ECE 271A: Homework #2

Due on October 30, 2023 at 11:59pm

Professor Vasconcelos

Ray Tsai

A16848188

Problem 1

Let $p(x|\omega_i) \sim N(\mu_i, \Sigma)$ for a two-catagory d-dimensional problem with the same covariances but arbitrary means and prior probabilities. Consider the squared Mahalanobis distance

$$r_i^2 = (x - \mu_i)^T \Sigma^{-1} (x - \mu_i).$$

Part A

Show that the gradient of r_i^2 is given by

$$\nabla r_i^2 = 2\Sigma^{-1}(x - \mu_i).$$

Solution

Notice that

$$
\begin{aligned}
r_i^2 &= (x - \mu_i)^T \Sigma^{-1}(x - \mu_i) \\
&= x^T \Sigma^{-1} x - 2\mu_i^T \Sigma^{-1} x + \mu_i^T \Sigma^{-1} \mu_i \\
&= \sum_k \sum_j \Sigma_{kj}^{-1} x_k x_j - 2\sum_k \sum_j \Sigma_{kj}^{-1} \mu_{i_k} x_j + \sum_k \sum_j \Sigma_{kj}^{-1} \mu_{i_k} \mu_{i_j}.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\frac{\partial r_i^2}{\partial x_m} &= \frac{\partial}{\partial x_m} \left(\sum_k \sum_j \Sigma_{kj}^{-1} x_k x_j - 2\sum_k \sum_j \Sigma_{kj}^{-1} \mu_{i_k} x_j + \sum_k \sum_j \Sigma_{kj}^{-1} \mu_{i_k} \mu_{i_j} \right) \\
&= \frac{\partial}{\partial x_m} \left(\sum_k \sum_j \Sigma_{kj}^{-1} x_k x_j \right) - 2\sum_k \Sigma_{km}^{-1} \mu_{i_k} \\
&= \frac{\partial}{\partial x_m} \left(\Sigma_{mm}^{-1} x_m^2 + \sum_{k \neq m} \Sigma_{km}^{-1} x_k x_m + \sum_{j \neq m} \Sigma_{mj}^{-1} x_m x_j + \sum_{k \neq m} \sum_{j \neq m} \Sigma_{kj}^{-1} x_k x_j \right) - 2\sum_k \Sigma_{km}^{-1} \mu_{i_k} \\
&= 2\Sigma_{mm}^{-1} x_m + \sum_{k \neq m} \Sigma_{km}^{-1} x_k + \sum_{j \neq m} \Sigma_{mj}^{-1} x_j - 2\sum_k \Sigma_{km}^{-1} \mu_{i_k} \\
&= \sum_k \Sigma_{km}^{-1} x_k + \sum_j \Sigma_{mj}^{-1} x_j - 2\sum_k \Sigma_{km}^{-1} \mu_{i_k} \\
&= 2\sum_k \Sigma_{mk}^{-1}(x_k - \mu_{i_k}).
\end{aligned}
$$

Notice that $\frac{\partial r_i^2}{\partial x_m}$ is simply dot product of the m-th row of Σ^{-1} and $(x - \mu)$ times 2, and so $\nabla r_i^2 = 2\Sigma^{-1}(x - \mu_i)$.

Part B

Show that at any position on a given line through μ_i the gradient ∇r_i^2 points in the same direction. Must this direction be parellel to that line?

Solution

Let $L = \{ax + \mu_i \,|\, a \in \mathbb{R}\}$, for some $x \in \mathbb{R}^d$. Let $y = lx + \mu_i \in L$. Then, $\nabla_i^2 r(y) = \nabla r_i^2 = 2\Sigma^{-1}(y - \mu_i) = 2l\Sigma^{-1}x$, and so ∇r_i^2 is not parellel to the line L unless $\Sigma^{-1} = \Sigma = I$.

