CS189 HW3

David Chen

February 2022

Contents

1 Write-up and Honor Code

Collaborated with: N/A.

I certify that all solutions are entirely in my own words and that I have not looked at another student's solutions. I have given credit to all external sources I consulted.

Signed: David Chen

2 Gaussian Classification

2.1 Part 1

Since the loss function is symmetric and the two classes are assumed to have the same variance, this is equivalent to Linear Discriminant Analysis. Therefore, we can first define

$$\begin{aligned} g(x) &= Q_{C_1}(x) - Q_{C_2}(x) \\ &= \frac{(\mu_{C_1} - \mu_{C_2}) \cdot x}{\sigma^2} - \frac{\|\mu_{C_1}\|^2 - \|\mu_{C_2}\|^2}{2\sigma^2} + \ln \pi_{C_1} - \ln \pi_{C_2} \\ &= \frac{(\mu_{C_1} - \mu_{C_2}) \cdot x}{\sigma^2} - \frac{\mu_{C_1}^2 - \mu_{C_2}^2}{2\sigma^2}. \end{aligned}$$

The Bayes optimal decision boundary is the vector X that satisfy $g(X) = 0$. Since this is a one-dimensional classification problem, this vector is just the scalar $X = \frac{\mu_1 + \mu_2}{2}$. Therefore, the Bayes decision rule is given by

$$\begin{aligned} r^* &= \begin{cases} C_1 & g(x) > 0 \\ C_2 & \text{otherwise} \end{cases} \\ &= \begin{cases} C_1 & x < \frac{\mu_1 + \mu_2}{2} \\ C_2 & \text{otherwise} \end{cases}. \end{aligned}$$

2.2 Part 2

Proof. First note that the two events are disjoint, since we the classification can not be both C_1 and C_2 at the same time. Therefore,

$$\begin{aligned} P_e &= P\left((C_1 \text{ misclassified as } C_2) \cup (C_2 \text{ misclassified as } C_1)\right) \\ &= P\left(C_1 \text{ misclassified as } C_2\right) + P\left(C_2 \text{ misclassified as } C_1\right). \end{aligned}$$

Expanding this result gives

$$\begin{aligned} P_e\left(b\right) &= P\left(C_1 \text{ misclassified as } C_2\right) + P\left(C_2 \text{ misclassified as } C_1\right) \\ &= P\left(y = C_1 \wedge x = C_2\right) + P\left(y = C_2 \wedge x = C_1\right) \\ &= P\left(y = C_1 \wedge x \geq b\right) + P\left(y = C_2 \wedge x \leq b\right) \\ &= P\left(x \geq b | y = C_1\right) P\left(y = C_1\right) + P\left(x \leq b | y = C_2\right) P\left(y = C_2\right) \\ &= \frac{1}{2}\left(P\left(x \geq b | y = C_1\right) + P\left(x \leq b | y = C_2\right)\right) \\ &= \frac{1}{2}\left(\int_b^\infty \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu_1)^2}{2\sigma^2}}\, dx + \int_{-\infty}^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu_2)^2}{2\sigma^2}}\, dx\right) \\ &= \frac{1}{2\sqrt{2\pi}\sigma}\left(\int_{-\infty}^b e^{-\frac{(x-\mu_2)^2}{2\sigma^2}}\, dx + \int_b^\infty e^{-\frac{(x-\mu_1)^2}{2\sigma^2}}\, dx\right). \end{aligned}$$

□

2.3 Part 3

To find the optimal decision boundary b^*, take the derivative $\frac{d}{db}P_e(b)$ and find the value b which sets this derivative to 0. Using the first fundamental rule of calculus to take this derivative gives

$$\begin{aligned}\frac{d}{db}P_e(b) &= \frac{d}{db}\left(\frac{1}{2\sqrt{2\pi}\sigma}\left(\int_{-\infty}^{b} e^{-\frac{(x-\mu_2)^2}{2\sigma^2}}\,dx + \int_{b}^{\infty} e^{-\frac{(x-\mu_1)^2}{2\sigma^2}}\,dx\right)\right)\\ &= \frac{1}{2\sqrt{2\pi}\sigma}\left(e^{-\frac{(b-\mu_2)^2}{2\sigma^2}} - e^{-\frac{(b-\mu_1)^2}{2\sigma^2}}\right)\\ &= 0.\end{aligned}$$

Solving this equation gives

$$\begin{aligned}e^{-\frac{(b-\mu_2)^2}{2\sigma^2}} &= e^{-\frac{(b-\mu_1)^2}{2\sigma^2}}\\ \therefore (b-\mu_2)^2 &= (b-\mu_1)^2\\ \therefore 2b(\mu_1-\mu_2) &= \mu_1^2-\mu_2^2\\ \therefore b^* &= \frac{\mu_1^2-\mu_2^2}{2(\mu_1-\mu_2)}\\ &= \frac{\mu_1+\mu_2}{2}.\end{aligned}$$

This result matches the Bayes optimal decision boundary we found in part (1).

