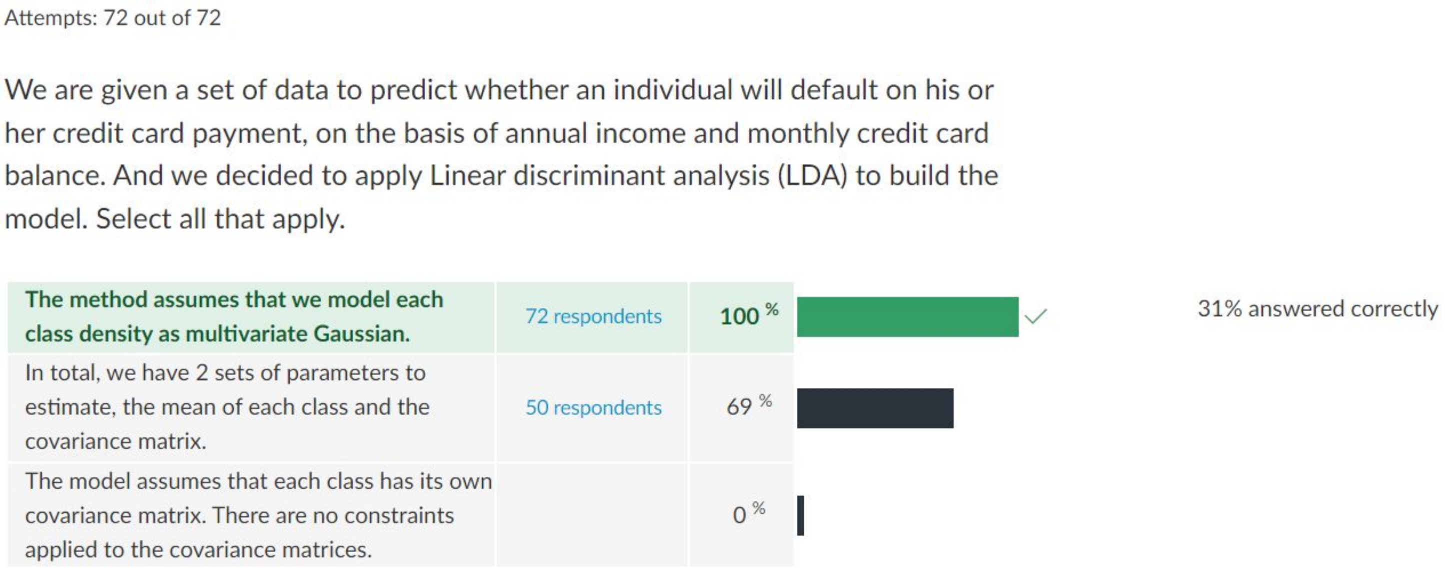
ML TA hours

HW2 Review

2024.10.8

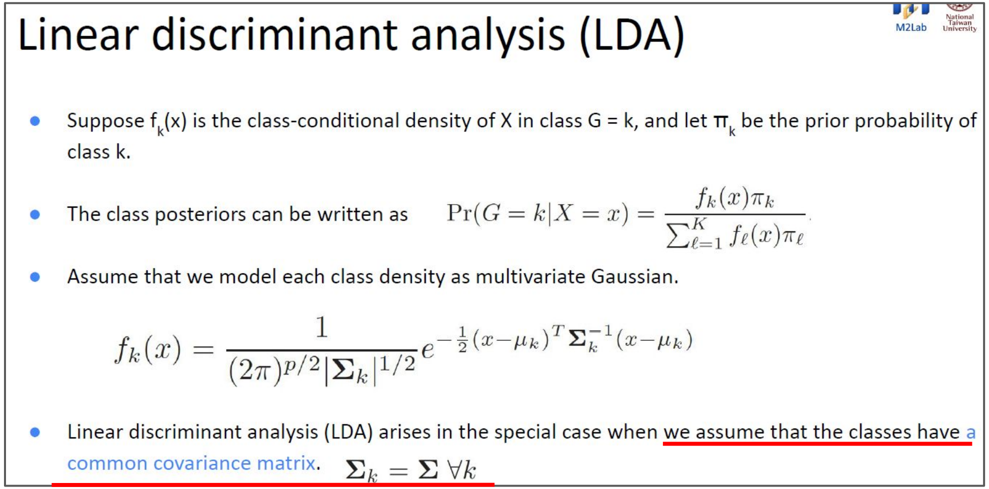


Attempts: 72 out of 72

We are given a set of data to predict whether an individual will default on his or her credit card payment, on the basis of annual income and monthly credit card balance. And we decided to apply Linear discriminant analysis (LDA) to build the model. Select all that apply.