Part C

Show that ∇r_1^2 and ∇r_2^2 point in opposite directions along the line from μ_1 to μ_2.

Solution

Let $p = a\mu_1 + (1-a)\mu_2$ be a point on the line from μ_1 to μ_2, for $a \in [0,1]$. Then,

$$\nabla r_1^2(p) = 2\Sigma^{-1}(p - \mu_1) = 2(a-1)\Sigma^{-1}(\mu_1 - \mu_2) \tag{1}$$

$$\nabla r_2^2(p) = 2\Sigma^{-1}(p - \mu_2) = 2a\Sigma^{-1}(\mu_1 - \mu_2). \tag{2}$$

When $a = 0$ or 1, a or $a-1$ become 0. When $a \in (0,1)$, a and $a-1$ has opposite signs. Thus, we conclude that $\nabla r_1^2(p)$ and $\nabla r_2^2(p)$ points toward opposite directions.

Part 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 μ_1 to μ_2.

Solution

We know the optimal decision rule is $i^*(x) = \arg\max_i \ln P_{X|Y}(x|i) + \ln P_Y(i)$. Assume that $P_{X|Y}(x|i)$ is some Gaussian distribution with mean μ_i and covariance matrix Σ, denoted as $G(x, \mu_i, \Sigma)$. Then, our decision rule becomes

$$i^*(x) = \arg\max_i \; -\frac{1}{2}(x - \mu_i)^T \Sigma^{-1}(x - \mu_i) + \ln P_Y(i)$$

$$= \arg\max_i \mu_i^T \Sigma^{-1} x - \frac{1}{2}\mu_i^T \Sigma^{-1}\mu_i + \ln P_Y(i).$$

Let $w_i = \mu_i^T \Sigma^{-1}$ and $w_{i0} = -\frac{1}{2}\mu_i^T \Sigma^{-1}\mu_i + \ln P_Y(i)$. Thus, the optimal decision place between class 1 and 2 is

$$\mu_1^T \Sigma^{-1} x + w_{i1} - \mu_1^T \Sigma^{-2} x - w_{i2} = (\Sigma^{-1}(\mu_1 - \mu_2))^T x + (w_{i1} - w_{i2}) = 0.$$

Let p be a point on the line from μ_1 to μ_2 where the decision plane cuts through. From part C, we know that $\nabla r_1^2(p) = 2(a-1)\Sigma^{-1}(\mu_1 - \mu_2)$ and $\nabla r_2^2(p) = 2a\Sigma^{-1}(\mu_1 - \mu_2)$. Notice that the direction of the plane $\Sigma^{-1}(\mu_1 - \mu_2)$ is parallel to both $\nabla r_1^2(p)$ and $\nabla r_1^2(p)$, which implies that the plane is orthogonal to the gradients at point p. Since gradients of a function are also orthogonal to the constant contours of its function, we thus know that the optimal decision plane is tangent to the constant probability density function hyperellipsoids at point p.

Part E

True or False: For a two-catagory problem involving normal densities with arbitrary means and covariances, and $P_Y(w_1) = P_Y(w_2) = \frac{1}{2}$, the Bayes decision boundary consists of the set of points of equal Mahalanobis distance from the respective sample means.

Solution

False. Although the decision boundary consists of the points of equal Mahalanobis distance from the two class means in this case, we cannot guarantee the sample means equal the "true" class means.

Problem 2

In this problem we will consider the ML estimate of the parameters of a multinomial distribution. Consider a random variable X such that $P_X(k) = \pi_k, k \in \{1, \ldots, N\}$. Suppose we draw n independent observations from X and form a random vector $C = (C_1, \ldots, C_N)^T$, where C_k is the number of times that the observed value is k (i.e. C is the histogram of the sample of observations). Then, C has multinomial distribution

$$P_{C_1,\ldots,C_N}(c_1,\ldots,c_N) = \frac{n!}{\prod_{k=1}^N c_k!} \prod_{j=1}^N \pi_j^{c_j}.$$

Part A

Derive the ML estimator for the parameters π_i, $i = 1, \ldots, N$.