3 Isocontours of Normal Distributions

Entire section is included in attached Python Notebook.

4 Eigenvectors of the Gaussian Covariance Matrix

Entire section is included in attached Python Notebook.

5 Classification and Risk

5.1 Part 1

Proof. First, consider the expected loss of a single data point using the predictor offered. Let class i be the class which maximizes the conditional posterior probability. Then, if $P(Y = i|x) < \frac{\lambda_r}{\lambda_s}$, the first condition does not hold, so we choose to doubt and receive a loss of λ_r. Otherwise, the first condition holds, so we choose class i and receive an expected loss of

$$\begin{aligned} E\left[L\left(r\left(x\right)=i,y=j\right)|\, P\left(Y=i|x\right) \geq 1-\frac{\lambda_r}{\lambda_s}\right] &= P\left(Y=i|x\right)\cdot 0 + P\left(Y\neq i|x\right)\cdot\lambda_s \\ &= P\left(Y\neq i|x\right)\cdot\lambda_s \\ &\leq \frac{\lambda_r}{\lambda_s}\lambda_s \\ &\leq \lambda_r. \end{aligned}$$

Therefore, the predictor we use always returns a class which has an expected loss value that is less than or equal to λ_r for each data point. This would only minimize risk if $\lambda_r < \lambda_s$, as choosing to never doubt would guarantee an expected loss value that is less than or equal to λ_s for each data point.

□

5.2 Part 2

First, we analyze the possibilities strictly using the definition of the predictor. If $\lambda_r = 0$, then $1 - \frac{\lambda_r}{\lambda_s} = 1$, which means we only guess class i when we are completely certain that the result is from class i. This also follows intuitively, as having a loss value of 0 for doubting would mean that doubting is as valuable(or as detrimental) as guessing the correct answer. Since doubting is a certain result while guessing the answer is not guaranteed to be correct, doubting would always be the choice if they have the same loss value. In fact, the only reason our predictor chooses to guess the answer when completely certain is because we use $P(Y = i|x) \geq \frac{\lambda_r}{\lambda_s}$ rather than $P(Y = i|x) > \frac{\lambda_r}{\lambda_s}$, which would be a more intuitive solution.

If $\lambda_r > \lambda_s$, then our predictor wouldn't activate, as the predictor requires $\lambda_r \leq \lambda_s$. Intuitively, this means that we would never doubt, because doubting would always give us a loss greater than what would be received if we had guessed. This is obvious if we would have guessed the right answer, as a correct answer would give us 0 loss. However, if we guess the wrong answer, we still receive less loss than doubting, as $\lambda_r > \lambda_s$ by definition.

6 Maximum Likelihood Estimation and Bias

6.1 Part 1

The likelihood function and log likelihood functions are given by

$$\begin{aligned} L(\mu, \sigma; X_1, ..., X_n) &= \Pi f_i(X_i) \\ \therefore l(\mu, \sigma; X_1, ..., X_n) &= \Sigma \ln(f_i(X_i)) \\ &= \Sigma \left(-\frac{i\|X_i - \mu\|^2}{2\sigma^2} - \ln\sqrt{2\pi} - \ln\frac{\sigma}{\sqrt{i}} \right) \\ &= \Sigma \left(-\frac{i\|X_i - \mu\|^2}{2\sigma^2} - \ln\sqrt{2\pi} - \ln\sigma - \ln\sqrt{i} \right). \end{aligned}$$

Since logarithm is a monotonically increasing function of its argument, the parameters which maximize the log likelihood function is equivalent to the parameters which maximize the likelihood function. Therefore, we can first take the gradient with respect to the mean and set it to zero, giving us

$$\begin{aligned} \nabla_\mu l &= \Sigma \frac{i(X_i - \mu)}{\sigma^2} \\ &= 0 \\ \therefore \hat{\mu} &= \frac{1}{n}\Sigma X_i. \end{aligned}$$

Now, we can take the partial of the log likelihood function with respect to σ and set it to zero, giving us

$$\begin{aligned} \frac{\partial l}{\partial \sigma} &= \Sigma \left(\frac{i \cdot \|X_i - \hat{\mu}\|^2 - \sigma^2}{\sigma^3} \right) \\ &= 0 \\ \therefore n\hat{\sigma}^2 &= \Sigma \left(i \cdot \|X_i - \hat{\mu}\|^2 \right) \\ \therefore \hat{\sigma}^2 &= \frac{1}{n}\Sigma \left(i \cdot \|X_i - \hat{\mu}\|^2 \right). \end{aligned}$$

6.2 Part 2

Proof.

$$\begin{aligned} E[\hat{\mu}] &= E\left[\frac{1}{n}\Sigma X_i\right] \\ &= \frac{1}{n}\Sigma E[X_i] \\ &= \frac{1}{n}\Sigma\mu \\ &= \frac{1}{n}n\mu \\ &= \mu. \end{aligned}$$

Therefore, since the expected value of the mean estimator $\hat{\mu}$ is equal to the mean, the mean estimator is unbiased.