The method assumes that we model each class density as multivariate Gaussian.	72 respondents	100 %	✓
In total, we have 2 sets of parameters to estimate, the mean of each class and the covariance matrix.	50 respondents	69 %	
The model assumes that each class has its own covariance matrix. There are no constraints applied to the covariance matrices.		0 %	

31% answered correctly



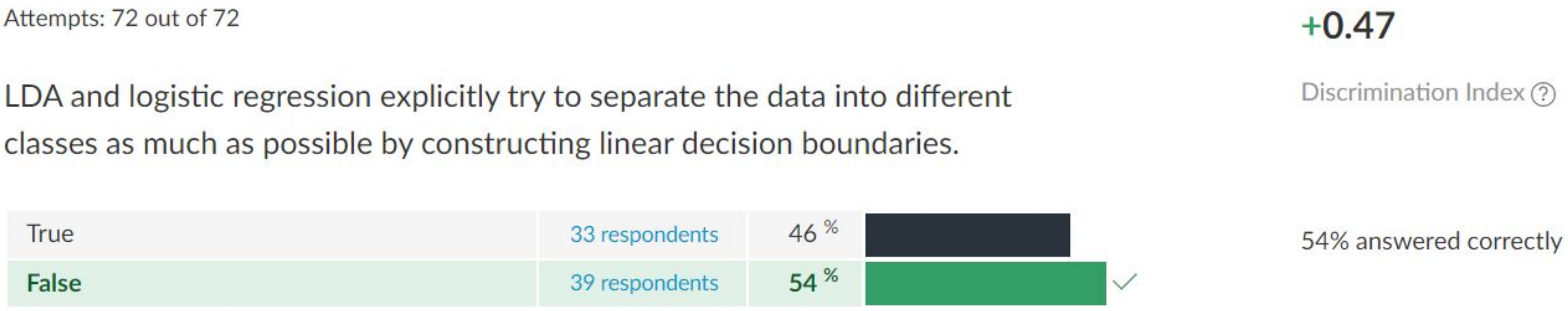
Linear discriminant analysis (LDA)

- Suppose $f_k(x)$ is the class-conditional density of X in class G = k, and let π_k be the prior probability of class k.
- The class posteriors can be written as $\Pr(G=k|X=x) = \frac{f_k(x)\pi_k}{\sum_{\ell=1}^{K} f_\ell(x)\pi_\ell}$.
- Assume that we model each class density as multivariate Gaussian.

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \boldsymbol{\Sigma}_k^{-1}(x-\mu_k)}$$

- Linear discriminant analysis (LDA) arises in the special case when we assume that the classes have a common covariance matrix. $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma} \; \forall k$

week4 p.17



Attempts: 72 out of 72

+0.47

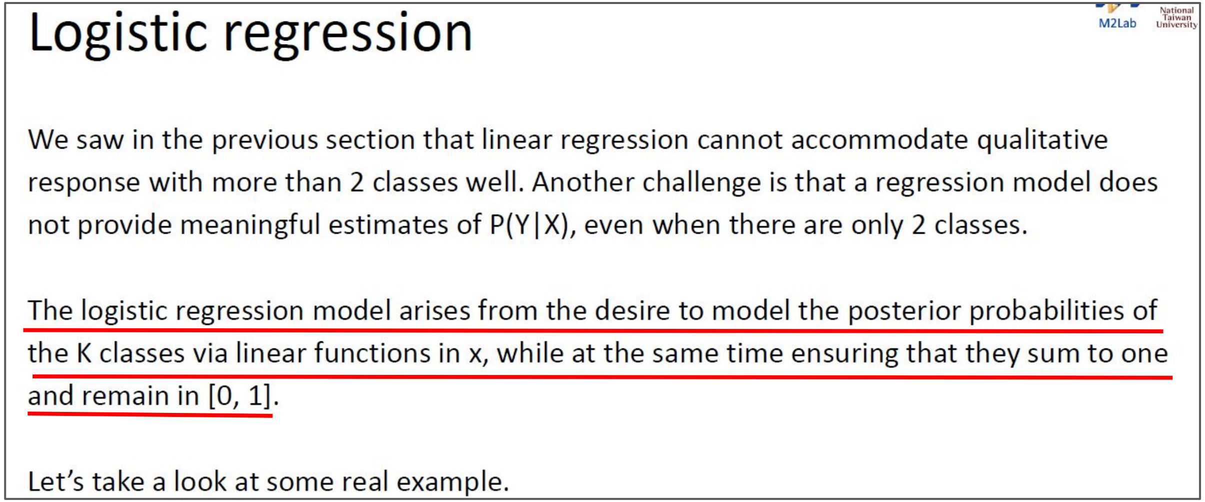
Discrimination Index ?