Solution

We first take the logarithm of $P_{C_1,\ldots,C_N}(c_1,\ldots,c_N)$ and get

$$\ln P_{C_1,\ldots,C_N}(c_1,\ldots,c_N) = \ln \frac{n!}{\prod_{k=1}^N c_k!} + \sum_{j=1}^N c_j \ln \pi_j.$$

Let $\theta = (\pi_1, \ldots, \pi_N)^T$, and let $L(\theta, \lambda) = \ln P_{C_1,\ldots,C_N}(c_1,\ldots,c_N) + \lambda \left(\sum_j^N \pi_j - 1 \right)$. Then,

$$\nabla_\theta L = \left(\frac{c_1}{\pi_1}, \ldots, \frac{c_N}{\pi_N} \right)^T + (\lambda, \ldots, \lambda)^T = 0$$

$$\frac{\partial}{\partial \lambda} L = \sum_{j=1}^N \pi_j - 1 = 0.$$

For $1 \leq j \leq N$, we get $\frac{c_j}{\pi_j} + \lambda = 0$, and so $c_j + \pi_j \lambda = 0$. Then, $\sum_{j=1}^N (c_j + \pi_j \lambda) = n + \lambda = 0$, and thus $\lambda = -n$. We also take the Hessian of L. Notice that

$$\frac{\partial^2 L}{\partial \pi_i \partial \pi_j} = \begin{cases} -\frac{c_j}{\pi_j^2}, & i = j \\ 0, & i \neq j \end{cases},$$

and so $\nabla_\theta^2 L = \mathrm{diag} \left(-\frac{c_i}{\pi_i^2}, \ldots, -\frac{c_N}{\pi_N^2} \right)$. Since the Hessian of L is obviously negative definite, the critical point we are looking for is indeed a maximum point. Therefore, we obtain the ML estamator $\hat\theta = \left(\frac{C_1}{n}, \ldots, \frac{C_N}{n} \right)$.

Part B

Is the estimator derived in part A unbiased? What is its variance? Is this a good estimator? Why?

Solution Since

$$Bias(\hat\pi_j) = E[\hat\pi_j - \pi_j]$$
$$= \frac{1}{n} E[C_j] - \pi_j$$
$$= \frac{1}{n} \cdot n\pi_j - \pi_j = 0,$$

4

$\hat{\theta}$ is unbiased.

We also know

$$Var(\hat{\pi}_j) = \frac{1}{n^2} Var(C_j) = \frac{\pi_j(1 - \pi_j)}{n}.$$

This shows that our estimator converges to the perfect estimation as we increase our sample size, as thus $\hat{\theta}$ is a good estimator.

Problem 3

Consider an extreme case of the general issue discussed 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%. Suppose our data in fact comes from two one-dimensional distribution of the forms

$$p(x|\omega_1) \sim [(1-k)\delta(x-1) + k\delta(x+X)] \tag{3}$$

$$p(x|\omega_2) \sim [(1-k)\delta(x+1) + k\delta(x-X)], \tag{4}$$

where X is positive, $0 \leq k < 0.5$ represents the portion of the total probability mass concentrated at the point $\pm X$, and $\delta(.)$ 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.

Part 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.

