COMP3206: Coursework

November 12, 2017

1 Due date

The hand-in date for the assignment is Wednesday, **December ~~12~~ 13, 2017**.

The correction is because the Hand-in day for Year III courseworks should be Wednesday.

2 Introduction

There are two sections in the coursework. You will use the same data sets for both. In the first you will explore how Fisher's Linear Discriminant Analysis seeks to separate classes. In the second you will perform logistic regression on the data to achieve the same tasks. Section 4.2 may end up being a bit more involved and challenging.

3 Fisher's Linear Discriminant Analysis (LDA)

[6+6 marks]

You will first explore Fisher's LDA for binary classification for class labels a and b. In Fisher's method a direction defined by vector $\boldsymbol{w} = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$ is chosen so that the data x^n projected on to $\boldsymbol{w}$ maximises the separation between the a and b type distributions. (n is a data index, ranging from 1 to n_a for class a, and from 1 to n_b for class b.) Setting $y_c^n \triangleq \boldsymbol{w} \cdot x_c^n$ for class label $c = a, b$ the scalar means and standard deviations of the projected data become

$$\mu_c = \frac{1}{n_c} \sum_{n=1}^{n_c} y_c^n, \ \sigma_c^2 = \frac{1}{n_c} \sum_{n=1}^{n_c} (y_c^n - \mu_c)^2, \ c = a, b.$$

The direction $\boldsymbol{w}$ is chosen in order to maximise the Fisher ratio $F(\boldsymbol{w})$:

$$F(\boldsymbol{w}) \triangleq \frac{|\mu_a - \mu_b|}{\frac{n_a}{n_a + n_b}\sigma_a^2 + \frac{n_b}{n_a + n_b}\sigma_b^2}.$$

3.1 Data 1: separate 2 Gaussians

1. Generate data from two 2-dimensional Gaussian distributions

$$x_a \sim \mathcal{N}(x|m_a, S_a), \ x_b \sim \mathcal{N}(x|m_b, S_b)$$

   where m_a, m_b are the (2×1) mean vectors and S_a and S_b are the (2×2) variance-covariance matrices that define normal distributions. Let the number of data points from each type a, b be n_a and n_b. You will have to choose appropriate numerical values for n_a, n_b, the means m_a, m_b and variances S_a, S_b so that the differences of the means is comparable to the standard deviations. The precise values will affect the results of the following tasks and you can experiment with numerical values that help you appreciate the pros and cons of the classification method.

2. Make a few illustrative choices for the direction w and plot the histograms of the values y_a^n and y_b^n.

3. Plot the dependence of $F(w)$ on the direction of w. Find the minimum value of $F(w)$ and the corresponding direction w^*:

$$w^* = \underset{w}{\operatorname{argmin}} F(w)$$

4. Since the generating distributions for the classes are known, plot the equi-probable contour lines for each class and draw the direction of the optimal choice vector w.

5. Use Bayes' theorem to write down the logarithm of the ratio of conditional probabilities

$$\ln\left(\frac{P(c = a|x^n)}{P(c = b|x^n)}\right)$$

   and plot the decision boundary where this quantity vanishes.

6. Write up your observations in no more than 2 pages, providing evidence based on your experiments, and discuss what you have learned. You should read Section 4.4 of [3] (or Section 4.3 of [2]) for guidance and ideas.

3.2 Data 2: Iris data

In this section you will perform the same LDA task on the famous Iris dataset `https://en.wikipedia.org/wiki/Iris_flower_data_set` which can be downloaded from `http://archive.ics.uci.edu/ml/datasets/Iris`.

With more features, you will need to compute the between-class and within-class variance-covariance matrices Σ_B and Σ_W:

$$\Sigma_B = \sum_c (\mu_c - \mu)(\mu_c - \mu)^T, \text{where } \mu \text{ is the mean of class means,}$$

and Σ_W is the sum of the covariance matrices for each class. In case there are different numbers of training data points from each class, you have to scale any class dependence by the corresponding fraction of class members in the population. (In the Iris data set all three classes have 50 members, so you can skip this step.)

(a) Find the optimal direction w^* for projecting the data onto. You will need to solve the generalised eigenvalue problem $\Sigma_B w = \lambda \Sigma_W w$. Section 16.2, 16.3 of [1] and eq. (4.15) of [2] has further details.

NB: *The standard libraries in numpy/scipy, MATLAB, Mathematica etc can give you the results. Since the covariance matrix is symmetric, the function you should call in* `numpy/scipy` *is* `eigh`, *and not* `eig`, *although for problems of this scale it won't make a difference and you can eyeball the eigenvalues. You must also check the answer provided by verifying that the generalised eigenvalue condition* $(\Sigma_B - \lambda\Sigma_W)w = 0$ *holds. This will clarify the notational conventions of the software used. Sometimes it is the transpose of the returned matrix of vectors that contains the eigenvectors, so please check it.*

(b) Display the histograms of the three classes in the reduced dimensional space defined by w^*.

(c) Present your results with reflections and evidence in no more than 2 pages.

4 Logistic regression

[5+3 marks]

In this section you can rely on the example in the first lab sheet *Introduction by examples* where the log-odds between two classes was fit to a linear function, and a logistic function applied to the answer.

4.1 Logistic regression on the 2-gaussian example

You will find Chapter 22 of [4] a useful reference to reflect on the connections between maximum likelihood, logistic regression and (when you get to it in the module) k-means clustering (see section 22.2).

1. Fit a linear model to the log-odds of your simulated data generated from from two gaussians and obtain the posterior class probabilities $P(c|\mathbf{x}^n)$ using

   $$\sigma(x, \beta) = \frac{1}{1 + \exp(-\beta x)},$$

   where x is the log-odds (logit). What role does β play in classification?

2. Split your data into test and training sets to measure training and test error for different numbers of samples n_a and n_b in the training set.

3. Write up your results in no more than 2 pages.

4.2 Logistic regression for the Iris dataset

NB:*This section is worth 3 marks. This is here to challenge you.*

In this section you are meant to introduce a log-odds for the multi-class case. The wikipedia page `https://en.wikipedia.org/wiki/Multinomial_logistic_regression` usefully breaks down the problem of registering ratios of probabilities with respect to a chosen reference class. You may choose any one of the three species of flowers as the reference class in the Iris dataset.

1. Perform multi-class classification of Iris dataset using logistic regression.

2. Find a way of comparing the results of logistic regression with those of Fisher's LDA. Even for this tiny dataset, you may want to consider a training-test set split.

3. Write up your methods, experimental evidence and your reflections in no more than 2 pages.

References

[1] D. Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 04-2011 edition, 2011.

[2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2nd edition, 2009.

[3] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014.

[4] David J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.