LDA and logistic regression explicitly try to separate the data into different classes as much as possible by constructing linear decision boundaries.

True	33 respondents	46 %	
False	39 respondents	54 %	✓

54% answered correctly

Logistic Regression can also generate linear decision boundaries under certain circumstances, but its essence and purpose is to estimate the probability that data points belong to a certain category, and does not emphasize separating data points as much as possible.


M2Lab
National Taiwan University

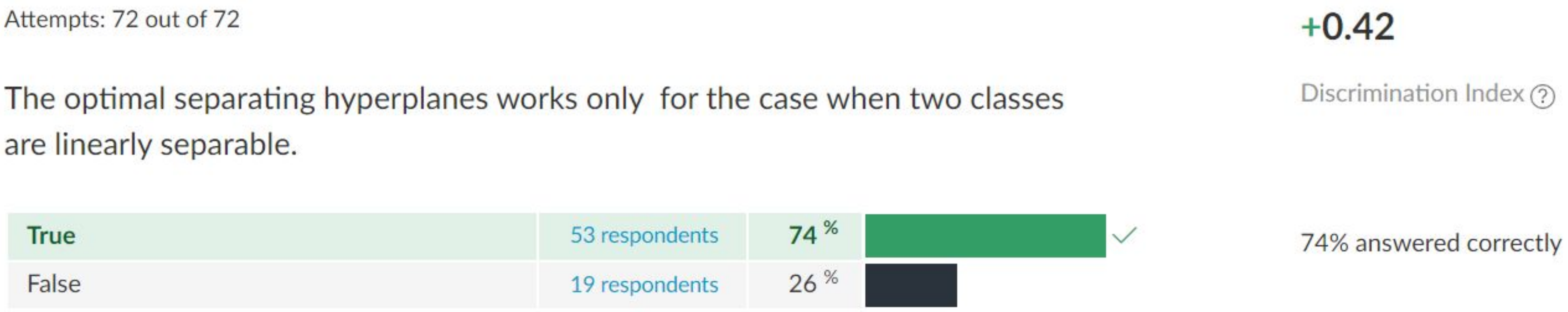
Logistic regression

We saw in the previous section that linear regression cannot accommodate qualitative response with more than 2 classes well. Another challenge is that a regression model does not provide meaningful estimates of P(Y|X), even when there are only 2 classes.

The logistic regression model arises from the desire to model the posterior probabilities of the K classes via linear functions in x, while at the same time ensuring that they sum to one and remain in [0, 1].

Let's take a look at some real example.

week4 p.24



Attempts: 72 out of 72

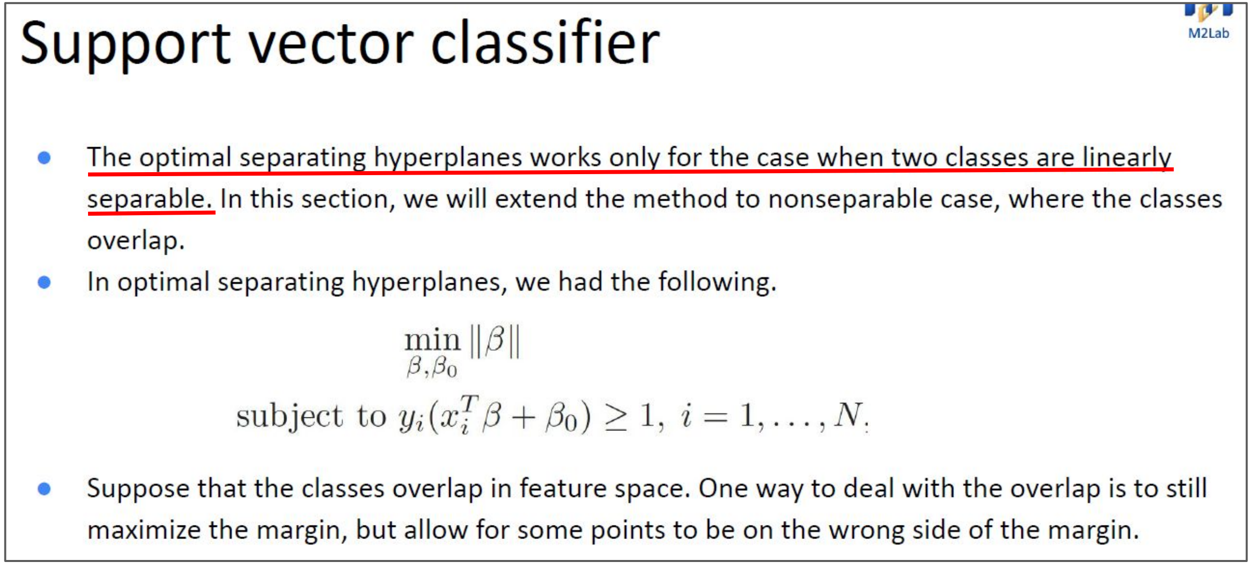
+0.42

Discrimination Index ?

