Machine Learning Fundamentals
Lab 2

Shubhm Agarwal
Master 2 MSIAM Data Science
shubhamagarwal92@gmail.com

1 Linear discriminant analysis (LDA)

Theoritical aspect

Let us consider two Gaussian populations in R^p with the same covariance structure. We have observations drawn from a mixture of these two populations. The conditional distributions of X given Y=+1 (respectively Y = -1) are multivariate Gaussian distributions $N_p(\mu_+, \Sigma)$ (respectively $N_p(\mu_-, \Sigma)$). We denote their respective probability density functions f_+ and f_-. The two vectors f_+ and f_+ both belong to R_p and σ is a symmetric matrix. We also denote $\pi = P[Y = +1]$. We recall that the p.d.f. of $N_p(\mu, \Sigma)$ reads :

$P(x) = f(x) = \frac{1}{(2\pi)^{p/2}\sqrt{det(\Sigma)}} exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$

and that the covariance matrix of a random vector X is defined as

$\Sigma = E((X - E(X))(X - E(X))^T)$

1. Use the Bayes formula to compute $P[Y = +1|X = x], P[Y = -1|X = x]$ as function of f_+, f_- and π_+

We have $f(x) = \frac{1}{(2\pi)^{p/2}\sqrt{det(\Sigma)}} exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$

and for the notation we have

$P(x|y = 1) = f_+(x) = \frac{1}{(2\pi)^{p/2}\sqrt{det(\Sigma)}} exp(-\frac{1}{2}(x-\mu_+)^T\Sigma^{-1}(x-\mu_+))$

$P(x|y = -1) = f_-(x) = \frac{1}{(2\pi)^{p/2}\sqrt{det(\Sigma)}} exp(-\frac{1}{2}(x-\mu_-)^T\Sigma^{-1}(x-\mu_-))$

with $\Sigma_+ = \Sigma_- = \Sigma$. Also as it is mixture of two Gaussians $\pi_- = 1 - \pi_+$

We have from Bayes rule that

$P(y|x) = \frac{\pi_y P(x|y)}{P(x)}$

for the posterior probability $P(y|x)$ of observing an instance of class y at point x. The unconditional probability P(x) in the denominator does not depend on y and we use the definition to finally have

$P(Y = 1|X) = \frac{\pi_+ f_+(x)}{\pi_+ f_+(x) + (1-\pi_+) f_-(x)}$

$P(Y = -1|X) = \frac{(1-\pi_+) f_-(x)}{\pi_+ f_+(x) + (1-\pi_+) f_-(x)}$

2. Express the log-ratio of the two classes :

$log(\frac{P[Y=+1|X=x]}{P[Y=-1|X=x]})$

in function of $\mu_+, \mu_-, \pi_+\Sigma$

$log(\frac{P[Y=+1|X=x]}{P[Y=-1|X=x]}) = log(\frac{\pi_+}{1-\pi_+}) - \frac{1}{2}log(\frac{|\Sigma_+|}{|\Sigma_-|}) + x^T(\Sigma_+^{-1}\mu_+ - \Sigma_-^{-1}\mu_-) - \frac{1}{2}x^T(\Sigma_+^{-1} - \Sigma_-^{-1})x - \frac{1}{2}(\mu_+^T\Sigma_+^{-1}\mu_+ - \mu_-^T\Sigma_-^{-1}\mu_-)$

Here $\Sigma_+ = \Sigma_-$ and we have used $\pi_- = 1 - \pi_+$. This simplifies to:

$log(\frac{P[Y=+1|X=x]}{P[Y=-1|X=x]}) = log(\frac{\pi_+}{1-\pi_+}) + x^T(\Sigma^{-1}\mu_+ - \Sigma^{-1}\mu_-) - \frac{1}{2}(\mu_+^T\Sigma^{-1}\mu_+ - \mu_-^T\Sigma^{-1}\mu_-)$

(The term $-\frac{1}{2}x^T(\Sigma^{-1} - \Sigma^{-1})x$ cancels out.)

3. We have some observations drawn from this mixture and we assume that $\mu_+, \mu_-, \pi_+, \Sigma$ are unknown. We assume that the sample contains n observations $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ and that $\sum_{i=1}^n 1_{\{y_i=+1\}} = m$. Use the moments method to propose parametric estimators of the unknown parameters.

