School of Electronics and Computer Science
University of Southampton

COMP6245(2021/22): Foundations of Machine Learning (MSc) Lab Three

Issue	22/10/2021
Deadline	3/11/2021 (10:00 AM)

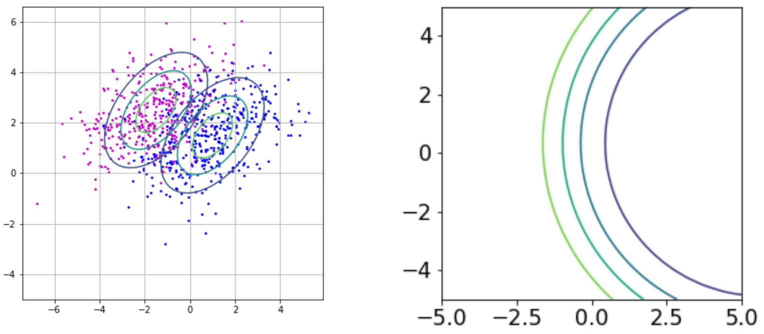
Objective

1. Class boundaries and posterior probabilities of Gaussian classifiers
2. Fisher Linear Discriminant Analysis
3. Receiver Operating Characteristic (ROC) Curve
4. Mahalanobis distance

1 Class Boundaries and Posterior Probabilities

Consider two-class classification problems in two dimensions in which features of the two classes are Gaussian distributed:

- $\boldsymbol{m}_1 = \begin{pmatrix} 0 \\ 3 \end{pmatrix}$ and $\boldsymbol{m}_1 = \begin{pmatrix} 3 \\ 2.5 \end{pmatrix}$, $\boldsymbol{C}_1 = \boldsymbol{C}_2 = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$, $P_1 = P2 = 0.5$;
- $\boldsymbol{m}_1 = \begin{pmatrix} 0 \\ 3 \end{pmatrix}$ and $\boldsymbol{m}_2 = \begin{pmatrix} 3 \\ 2.5 \end{pmatrix}$, $\boldsymbol{C}_1 = \boldsymbol{C}_2 = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$, $P_1 = 0.7,\ P_2 = 0.3$;
- $\boldsymbol{m}_1 = \begin{pmatrix} 0 \\ 3 \end{pmatrix}$ and $\boldsymbol{m}_2 = \begin{pmatrix} 3 \\ 2.5 \end{pmatrix}$, $\boldsymbol{C}_1 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$, $\boldsymbol{C}_2 = \begin{pmatrix} 1.5 & 0 \\ 0 & 1.5 \end{pmatrix}$, $P_1 = P2 = 0.5$.



(a) (b)

Figure 1: (a) Example of probability densities and data; (b) Example of contours on the posterior probability $P[\omega_1 \mid \boldsymbol{x}]$. Note these are illustrations and (b) is not the posterior probability for the problem in (a)!

In each of the above cases, plot contours on the likelihoods of the two classes, a scatter of 200 data sampled from each of the classes and contours on the posterior probability of one of the classes. Discuss (in your report) if what you plot is consistent with your expectation from analytical derivation of the class boundaries. Note you are free to change the means and covariance matrices given above to illustrate the differences we are learning about. Examples of the graphs to aim for are given in Fig. 1

2 Fisher LDA and ROC Curve

Define a two class pattern classification problem in two dimensions, in which the two classes are Gaussian distributed with means $\boldsymbol{m}_1 = [0 \;\; 3]^t$ and $\boldsymbol{m}_1 = [3 \;\; 2.5]^t$, and have a common covariance matrix

$$\boldsymbol{C}_1 = \boldsymbol{C}_2 = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

