2.6.26

26. Let $p(\mathbf{x}|\omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ for a two-category d-dimensional problem with the same covariances but arbitrary means and prior probabilities. Consider the squared Mahalanobis distance

$$r_i^2 = (\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_i).$$

(a) Show that the gradient of r_i^2 is given by

$$\nabla r_i^2 = 2\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i).$$

(b) Show that at any position on a given line through $\boldsymbol{\mu}_i$ the gradient ∇r_i^2 points in the same direction. Must this direction be parallel to that line?

(c) Show that ∇r_1^2 and ∇r_2^2 point in opposite directions along the line from $\boldsymbol{\mu}_1$ to $\boldsymbol{\mu}_2$.

(d) Show that the optimal separating hyperplane is tangent to the constant probability density hyperellipsoids at the point that the separating hyperplane cuts the line from $\boldsymbol{\mu}_1$ to $\boldsymbol{\mu}_2$.

(e) True of False: For a two-category problem involving normal densities with arbitrary means and covariances, and $P(\omega_1) = P(\omega_2) = 1/2$, the Bayes decision boundary consists of the set of points of equal Mahalanobis distance from the respective sample means. Explain.

3.2.8

8. Consider an extreme case of the general issue discussd in Problem 7, one in which it is possible that the maximum likelihood solution leads to the *worst* possible classifier, i.e., one with an error that approaches 100% (in probability). Suppose our data in fact comes from two one-dimensional distributions of the forms

$$\begin{aligned} p(x|\omega_1) &\sim [(1-k)\delta(x-1) + k\delta(x+X)] \quad \text{and} \\ p(x|\omega_2) &\sim [(1-k)\delta(x+1) + k\delta(x-X)], \end{aligned}$$

where X is positive, $0 \leq k < 0.5$ represents the portion of the total probability mass concentrated at the point $\pm X$, and $\delta(\cdot)$ is the Dirac delta function. Suppose our poor models are of the form $p(x|\omega_1, \mu_1) \sim N(\mu_1, \sigma_1^2)$ and $p(x|\omega_2, \mu_2) \sim N(\mu_2, \sigma_2^2)$ and we form a maximum likelihood classifier.

(a) Consider the symmetries in the problem and show that in the infinite data case the decision boundary will always be at $x = 0$, regardless of k and X.

(b) Recall that the maximum likelihood estimate of either mean, $\hat{\mu}_i$, is the mean of its distribution. For a fixed k, find the value of X such that the maximum likelihood estimates of the means "switch," i.e., where $\hat{\mu}_1 \geq \hat{\mu}_2$.

(c) Plot the true distributions and the Gaussian estimates for the particular case $k = .2$ and $X = 5$. What is the classification error in this case?

(d) Find a dependence $X(k)$ which will guarantee that the estimated mean $\hat{\mu}_1$ of $p(x|\omega_1)$ is less than zero. (By symmetry, this will also insure $\hat{\mu}_2 > 0$.)

(e) Given your $X(k)$ just derived, state the classification error in terms of k.

(f) Suppose we constrained our model space such that $\sigma_1^2 = \sigma_2^2 = 1$ (or indeed any other constant). Would that change the above results?

(g) Discuss how if our model is wrong (here, does not include the delta functions), the error can approaches 100% (in probability). Does this surprising answer arise because we have found some local minimum in parameter space?

3.2.10

10. Suppose we employ a novel method for estimating the mean of a data set $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$: we assign the mean to be the value of the first point in the set, i.e., $\mathbf{x}_1$.

(a) Show that this method is unbiased.

(b) State why this method is nevertheless highly undesirable.

3.4.13

13. Let $p(\mathbf{x}|\boldsymbol{\Sigma}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu}$ is known and $\boldsymbol{\Sigma}$ is unknown. Show that the maximum likelihood estimate for $\boldsymbol{\Sigma}$ is given by

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu})^t$$

by carrying out the following argument:

(a) Prove the matrix identity $\mathbf{a}^t\mathbf{A}\mathbf{a} = \mathrm{tr}[\mathbf{A}\mathbf{a}\mathbf{a}^t]$, where the trace, $\mathrm{tr}[\mathbf{A}]$, is the sum of the diagonal elements of $\mathbf{A}$.

(b) Show that the likelihood function can be written in the form

$$p(\mathbf{x}_1, ..., \mathbf{x}_n|\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{nd/2}}|\boldsymbol{\Sigma}^{-1}|^{n/2}\exp\left[-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\sum_{k=1}^{n}(\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu})^t\right]\right].$$

(c) Let $\mathbf{A} = \boldsymbol{\Sigma}^{-1}\widehat{\boldsymbol{\Sigma}}$ and $\lambda_1, ..., \lambda_n$ be the eigenvalues of $\mathbf{A}$; show that your result above leads to

$$p(\mathbf{x}_1, ..., \mathbf{x}_n|\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{nd/2}|\widehat{\boldsymbol{\Sigma}}|^{n/2}}(\lambda_1 \cdots \lambda_d)^{n/2}\exp\left[-\frac{n}{2}(\lambda_1 + \cdots + \lambda_d)\right].$$

(d) Complete the proof by showing that the likelihood is maximized by the choice $\lambda_1 = \cdots = \lambda_d = 1$. Explain your reasoning.