu})^T$$

That is, the total variability (i.e., total covariance) in the data is the sum of individual class variability (i.e., within-covariance) and between-class variability (i.e., between class-covariance).

3. The purpose of this problem is to derive the Bayesian classifier for the d-dimensional multivariate Bernoulli case. Let the conditional probability mass function for a given category be given by

$$p(\underline{x}|\underline{\theta}) = \prod_{i=1}^{d} \theta_i^{x_i}(1-\theta_i)^{1-x_i}$$

Let $D = \{\underline{x}_1, \underline{x}_2, \ldots\ldots, \underline{x}_n\}$ be a set of n samples independently drawn according to this probability mass function.

a. If $\underline{s} = [s_1, s_2, \ldots, s_d]^T$ is the sum of the n samples along each dimension, show that

$$p(D|\underline{\theta}) = \prod_{i=1}^{d} \theta_i^{s_i}(1-\theta_i)^{n-s_i}$$

b. Assuming a uniform prior distribution for $\underline{\theta}$ and using the identity

$$\int_0^1 \theta^m (1\text{-}\theta)^n d\theta = \frac{m!\,n!}{(m+n+1)!}$$

show that

$$p(\underline{\theta}|D) = \prod_{i=1}^{d} \frac{(n+1)!}{s_i!\,(n-s_i)!}\theta_i^{s_i}(1-\theta_i)^{n-s_i}$$

Sketch this density for the case d=1, n =1 and for the two resulting possibilities for s_1 of 0 and 1.

c. Using $p(\underline{x}|D) = \int p(\underline{x}|\underline{\theta})p(\underline{\theta}|D)d\underline{\theta}$, show that

$$p(\underline{x}|D) = \prod_{i=1}^{d} \left(\frac{s_i+1}{n+2}\right)^{x_i} \left(1 - \frac{s_i+1}{n+2}\right)^{1-x_i}$$

d. What is the effective Bayesian estimate for $\underline{\theta}$ based on observed data?

4. Suppose in a C-category supervised learning environment, we sample the full distribution $p(\underline{x})$ and subsequently train a PNN classifier.
   a. Show that even if there are unequal category priors and hence unequal numbers of points in each category, PNN properly accounts for such priors.
   b. Suppose we have trained a PNN with the assumption of equal category priors, but later wish to use it for a problem having the cost matrix $[\lambda_{ij}]$, representing the cost of choosing category i when in fact the pattern came from j. How should we do this?
   c. Suppose instead we know the cost matrix $[\lambda_{ij}]$ *before* training. How should we train PNN for minimum risk?

5. Bishop, Chapter 9, Page 456, Problem 9.10

6. Bishop, Chapter 9, Page 458, Problem 9.19.

7. Bishop, Chapter 10, Page 518, Problem 10.12.

8. Bishop, Chapter 10, Page 519, Problem 10.14

9. (Computational. Due October 22, 2018)

On the four sample data sets of your choice from the UCI data, experiment with the following classifiers:

a. Probabilistic Neural Network
b. K-nearest Neighbor (K=1,3,5)