Statistical Learning and Data Mining

Module 11: Classification II

Semester 2, 2017

Discipline of Business Analytics, The University of Sydney Business School

Module 11: Classification II

1. Logistic regression

2. Gaussian discriminant analysis

3. Empirical example

4. Comparison of classification methods

Classification

In this module, we study two essential classification techniques: logistic regression (a discriminative method) and Gaussian discriminant analysis (a class of generative models).

We also consider applications and discuss the advantages and disadvantages of discriminative versus generative classifiers.

Business application: customer churn

Customer churn or attrition occurs when a customer leaves the current service provider and switch to a different merchandiser.

- What are the drivers of customer attrition?
- When is a current customer likely to end the relationship?
- What strategies improve the retention of profitable customers?
- Is it more profitable to invest additional resources into customer acquisition or retention?

Customer attrition can be very costly for companies that do not have good answers to these questions.

Issues addressed in customer retention modelling

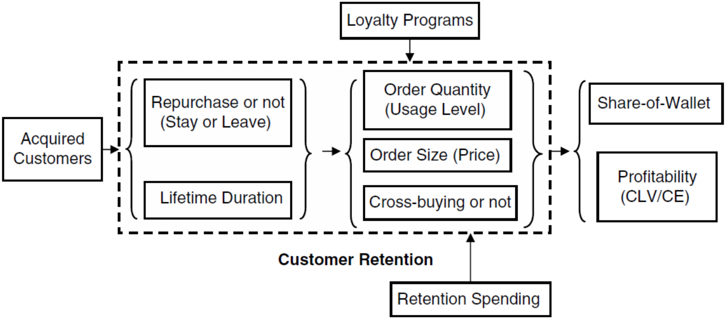


Figure from Kumar and Petersen, 2012.

Balancing acquisition and retention

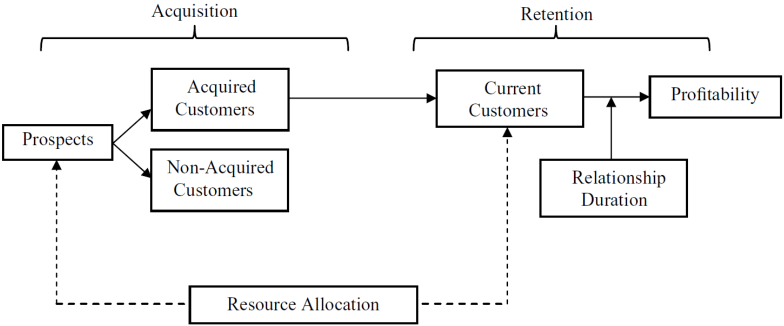


Figure from Kumar and Petersen, 2012.

Customer churn data

```
Response: whether the customer had churned by the end
of the observation period.

Predictors

1. Average number of dollars spent on marketing efforts
   to try and retain the customer per month.
2. Total number of categories the customer has purchased from.
3. Number of purchase occasions.
4. Industry: 1 if the prospect is in the B2B industry,
   0 otherwise.
5. Revenue: annual revenue of the prospect's firm.
6. Employees: number of employees in the prospect's firm.

Observations: 500.

Source: Kumar and Petersen (2012).
```

Logistic regression

Regression models for classification

Suppose that we want the specify a discriminative model for binary classification. The response Y follows the Bernoulli distribution,

$$Y = \begin{cases} 1 \text{ with probability } P(Y=1|X=\boldsymbol{x}) \\ 0 \text{ with probability } 1-P(Y=1|X=\boldsymbol{x}) \end{cases}$$

How to model the conditional probability $P(Y=1|X=\boldsymbol{x})$ as a function of the predictors?

Regression models for classification

Since $P(Y = 1|X = \boldsymbol{x}) = E(Y|X = \boldsymbol{x})$, one option is to specify a linear regression model

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \varepsilon.$$

This is called the **linear probability model**. However, there are a few reasons why we want to move beyond this framework.

Why not the linear probability model?

1. There is no guarantee that a linear probability model will generate probabilities between zero and one, since the regression function $\beta_0 + \sum_{j=1}^{p} \beta_j x_j$ is unconstrained. In other words, the linearity assumption does not hold.

2. The Bernoulli distribution has variance $p(\boldsymbol{x})(1 - p(\boldsymbol{x}))$. Hence, the linear probability model violates the classical assumption of constant error variance.