Solution

We first calculate the means and variances of $p(x|\omega_1)$ and $p(x|\omega_2)$ and get

$$\mu_1 = (1-k) - kX \tag{5}$$

$$\mu_2 = kX - (1-k) \tag{6}$$

$$\sigma_1^2 = (1-k)(\mu_1 - 1)^2 + k(\mu_1 + X)^2 \tag{7}$$

$$\sigma_2^2 = (1-k)(\mu_2 + 1)^2 + k(\mu_2 - X)^2. \tag{8}$$

Notice that $\sigma_2^2 = \sigma_1^2 = \sigma^2$, as $\mu_1 = -\mu_2$. In the infinite data case, the sample means and variances converges to the the theoretic mean of our model, and so the decision boundary is

$$x = \frac{\mu_1 + \mu_2}{2} - \frac{\sigma^2}{\mu_2 - \mu_1} \ln \frac{P_Y(\omega_1)}{P_Y(\omega_2)} = 0.$$

Part 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 $\hat{\mu}_1 \geq \hat{\mu}_2$.
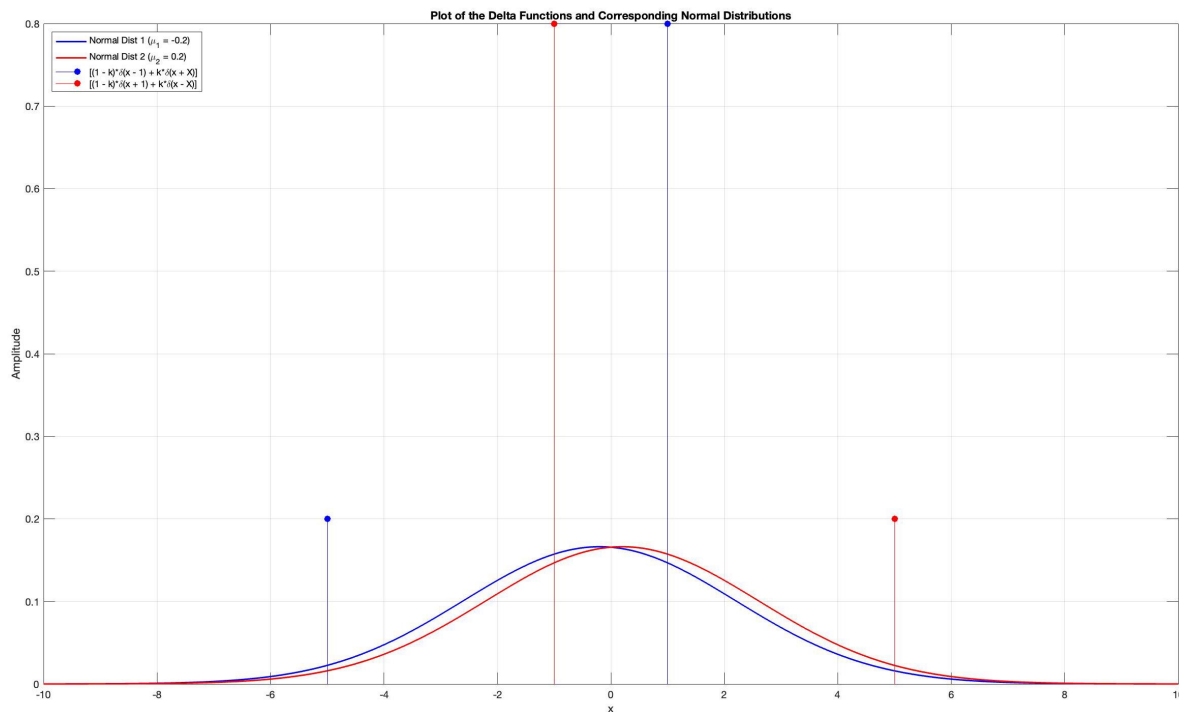
Solution

$$\hat{\mu}_1 \geq \hat{\mu}_2 = -\hat{\mu}_1$$

$$\hat{\mu}_1 \geq 0$$

$$(1-k) - kX \geq 0$$

$$\frac{1-k}{k} \geq X.$$

Thus, the means "switch" when $X = \frac{1-k}{k}$.

6

Part 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?