□

6.3 Part 3

Proof.

$$\begin{aligned} E\left[\hat{\sigma}^2\right] &= E\left[\frac{1}{n}\Sigma\left(i \cdot \|X_i - \hat{\mu}\|^2\right)\right] \\ &= \frac{1}{n}\Sigma\left(i \cdot E\left[\left(X_i - \hat{\mu}\right)^2\right]\right) \\ &= \frac{1}{n}\Sigma\left(i \cdot E\left[X_i^2 - 2\hat{\mu}X_i + \hat{\mu}^2\right]\right) \\ &= \frac{1}{n}\Sigma\left(i \cdot E\left[X_i^2 - \hat{\mu}^2\right]\right) \\ &= \frac{1}{n}\left(\Sigma\left(i \cdot E\left[X_i^2\right]\right) - \Sigma\left(i \cdot E\left[\hat{\mu}^2\right]\right)\right). \end{aligned}$$

First, we use the equality $E\left[X^2\right] = Var\left(X\right) + E\left[X\right]^2$ to find the value of $\Sigma\left(i \cdot E\left[X_i^2\right]\right)$, which gives us

$$\begin{aligned} \Sigma\left(i \cdot E\left[X_i^2\right]\right) &= \Sigma\left(i \cdot \left(Var\left(X_i\right) + E\left[X_i\right]^2\right)\right) \\ &= \Sigma\left(i \cdot \left(\sigma_i^2 + \hat{\mu}^2\right)\right) \\ &= \Sigma\left(i \cdot \left(\frac{\sigma^2}{i} + \hat{\mu}^2\right)\right) \\ &= \Sigma\left(\sigma^2 + i\hat{\mu}^2\right) \\ &= n\sigma^2 + \left(\frac{n\left(n+1\right)}{2}\right)\hat{\mu}^2. \end{aligned}$$

Similarly, we can expand $\Sigma\left(i \cdot E\left[\hat{\mu}^2\right]\right)$, which gives us

$$
\begin{aligned}
\Sigma\left(i \cdot E\left[\hat{\mu}^2\right]\right) &= \Sigma\left(i \cdot \left(Var\left(\hat{\mu}\right) + E\left[\hat{\mu}\right]^2\right)\right) \\
&= \Sigma\left(i \cdot \left(Var\left(\hat{\mu}\right) + \hat{\mu}^2\right)\right) \\
&= \Sigma\left(i \cdot \left(Var\left(\frac{1}{n}\Sigma X_i\right) + \hat{\mu}^2\right)\right) \\
&= \Sigma\left(i \cdot \left(\frac{1}{n^2}Var\left(\Sigma X_i\right) + \hat{\mu}^2\right)\right) \\
&= \Sigma\left(i \cdot \left(\frac{1}{n^2}\Sigma Var\left(X_i\right) + \hat{\mu}^2\right)\right) \\
&= \Sigma\left(i \cdot \left(\frac{1}{n^2}\Sigma\sigma_i^2 + \hat{\mu}^2\right)\right) \\
&= \frac{n\left(n+1\right)}{2}\left(\frac{1}{n^2}\Sigma\sigma_i^2 + \hat{\mu}^2\right) \\
&\geq \frac{n\left(n+1\right)}{2}\left(\frac{1}{n^2}\Sigma\frac{\sigma^2}{n} + \hat{\mu}^2\right) \\
&\geq \frac{n\left(n+1\right)}{2}\left(\frac{\sigma^2}{n^2} + \hat{\mu}^2\right) \\
&> \frac{\sigma^2}{2} + \left(\frac{n\left(n+1\right)}{2}\right)\hat{\mu}^2.
\end{aligned}
$$

Plugging this result back into our original equation gives

$$
\begin{aligned}
E\left[\hat{\sigma}^2\right] &= \frac{1}{n}\left(\Sigma\left(i \cdot E\left[X_i^2\right]\right) - \Sigma\left(i \cdot E\left[\hat{\mu}^2\right]\right)\right) \\
&< \frac{1}{n}\left(n\sigma^2 + \left(\frac{n\left(n+1\right)}{2}\right)\hat{\mu}^2 - \frac{\sigma^2}{2} - \left(\frac{n\left(n+1\right)}{2}\right)\hat{\mu}^2\right) \\
&< \frac{1}{n}\left(\left(n - \frac{1}{2}\right)\sigma^2\right) \\
&< \frac{1}{n}\left(n\sigma^2\right) \\
&< \sigma^2.
\end{aligned}
$$

Therefore, since the expected value of our variance estimator is strictly less than the actual variance, our estimator is biased.