3. The linear probability approach does not generalise to categorical responses with more than two classes.

Logistic regression (key concept)

The **logistic regression model** is

$$Y|X = \boldsymbol{x} \sim \text{Ber}(p(\boldsymbol{x})),$$

where

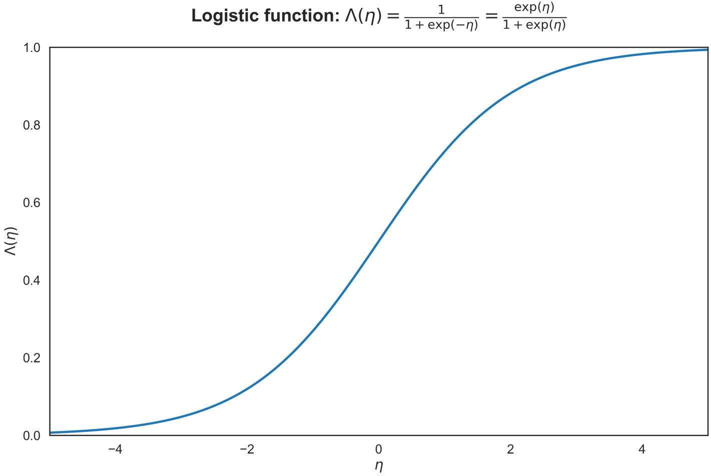
$$p(\boldsymbol{x}) = \frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_j)}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_j)}.$$

The **logistic function**

$$\frac{\exp(a)}{1 + \exp(a)} = \frac{1}{1 + \exp(-a)}$$

constrains the probability to be between zero and one.

Logistic function



Logistic Regression

Define the **odds ratio** as

$$\frac{p(\boldsymbol{x})}{1 - p(\boldsymbol{x})}.$$

We can show that

$$\frac{p(\boldsymbol{x})}{1 - p(\boldsymbol{x})} = \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j\right).$$

The logistic regression model therefore specifies a linear model for the log odds,

$$\log\left(\frac{p(\boldsymbol{x})}{1 - p(\boldsymbol{x})}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j,$$

where we call the left-hand side the logit transformation of the probability.

Maximum likelihood estimation

We estimate the logistic regression model by maximum likelihood. Recall that a Bernoulli random variable Y has probability mass function

$$p(y;\pi) = \pi^{y}(1-\pi)^{1-y}.$$

In the context of the logistic regression model, the probability mass function for a training case i is therefore

$$p(y_i|x_i) = p(\boldsymbol{x}_i)^{y_i}(1-p(\boldsymbol{x}_i))^{1-y_i}.$$

Maximum likelihood estimation

The likelihood function is

$$\begin{aligned}\ell(\boldsymbol{\beta}) &= p(y_1|\boldsymbol{x}_1)\, p(y_2|\boldsymbol{x}_2)\ \ldots\ p(y_N|\boldsymbol{x}_N) \\ &= \prod_{i=1}^{N} p(y_i|\boldsymbol{x}_i) \\ &= \prod_{i=1}^{N} p(\boldsymbol{x}_i)^{y_i}(1-p(\boldsymbol{x}_i))^{1-y_i} \\ &= \prod_{i=1}^{N} \left(\frac{p(\boldsymbol{x}_i)}{1-p(\boldsymbol{x}_i)}\right)^{y_i}(1-p(\boldsymbol{x}_i))\end{aligned}$$

Maximum likelihood estimation

The log-likelihood is

$$
\begin{aligned}
L(\boldsymbol{\beta}) &= \log\left(\prod_{i=1}^{N}\left(\frac{p(\boldsymbol{x}_i)}{1-p(\boldsymbol{x}_i)}\right)^{y_i}(1-p(\boldsymbol{x}_i))\right)\\
&= \sum_{i=1}^{N} y_i \log\left(\frac{p(\boldsymbol{x}_i)}{1-p(\boldsymbol{x}_i)}\right) + \log\left(1-p(\boldsymbol{x}_i)\right)\\
&= \sum_{i=1}^{N} y_i\left(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}\right) - \log\left(1+\exp\left(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}\right)\right)
\end{aligned}
$$

The negative log-likelihood $-L(\boldsymbol{\beta})$ is known as the **cross-entropy loss function** or log loss in machine learning.

Maximum likelihood estimation

The MLE for the logistic regression model is

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\ L(\boldsymbol{\beta}),$$

where

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{N} y_i \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right) - \log \left(1 + \exp \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right) \right)$$

Maximum likelihood estimation

Optimisation. Setting the partial derivatives of the log-likelihood to zero leads to estimation equations that are nonlinear in the coefficients. We therefore use numerical optimisation routines to obtain the estimates.