Solution

The error is

$$
\begin{aligned}
P_{X,Y}(i(x) \neq y) &= P_{X|Y}(i(x) = 1|y = 2)P_Y(2) + P_{X|Y}(i(x) = 2|y = 1)P_Y(1) \\
&= P_{X|Y}(x \leq 0|y = 2)P_Y(2) + P_{X|Y}(x \geq 0|y = 1)P_Y(1) \\
&= P_{X|Y}(x \leq 0|y = 2)(P_Y(1) + P_Y(2)) \\
&= P_{X|Y}(x \leq 0|y = 2) \\
&= 1 - k = 0.8.
\end{aligned}
$$

Part 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 $\hat{\mu}_2 > 0$.)

Solution

Consider $X = \frac{1}{k}$. Since, $\hat{\mu}_1 = (1 - k) - kX = -k < 0$, X meets the requirements.

7

Part E

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

Solution

Did it in part C already bruh. $1 - k$.

Part 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?

Solution

No. The decision boundary would not change unless $\sigma_1^2 \neq \sigma_2^2$, and thus the classification error would also not be affected.

Part 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?

Solution

Consider $k \to 0$. Then, according to the result we concluded in part E, the classification error approaches 100%. This is not due to finding the local minimum but simply a class classification mismatch.

Problem 4

Suppose we employ a novel method for estimating the mean of a data set $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$: we assign the mean to be the value of the first point in the set, i.e., x_1.

Part A

Show that this method is unbiased.

Solution

Suppose that $\hat{\mu} = x_1$. Then, $Bias(\hat{\mu}) = E[\hat{\mu} - \mu] = E[x_1] - \mu = \mu - \mu = 0$, and thus $\hat{\mu}$ is unbiased.

Part B

State why this method is nevertheless highly undesirable.

Solution

We check the variance of this estimator. Since

$$Var(\hat{\mu}) = Var(x_1),$$

the variance of the estimator does not decrease even if we increase the sample size, and so $\hat{\mu}$ is not a good estimator.

Problem 5

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

$$\hat{\Sigma} = \frac{1}{n} \sum_{k=1}^{n} (x_k - \mu)(x_k - \mu)^T$$

by carrying out the following argument:

Part A

Prove the matrix identity $a^T A a = tr[Aaa^T]$, where the trace, $tr[A]$, is the sum of the diagonal elements of A.

Solution

On the RHS, we get $a^T A a = \sum_i \sum_j A_{ij} a_i a_j$. On the LHS, we get

$$\begin{aligned}
tr[Aaa^T] &= \sum_i (Aaa^T)_{ii} \\
&= \sum_i A_{i\cdot}(aa^T)_{\cdot i} \\
&= \sum_i \sum_j A_{ij}(aa^T)_{ji} \\
&= \sum_i \sum_j A_{ij} a_i a_j.
\end{aligned}$$

Thus, we know $a^T A a = \sum_i \sum_j A_{ij} a_i a_j = tr[Aaa^T]$.

Part B

Show that the likelihood function can be written in the form

$$p(x_1, \ldots, x_n|\Sigma) = \frac{1}{((2\pi)^d|\Sigma|)^{\frac{n}{2}}} \exp\left[-\frac{1}{2}tr\left[\Sigma^{-1}\sum_{k=1}^{n}(x_k - \mu)(x_k - \mu)^T\right]\right].$$

Solution

$$\begin{aligned}
p(x_1, \ldots, x_n|\Sigma) &= \prod_{k=1}^{n} \frac{1}{((2\pi)^d|\Sigma|)^{\frac{n}{2}}} \exp\left[-\frac{1}{2}(x_k - \mu)^T\Sigma^{-1}(x_k - \mu)\right] \\
&= \frac{1}{((2\pi)^d|\Sigma|)^{\frac{n}{2}}} \exp\left[-\frac{1}{2}\sum_{k=1}^{n}(x_k - \mu)^T\Sigma^{-1}(x_k - \mu)\right] \\
&= \frac{1}{((2\pi)^d|\Sigma|)^{\frac{n}{2}}} \exp\left[-\frac{1}{2}\sum_{k=1}^{n}tr\left[\Sigma^{-1}(x_k - \mu)(x_k - \mu)^T\right]\right] \\
&= \frac{1}{((2\pi)^d|\Sigma|)^{\frac{n}{2}}} \exp\left[-\frac{1}{2}tr\left[\Sigma^{-1}\sum_{k=1}^{n}(x_k - \mu)(x_k - \mu)^T\right]\right].
\end{aligned}$$