The parameters of our model are $\mu_+, \mu_-, \pi_+, \Sigma$. (Note that while there're two different mean vectors μ_+, μ_-, this model is usually applied using only one covariance matrix Σ.) The log-likelihood of the data is given by

$l(\mu_+, \mu_-, \pi_+, \Sigma) = log\Pi_i p(x^{(i)}, y^{(i)}; \mu_+, \mu_-, \pi_+, \Sigma)$

$= log\Pi_i p(x^{(i)}|y^{(i)}; \mu_+, \mu_-, \Sigma)p(y^{(i)}; \pi_+)$

By maximizing l with respect to the parameters, we find the maximum likelihood estimate of the parameters to be:

$\pi_+ = \frac{1}{n}\sum_{i=1}^n 1_{\{y_i=+1\}} = \frac{m}{n}$

$\mu_+ = \frac{\sum_{i=1}^n 1_{\{y_i=+1\}}x^{(i)}}{\sum_{i=1}^n 1_{\{y_i=+1\}}} = \frac{\sum_{i=1}^n 1_{\{y_i=+1\}}x^{(i)}}{m}$

$\mu_- = \frac{\sum_{i=1}^n 1_{\{y_i=-1\}}x^{(i)}}{\sum_{i=1}^n 1_{\{y_i=-1\}}} = \frac{\sum_{i=1}^n 1_{\{y_i=-1\}}x^{(i)}}{1-m}$

$\Sigma = \frac{1}{n}\sum_{i=1}^n (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T$

4. Justify the following choice of the classifer

$$\begin{cases} 1 & \text{if } x^T\hat{\Sigma}^{-1}(\widehat{mu_+} - \hat{\mu}_-) > \frac{1}{2}\widehat{mu_+}\hat{\Sigma}^{-1}\widehat{mu_+} - \frac{1}{2}\widehat{mu_-}\hat{\Sigma}^{-1}\widehat{mu_-} + log(1-m/n) - log(m/n) \\ -1 & \text{otherwise} \end{cases}$$

The decision boundary for this linear classifier is given when $P[Y = +1|X = x] = P[Y = -1|X = x]$. Taking log on both sides and using terms of Question 2, we have

$log(\frac{P[Y=+1|X=x]}{P[Y=-1|X=x]}) = 0$

$log(\frac{P[Y=+1|X=x]}{P[Y=-1|X=x]}) = log(\frac{\pi_+}{1-\pi_+}) + x^T(\Sigma^{-1}\mu_+ - \Sigma^{-1}\mu_-) - \frac{1}{2}(\mu_+^T\Sigma^{-1}\mu_+ - \mu_-^T\Sigma^{-1}\mu_-) = 0$

We also know that number of positive observations are m (i.e $\sum_{i=1}^n 1_{\{y_i=+1\}} = m$). Thus $log(\pi_+) = log(m/n)$ and $log(\pi_-) = log(1 - m/n)$.

We classify as class 1 when $P[Y = +1|X = x] > P[Y = -1|X = x]$ or $log(\frac{P[Y=+1|X=x]}{P[Y=-1|X=x]}) > 0$ and class -1 otherwise.

Thus, we get

$x^T(\Sigma^{-1}\mu_+ - \Sigma^{-1}\mu_-) > \frac{1}{2}(\mu_+^T\Sigma^{-1}\mu_+ - \mu_-^T\Sigma^{-1}\mu_-) + log(\pi_-) - log(\pi_+)$ or

$x^T\Sigma^{-1}(\mu_+ - \mu_-) > \frac{1}{2}(\mu_+^T\Sigma^{-1}\mu_+) - \frac{1}{2}(\mu_-^T\Sigma^{-1}\mu_-) + log(1-m/n) - log(m/n)$

as the choice of classifier

5. What happens when the two covariance matrices differ

If we assume that each class has its own correlation structure, the discriminant functions are no longer linear. Instead, we get them as

$\delta_k(x) = -\frac{1}{2}log|\Sigma_k|^{-1} - \frac{1}{2}(x - \mu_k)^T\Sigma_k^{-1}(x - \mu_k)$

The decision boundary is now described with a quadratic function. This is therefore called quadratic discriminant analysis (QDA).