Statistical inference. We can conduct statistical inference for the logistic regression model using the large sample theory for ML estimation or the Bootstrap method.

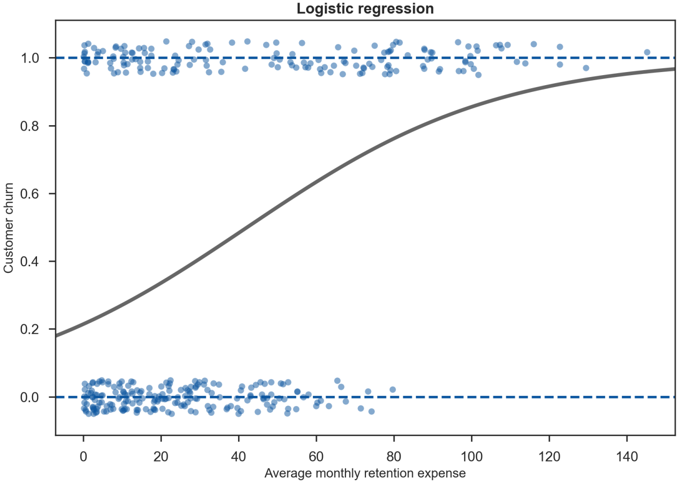
Predicted probabilities

The predicted probability given observed input values $x_1, \ldots, x_p$ is

$$\widehat{p}(\boldsymbol{x}) = \frac{\exp(\widehat{\beta}_0 + \sum_{j=1}^{p} \widehat{\beta}_j x_j)}{1 + \exp(\widehat{\beta}_0 + \sum_{j=1}^{p} \widehat{\beta}_j x_j)},$$

where $\widehat{\beta}_0$, $\widehat{\beta}_1, \ldots,$ $\widehat{\beta}_p$ are the maximum likelihood estimates.

Example: customer churn data



Example: customer churn data

```
                           Logit Regression Results
=============================================================================
Dep. Variable:                  Churn  No. Observations:                  350
Model:                          Logit  Df Residuals:                      348
Method:                           MLE  Df Model:                            1
Date:                                  Pseudo R-squ.:                  0.1319
Time:                                  Log-Likelihood:                -208.99
converged:                       True  LL-Null:                       -240.75
                                       LLR p-value:                 1.599e-15
==============================================================================
                  coef    std err          z      P>|z|     [0.025      0.975]
------------------------------------------------------------------------------
Intercept      -1.2959      0.189     -6.866      0.000     -1.666      -0.926
Avg_Ret_Exp     0.0308      0.004      7.079      0.000      0.022       0.039
==============================================================================
```

Example: customer churn

Suppose that we want to predict the probability that a customer with average retention expenses of 100 will churn.

$$\widehat{p} = \frac{\exp(\widehat{\beta}_0 + \widehat{\beta}_1 \times 100)}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 \times 100)}$$

$$= \frac{\exp(-1.296 + 0.031 \times 100)}{1 + \exp(-1.296 + 0.031 \times 100)}$$

$$= 0.856$$

Regularised logistic regression

Regularised risk minimisation applies to logistic regression. With an ℓ_1 penalty as in the lasso, we solve the minimisation problem

$$\min_{\boldsymbol{\beta}} \; -L(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|,$$

where

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{N} y_i \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right) - \log \left(1 + \exp \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right) \right).$$

Subset selection and dimension reduction with principal components also extend to logistic regression in a straightforward way.

Multinomial logistic regression (key concept)

The **multinomial logistic regression** is a generalisation of logistic regression to multiple classes. The model specifies

$$p(y=c|\boldsymbol{x}) = \frac{\exp\left(\beta_{0c} + \boldsymbol{\beta}_c^T \boldsymbol{x}\right)}{\sum_{c'=1}^{C} \exp\left(\beta_{0c'} + \boldsymbol{\beta}_{c'}^T \boldsymbol{x}\right)},$$

where $\boldsymbol{\beta}_c$ is the vector of coefficients for class c.

The **softmax function**

$$\mathcal{S}_c(a_1, \ldots, a_C) = \frac{\exp\left(a_c\right)}{\sum_{c'=1}^{C} \exp\left(a_{c'}\right)}$$

ensures that the conditional class probabilities are in the $(0,1)$ interval and add up to one.