□

7 Covariance Matrices and Decompositions

7.1 Part 1

$\hat{\Sigma}$ is not invertible when it is a singular matrix. This can happen occur when any of the points X_i is exactly equal to the mean vector. If there exists such a point, then the variance of this point would be zero and the covariance of this point with any other point would be zero. Therefore, there would be an empty row and an empty column, which would make the matrix non-invertible. A similar result would occur if there exists a linear combination of the points such that the variance of these points is zero. This would also cause, through Gaussian elimination, one of the rows and columns to be zero.

7.2 Part 2

Suggestion: add the identity matrix to our estimate to get $\hat{\Sigma}' = \hat{\Sigma} + I$.

First, recall that the covariance matrix is positive semi-definite, which implies $x^T \hat{\Sigma} x \geq 0$ for any vector x. By adding the identity matrix, we get

$$\begin{aligned} x^T \hat{\Sigma}' x &= x^T \left(\hat{\Sigma} + I \right) x \\ &= x^T \hat{\Sigma} x + x^T I x \\ &\geq 0 + x^T I x \\ &> 0 \end{aligned}$$

for any non-zero vector x. Therefore, our modified covariance matrix $\hat{\Sigma}'$ is positive definite, so it is guaranteed to be invertible.

7.3 Part 3

First, let the eigendecomposition of $\hat{\Sigma}^{'}$ be given by

$$\hat{\Sigma}^{'} = V^T \Lambda V,$$

where V is the orthonormal matrix of eigenvectors and Λ is the diagonal matrix with its elements along the diagonal being the eigenvalues. Now, consider the multivariate normal PDF

$$f(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} e^{-\frac{1}{2} x^T \Sigma^{-1} x}.$$

Holding everything but the vector x constant, maximizing $f(x)$ is equivalent to minimizing $\Sigma^{-1}x$. We first expand this to get

$$\begin{aligned} x^T \Sigma^{-1} x &= x^T V^T \Lambda V^{-1} x \\ &= x^T V^T \Lambda^{-1} V x \\ &= (Vx)^T \Lambda^{-1} (Vx). \end{aligned}$$

Since x is defined to have length 1, we can minimize this value by choosing x such that the smallest diagonal element in Λ^{-1} is multiplied by 1 and every other diagonal is multiplied by 0. Let the smallest diagonal element be in row i. Then, we would set x to be equal to the ith column of V, as V is an orthonormal matrix, so Vx would give a vector filled with 0s, except for a 1 in the ith spot. Finally, we can conclude that this would give the minimum possible value, as every value must be a linear combination of the diagonal elements of Λ^{-1} with total weight equal to 1.

A similar approach is used for finding the vector x which would minimize $f(x)$, except x would be the eigenvector corresponding to the largest element in Λ^{-1} instead of the smallest.

Finally, we can conclude that this result gives the same vector x which minimizes the PDF of $\hat{\Sigma}$, as the eigenvalues of $\hat{\Sigma}^{'}$ are equal to 1 plus the eigenvalues of $\hat{\Sigma}$. Therefore, since this is a monotonically increasing function of its argument, optimizing the PDF of $\hat{\Sigma}^{'}$ by finding the optimal vector x is equivalent to optimizing the PDF of $\hat{\Sigma}$.

7.4 Part 4

$$\begin{aligned} Var\left(p\right) &= Var\left(y^T X\right) \\ &= E\left[\left(y^T X\right)\left(y^T X\right)^T\right] \\ &= E\left[y^T X X^T y\right] \\ &= y^T E\left[X X^T\right] y \\ &= y^T Var\left(X\right) y \\ &= y^T \Sigma y. \end{aligned}$$

This becomes the optimization problem we had in part (3). First note that the largest eigenvalue of Σ corresponds to the smallest eigenvalue of Σ^{-1}, as $\lambda_\Sigma = \frac{1}{\lambda_{\Sigma^{-1}}}$. Combining this with the proof that the eigenvector corresponding to the smallest eigenvalue of Σ^{-1} maximizes the PDF gives the result that the eigenvector corresponding to the largest eigenvalue of Σ would maximize the PDF $f\left(x\right)$.

8 Gaussian Classifiers for Digits and Spam

Entire section is included in attached Python Notebook.