The optimal separating hyperplanes works only for the case when two classes are linearly separable.

True	53 respondents	74 %	🟩 ✓
False	19 respondents	26 %	■

74% answered correctly



Support vector classifier

- The optimal separating hyperplanes works only for the case when two classes are linearly separable. In this section, we will extend the method to nonseparable case, where the classes overlap.
- In optimal separating hyperplanes, we had the following.

$$\min_{\beta,\beta_0} \|\beta\|$$

$$\text{subject to } y_i(x_i^T\beta + \beta_0) \geq 1, \ i = 1, \ldots, N.$$

- Suppose that the classes overlap in feature space. One way to deal with the overlap is to still maximize the margin, but allow for some points to be on the wrong side of the margin.

week4 p.45

ML TA hours

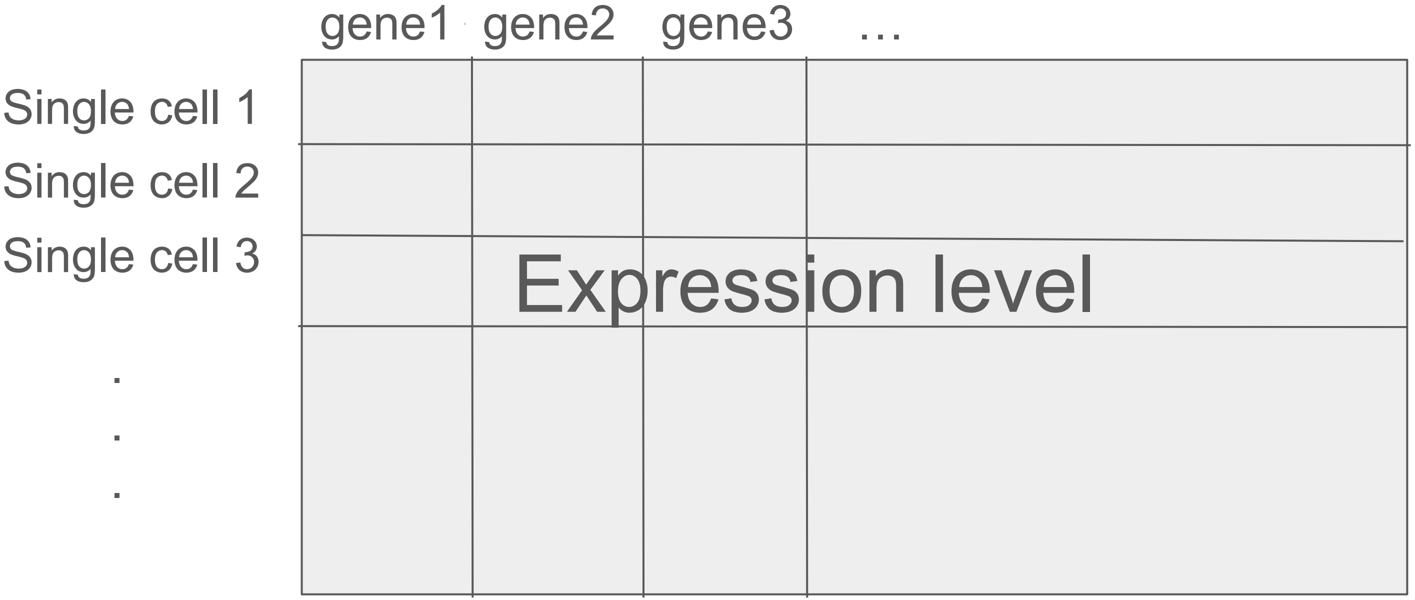
HW4

Colab experiment : Single cell data visualization

2024.10.8

Single-cell sequencing data

- 23822 x 3000 matrix



gene1 gene2 gene3 …

Single cell 1

Single cell 2

Single cell 3

.

.

.