Multinomial logistic regression

To avoid redundancy in the parameters, we usually specify $\beta_{0C} = 0$ and $\boldsymbol{\beta}_C = \mathbf{0}$. In this case,

$$p(y = c|\boldsymbol{x}) = \frac{\exp\left(\beta_{0c} + \boldsymbol{\beta}_c^T \boldsymbol{x}\right)}{1 + \sum_{c'=1}^{C-1} \exp\left(\beta_{0c'} + \boldsymbol{\beta}_{c'}^T \boldsymbol{x}\right)},$$

for $c = 1, \ldots, C - 1$, and

$$p(y = C|\boldsymbol{x}) = \frac{1}{1 + \sum_{c=1}^{C-1} \exp\left(\beta_{0c} + \boldsymbol{\beta}_c^T \boldsymbol{x}\right)}.$$

The choice of baseline label does not affect the predictions. It is not necessary restrict the model in this way with regularised estimation.

Gaussian discriminant analysis

Gaussian discriminant analysis (key concept)

In **Gaussian discriminant analysis**, we assume that the predictors are normally distributed conditional on the class

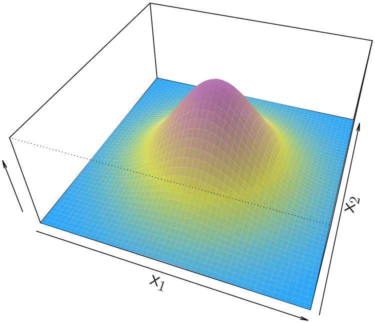
$$X|Y = c \sim N(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c),$$

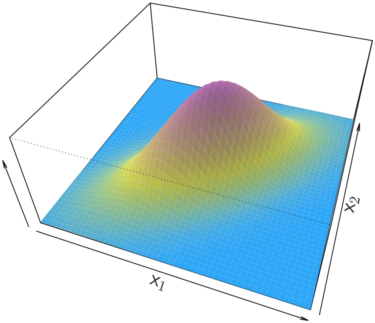
where μ_c is the mean of the predictors in class c and Σ_c is the covariance matrix.

The class conditional density is

$$p(\boldsymbol{x}|y = c) = \frac{1}{\sqrt{(2\pi)^p|\boldsymbol{\Sigma}_c|}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_c)^{\mathrm{T}}\boldsymbol{\Sigma}_c^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_c)\right).$$

Multivariate normal distribution





Bayes theorem for classification (review)

Let $\pi_c = P(Y = c)$ be the prior probability for class c. Using Bayes' theorem, the conditional probability is

$$p(y = c|\boldsymbol{x}) = \frac{\pi_c \, p(\boldsymbol{x}|y = c)}{\sum_{c'=1}^{C} \pi_{c'} \, p(\boldsymbol{x}|y = c')}.$$

Quadratic discriminant analysis

Using the multivariate normal density,

$$p(y=c|\boldsymbol{x}) = \frac{\pi_c \left(|2\pi \boldsymbol{\Sigma}_c|\right)^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_c)^{\mathrm{T}} \boldsymbol{\Sigma}_c^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_c)\right]}{\sum_{c'=1}^{C} \pi_{c'} \left(|2\pi \boldsymbol{\Sigma}_{c'}|\right)^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_{c'})^{\mathrm{T}} \boldsymbol{\Sigma}_{c'}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_{c'})\right]}.$$

We call this result **quadratic discriminant analysis** (QDA). If we assume that $\boldsymbol{\Sigma}_c$ is diagonal, QDA becomes equivalent to Naive Bayes.

Quadratic discriminant analysis

We define the **discriminant function** as

$$\delta_c(\boldsymbol{x}) = \log \pi_c - \frac{1}{2} \log |\boldsymbol{\Sigma}_c| - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_c)^{\mathrm{T}} \boldsymbol{\Sigma}_c^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_c).$$

We can write the conditional probability in terms of the discriminant function as

$$p(y = c|\boldsymbol{x}) = \frac{\exp\left[\delta_c(\boldsymbol{x})\right]}{\sum_{c'=1}^{C} \exp\left[\delta_{c'}(\boldsymbol{x})\right]}.$$

Quadratic discriminant analysis

We can describe any decision rule in terms of the discriminant score. For example, under the zero-one loss function the decision rule is

