Multivariate Analysis - Homework 4

Please upload your homework on SPOC before 8:00pm, April 20, including all details needed. For R exercises, R markdown is highly encouraged; for other parts, try to use LaTex.

1. Verify that the two-population Fisher's LDA is a special case of the g-group case.

2. For two-group Bayes classification rule, express the population-level TPM using prior probabilities p_1 and p_2, densities of the two populations $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$,and the decision rules R_1 and R_2 derived in class. Assuming equal prior probability and equal misclassification cost, prove that with such choices of R_1 and R_2, for two normal distributions $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$,

$$\text{TPM} = \Phi(\frac{-\Delta}{2}),$$

   where $\Delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, and $\Phi(\cdot)$ is the CDF of standard normal random variables. (Hint: Consider the distribution of discriminant function.)

3. Derive the g-group Bayes rule with equal misclassification costs based on normal distributions $f_k(\mathbf{x}) \sim N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, $k = 1, \ldots, g$, with unequal covariance matrices. Also provide the estimated version. Show that it's a quadratic classification rule.

4. Show that under the multiple linear regression setting, (all notations follow the slides)

   (a) The hat/projection matrix $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ of the space spanned by $\mathbf{X}$ is both symmetric and idempotent, i.e., $\mathbf{P}^2 = \mathbf{P}$. So is **I**-**P**, where $\mathbf{I}$ is $n \times n$ identity matrix.

   (b) $tr(\mathbf{P}) = q + 1$.

   (c) $\hat{\mathbf{y}} = \mathbf{P}\mathbf{y}$, and the residual vector $\hat{\boldsymbol{\varepsilon}} = (\mathbf{I} - \mathbf{P})\mathbf{y}$.

   (d) $SSE = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \boldsymbol{\varepsilon}'(\mathbf{I} - \mathbf{P})\boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ is the error vector. Note the difference between $\hat{\boldsymbol{\varepsilon}}$ and $\boldsymbol{\varepsilon}$.

   (e) The mean squared error $MSE = SSE/(n - q - 1)$ is an unbiased estimator of error variance σ^2. Hint: (1) $a = tr(a)$ if a is a scalar. (2) $tr(\mathbf{AB}) = tr(\mathbf{BA})$.

5. Show that the coefficient of determination R^2 is the square of the multiple correlation, that is, $R^2 = \widehat{corr}^2(Y, \hat{Y})$, where $\widehat{corr}(\cdot)$ stands for the sample/Pearson correlation.

6. Suppose a univariate random variable Y has a normal distribution with variance 4. If Y is from population G_1, its mean is 10; if it is from population G_2, its mean is 14.

(a) Assume equal prior probabilities for the events $A_1 = \{Y \text{ is from population } G_1\}$ and $A_2 = \{Y \text{ is from population } G_2\}$, and assume that the misclassification costs $c(1|2)$ and $c(2|1)$ are equal, both \$10. We decide that we shall allocate Y to population G_1 if $Y \leq c$, for some c to be determined, and to population G_2 if $Y > c$. Let event $B_1 = \{Y \text{ is classified into } G_1\}$ and $B_2 = \{Y \text{ is classified into } G_2\}$. Fill in the following table. And determine the optimal c that yields smallest ECM among the following choices of c.

c	$P(B_1\|A_2)$	$P(B_2\|A_1)$	$P(A_1 \text{ and } B_2)$	$P(A_2 \text{ and } B_1)$	TPM	ECM
9						
10						
⋮						
14						

(b) Repeat (a) if the prior probabilities of A_1 and A_2 are equal, but $c(2|1) = \$5$ and $c(1|2) = \$15$.

7. (R exercise.) Satellite applications motivated the development of a silver-zinc battery. The following table contains failure data collected to characterize the performance of the battery during its life cycle. (Data attached.)

Z_1 Charge rate (amps)	Z_2 Discharge rate (amps)	Z_3 Depth of discharge (% of rated ampere-hours)	Z_4 Temperature (°C)	Z_5 End of charge voltage (volts)	Y Cycles to failure
.375	3.13	60.0	40	2.00	101
1.000	3.13	76.8	30	1.99	141
1.000	3.13	60.0	20	2.00	96
1.000	3.13	60.0	20	1.98	125
1.625	3.13	43.2	10	2.01	43
1.625	3.13	60.0	20	2.00	16
1.625	3.13	60.0	20	2.02	188
.375	5.00	76.8	10	2.01	10
1.000	5.00	43.2	10	1.99	3
1.000	5.00	43.2	30	2.01	386
1.000	5.00	100.0	20	2.00	45
1.625	5.00	76.8	10	1.99	2
.375	1.25	76.8	10	2.01	76
1.000	1.25	43.2	10	1.99	78
1.000	1.25	76.8	30	2.00	160
1.000	1.25	60.0	0	2.00	3
1.625	1.25	43.2	30	1.99	216
1.625	1.25	60.0	20	2.00	73
.375	3.13	76.8	30	1.99	314
.375	3.13	60.0	20	2.00	170

(a) Find the estimated linear regression of Y on Z's. Plot the residuals from the fitted model to check model assumptions.

(b) Find the estimated linear regression of $\ln(Y)$ on an appropriate subset of predictor variables.

(c) Plot the residuals from the fitted model chosen in (b) to check the model assumption, including the normality assumption.

(d) Conduct statistical inference for the model in (b).