10

Part C

Let $A = \Sigma^{-1}\hat{\Sigma}$ and $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of A. Show that your result above leads to

$$p(x_1, \ldots, x_n | \Sigma) = \left(\frac{\prod_{k=1}^d \lambda_k}{(2\pi)^d |\hat{\Sigma}|} \right)^{\frac{n}{2}} \exp\left[-\frac{n}{2}(\lambda_1 + \cdots + \lambda_d) \right].$$

Solution

Since $|A| = \prod_{k=1}^d \lambda_k$ and $\mathrm{tr}(A) = \sum_{k=1}^n \lambda^k$,

$$\begin{aligned}
p(x_1, \ldots, x_n | \Sigma) &= \frac{1}{((2\pi)^d |\Sigma|)^{\frac{n}{2}}} \exp\left[-\frac{1}{2}\mathrm{tr}\left[\Sigma^{-1} \sum_{k=1}^n (x_k - \mu)(x_k - \mu)^T \right] \right] \\
&= \left(\frac{|A\hat{\Sigma}^{-1}|}{(2\pi)^d} \right)^{\frac{n}{2}} \exp\left[-\frac{1}{2}\mathrm{tr}\left[n\Sigma^{-1}\hat{\Sigma} \right] \right] \\
&= \left(\frac{\prod_{k=1}^d \lambda_k}{(2\pi)^d |\hat{\Sigma}|} \right)^{\frac{n}{2}} \exp\left[-\frac{n}{2}(\lambda_1 + \cdots + \lambda_d) \right].
\end{aligned}$$

Part D

Complete the proof by showing that the likelihood is maximized by the choice $\lambda_1 = \cdots = \lambda_d = 1$.

Solution

Let $\lambda = (\lambda_1, \ldots, \lambda_d)^T$

$$\begin{aligned}
g(x_1, \ldots, x_n | \lambda) &= \ln p(x_1, \ldots, x_n | \Sigma) \\
&= \frac{n}{2} \sum_{k=1}^d \ln \lambda_k - \frac{dn}{2} \ln 2\pi - \frac{n}{2} \ln |\hat{\Sigma}| - \frac{n}{2} \sum_{i=1}^d \lambda_i \\
&= \frac{n}{2} \left[\sum_{k=1}^n \ln(\lambda_k) - \lambda_k \right] - \frac{n}{2} \left[d \ln 2\pi - \ln |\hat{\Sigma}| \right].
\end{aligned}$$

Then,

$$\nabla_\lambda g = \frac{n}{2} \begin{bmatrix} \lambda_1^{-1} - 1 \\ \vdots \\ \lambda_d^{-1} - 1 \end{bmatrix} = 0,$$

and so the $\lambda = (1, \ldots, 1)^T$ is a critical point. We now take the Hessian of g. Notice that

$$\frac{\partial^2 g}{\partial \lambda_i \lambda_j} = \begin{cases} -\lambda_i^{-2} & i = j \\ 0 & i \neq j \end{cases},$$

and so $\nabla_\lambda^2 = \mathrm{diag}(-\lambda_1^{-2}, \ldots, -\lambda_d^{-2})$, which is obviously negative definite. The likelihood is indeed maximized by the choice $\lambda = (1, \ldots, 1)^T$. It immediately follows that $A = \Sigma^{-1}\hat{\Sigma} = I$, so $\Sigma = \hat{\Sigma}$, and we are done.