Expression level

Data preprocessing

- First normalize the single-cell sequencing data by the total counts
- And apply the log transformation
- The result of preprocessing is in *logCPM*

TODO - Implement PCA

- Goal : Use dimensionality reduction techniques such as PCA to simplify data dimensions for easier visualization and analysis

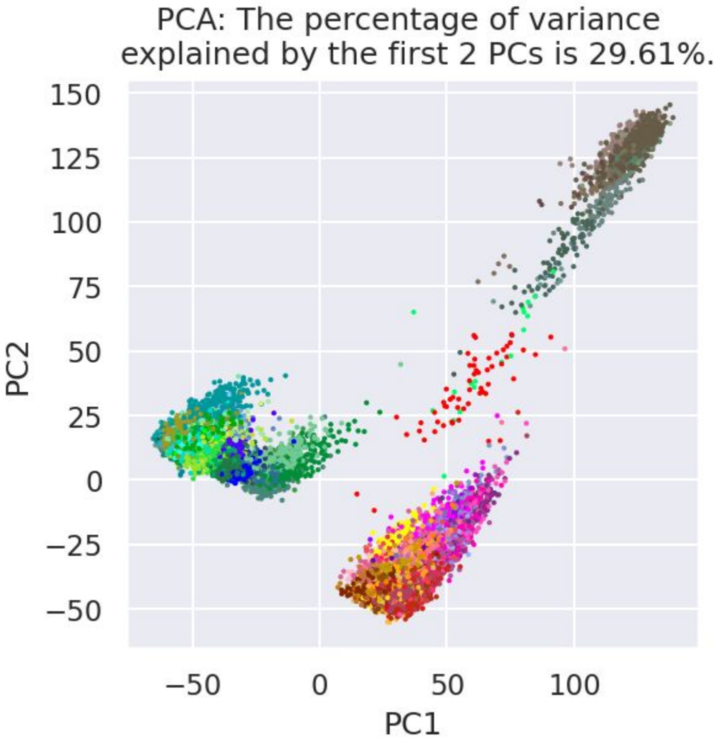
- # TODO
- Reduce the dimension by taking the first **2** principal components
- PCA

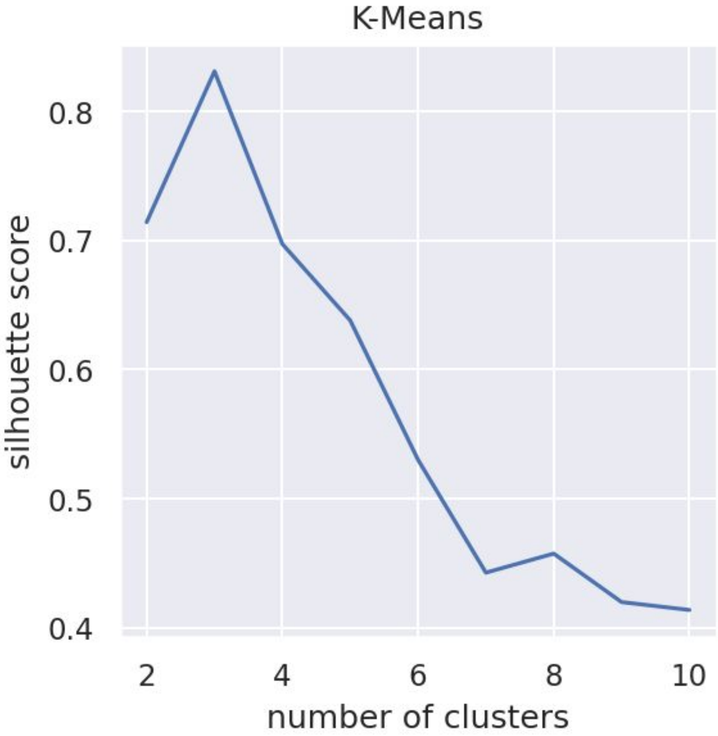
TODO - Implement K-Means

- Goal : Use K-Means to cluster the dimensionally reduced data under different cluster numbers

- # TODO
- Use the result of PCA to implement **K-Means**
- The clustering performance of each cluster is assessed by **silhouette score**
- under different cluster numbers (*k_grid* = 2~10)
- K-Means, silhouette_score

Baseline results





Summarize

1. Implement **PCA** (*n_components* = 2)
2. Use the result of PCA to implement **K-means** with different *k_grid*
3. Compute the **silhouette_score** with different *k_grid* (stored in score list)
4. Finish the Notebook (Methods && Conclusion)

NOTE

- After executing your code, download the .ipynb file and submit it to NTU COOL
- Submitted file name: student ID_week6_colab_homework.ipynb
- HW4 Deadline : 2024/10/14 23:59
- If there are any questions :

  Email: r12945048@ntu.edu.tw (add “[ML HW4]” to the beginning of the title.)