6. How can we generalize the linear discriminant analysis to the multiclass setting?

For multiclass linear discriminant analysis we have Mixtures of Gaussians with Gaussian densities. The prior probability of class k is π_k, $\sum_{k=1}^{K} \pi_k = 1$

π_k is usually estimated simply by empirical frequencies of the training set

$\hat{\pi_k} = \frac{\#SamplesInClassK}{Total\#ofsamples}$

The class-conditional density of X in class G = k is $f_k(x)$

$f_k(x) = \frac{1}{(2\pi)^{p/2}\sqrt{det(\Sigma_k)}} exp(-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k))$

For linear discriminant analysis (LDA) $\Sigma_k = \Sigma \; \forall k$.

The Gaussian distributions are shifted versions of each other.

We compute the posterior probability as

$P(G = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$

By MAP, we get

$\hat{G}(x) = argmax_k P(G = k|X = x) = argmax_k \pi_k f_k(x)$

LDA in practice

We now apply LDA on synthetic data and thereafter to real data. In this last case, we split randomly the dataset into two parts : a training set (around 70% of the data) and a validation set (the 30% remaining).

1. Import sklearn package
from sklearn.lda import LDA

2. Create a LDA model
my_lda = LDA()

3. Learn the model from the data dataX and their corresponding labels dataY
my_lda.fit(dataX,dataY)

4. Apply the LDA on the mixture generated by the function rand bi gauss. Estimate the prediction error using the test sample

Accuracy recorded for rand bi gauss: 97.49

```
LDA coef:[[ 2.65934084  2.81636903]]
LDA intercept:[ 0.52482839]
LDA results:
             precision    recall  f1-score   support

       -1.0       0.95      0.95      0.95        20
        1.0       0.95      0.95      0.95        20

avg / total       0.95      0.95      0.95        40
```

Figure 1: Results for using LDA on data generated by rand bi gauss

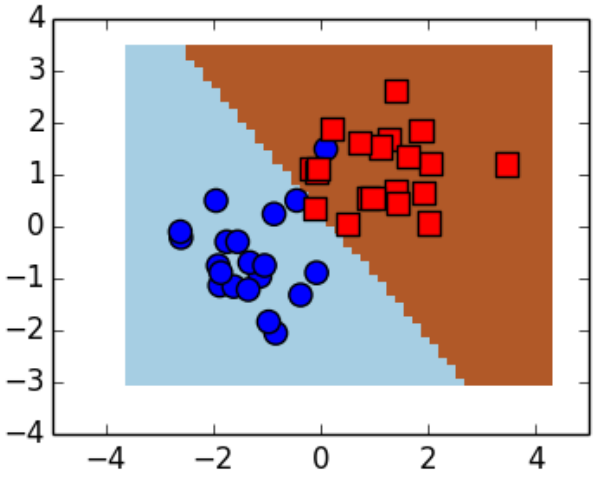

Figure 2: Decision boundary for rand bi gauss function

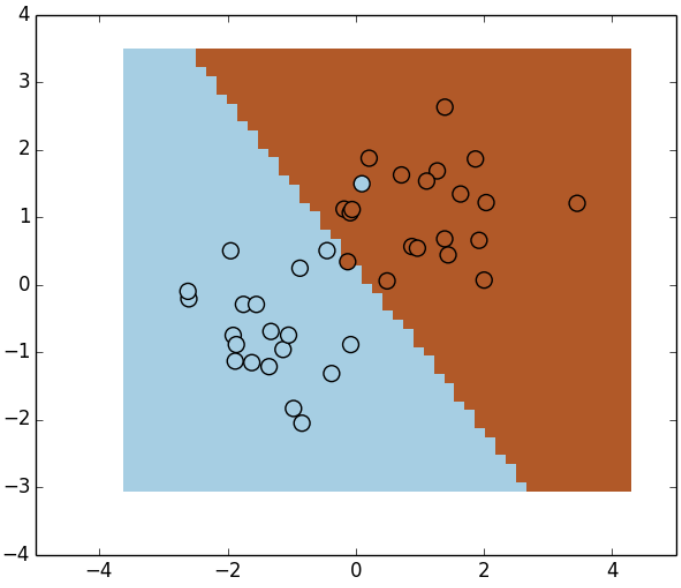

Figure 3: Another visualization of decision boundary for rand bi gauss function