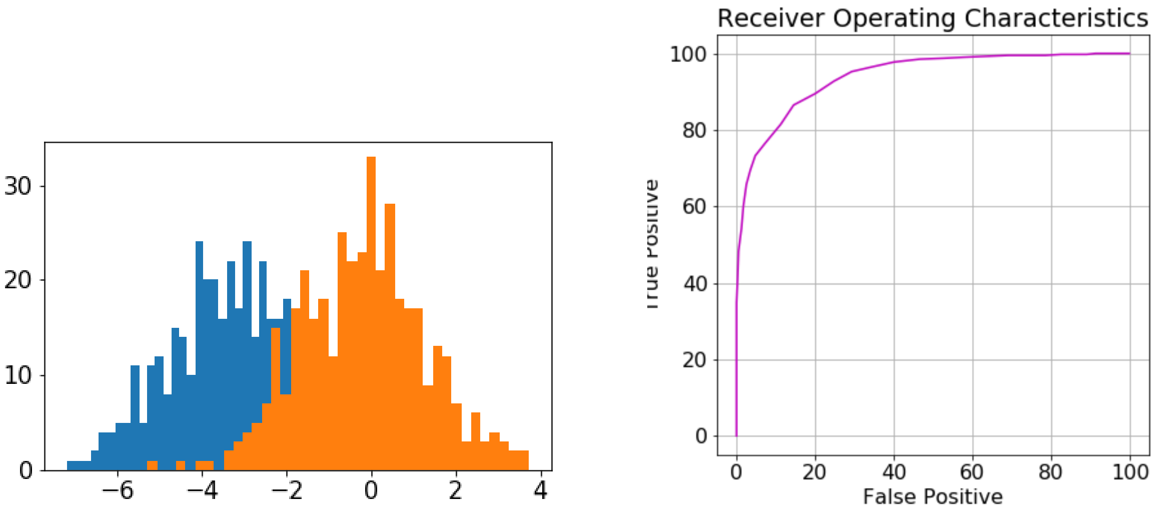
1. Plot contours on the two densities.
2. Draw 200 samples from each of the two distributions and plot them on top of the contours.
3. Compute the Fisher Linear Discriminant direction using the means and covariance matrices of the problem, and plot the discriminant direction: $\boldsymbol{w}_F = (\boldsymbol{C}_1 + \boldsymbol{C}_2)^{-1} (\boldsymbol{m}_1 - \boldsymbol{m}_2)$
4. Project the data onto the Fisher discriminant directions and plot histograms of the distribution of projections (an example of this is in Fig. 2(a);
5. Compute and plot the Receiver Operating Characteristic (ROC) curve, by sliding a decision threshold, and computing the True Positive and False Positive rates (see code snippet in Appendix and example of an ROC curve in Fig. 2(b).
6. Compute the area under the ROC curve (Hint: try `numpy.trapz`)
7. For a suitable choice of decision threshold, compute the classification accuracy.
8. Plot the ROC curve (on the same scale) for

   - A random direction (instead of the Fisher discriminant direction).
   - Projections onto the direction connecting the means of the two classes.

   Compute the area under the ROC curve (AUC) for these two cases. Your report should explain what the precise statistical interpretation of AUC is and when it is used.

3 Mahalanobis Distance

Using a suitable classification problem (two-class in two dimensions, adapted from one of the above examples), illustrate the difference between a distance-to-mean classifier and a Mahalanobis distance-to-mean classifier. Your report should give a clear and succinct description of the differences.



(a)

(b)

Figure 2: (a) Histograms of projections of the two classes onto the Fisher discriminant direction; (b) An ROC Curve

Appendix: Snippets of code

1. To compute the posterior probability (Note: `gauss2D` was used in previous Labs.)

```
def posteriorPlot(nx, ny, m1, C1, m2, C2, P1, P2):
    x = np.linspace(-5, 5, nx)
    y = np.linspace(-5, 5, ny)
    X, Y = np.meshgrid(x, y, indexing=’ij’)

    Z = np.zeros([nx, ny])
    for i in range(nx):
        for j in range(ny):
            xvec = np.array([X[i,j], Y[i,j]])
            num = P1*gauss2D(xvec, m1, C1)
            den = P1*gauss2D(xvec, m1, C1) + P2*gauss2D(xvec, m2, C2)
            Z[i,j] = num / den
    return X, Y, Z
```

2. To compute the Fisher discriminant direction, project data and plot histograms of the two projected classes:

```
Ci = np.linalg.inv(2*C)
uF = Ci @ (m2-m1)

yp1 = Y1 @ uF
yp2 = Y2 @ uF

matplotlib.rcParams.update({’font.size’: 16})
plt.hist(yp1, bins=40)
plt.hist(yp2, bins=40)
plt.savefig(’histogramprojections.png’)
```

3. To compute and plot a ROC curve:

```
# Define a range over which to slide a threshold
#
pmin = np.min( np.array( (np.min(yp1), np.min(yp2) )))
pmax = np.max( np.array( (np.max(yp1), np.max(yp2) )))
print(pmin, pmax)

# Set up an array of thresholds
#
nRocPoints = 50;
thRange = np.linspace(pmin, pmax, nRocPoints)
ROC = np.zeros( (nRocPoints, 2) )

# Compute True Positives and False positives at each threshold
#
for i in range(len(thRange)):
    thresh = thRange[i]
    TP = len(yp2[yp2 > thresh]) * 100 / len(yp2)
    FP = len(yp1[yp1 > thresh]) * 100 / len(yp1)
    ROC[i,:] = [TP, FP]

# Plot ROC curve
#
fig, ax = plt.subplots(figsize=(6,6))
ax.plot(ROC[:,1], ROC[:,0], c=’m’)
ax.set_xlabel(’False Positive’)
ax.set_ylabel(’True Positive’)
ax.set_title(’Receiver Operating Characteristics’)
ax.grid(True)
plt.savefig(’rocCurve.png’)
```

Report

Describe the work you have done as a short report. Upload a *pdf* file **no longer than four pages** using the ECS handin system: `http://handin.ecs.soton.ac.uk`. Please use LaTeX to typeset your report. Please make sure your name and email are included.