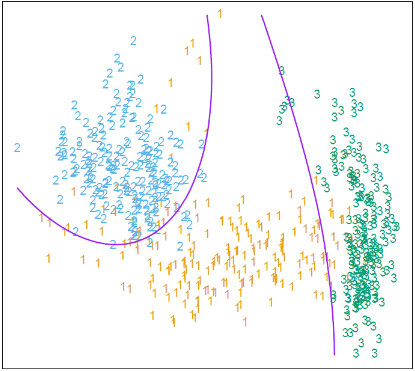
$$\widehat{Y}(\boldsymbol{x}) = \underset{c}{\operatorname{argmax}} \delta_c(\boldsymbol{x}).$$

The log-odds between two classes is

$$\log \frac{P(Y = c | X = \boldsymbol{x})}{P(Y = c' | X = \boldsymbol{x})} = \delta_c(\boldsymbol{x}) - \delta_{c'}(\boldsymbol{x}).$$

Because $\delta_c(\boldsymbol{x})$ is a quadratic function in $\boldsymbol{x}$, the decision boundary between two classes $\{\boldsymbol{x} : \delta_c(\boldsymbol{x}) = \delta_{c'}(\boldsymbol{x})\}$ is described by a quadratic equation.

Quadratic discriminant analysis



Maximum likelihood estimation

Let N_c be number of training observations in class c,

$$N_c = \sum_{i=1}^{N} I(y_i = c).$$

We estimate the model parameters as

$$\begin{aligned} \widehat{\pi}_c &= \frac{N_c}{N}, \\ \widehat{\boldsymbol{\mu}}_c &= \frac{1}{N_c} \sum_{i:\, y_i=c} \boldsymbol{x}_i, \\ \widehat{\boldsymbol{\Sigma}}_c &= \frac{1}{N_c} \sum_{i:\, y_i=c} (x_i - \widehat{\mu}_c)(x_i - \widehat{\mu}_c)^T, \end{aligned}$$

for $c = 1, \ldots, C$.

Linear discriminant analysis (key concept)

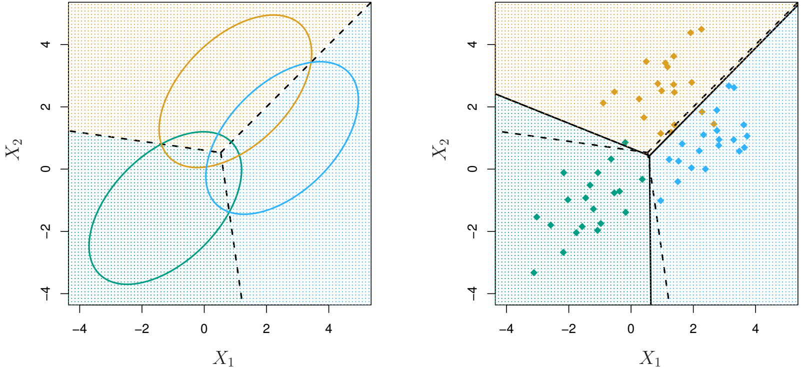
In **linear discriminant analysis** (LDA), we assume that the classes have a common covariance matrix. That is, $\boldsymbol{\Sigma}_c = \boldsymbol{\Sigma}$ for $c = 1, \ldots, C$.

This assumption leads to the linear discriminant function

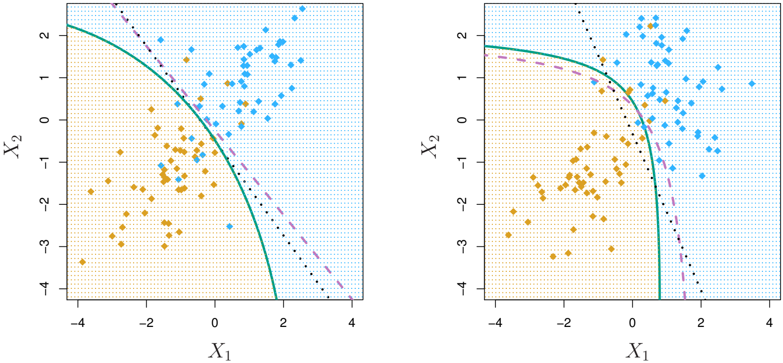
$$\delta_c(\boldsymbol{x}) = \log \pi_c + \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_c - \frac{1}{2} \boldsymbol{\mu}_c^{\mathrm{T}} \boldsymbol{\Sigma}_c^{-1} \boldsymbol{\mu}_c,$$

and therefore linear decision boundaries.

Linear discriminant analysis



Linear and quadratic discriminant analysis



Maximum likelihood estimation

The estimation algorithm LDA is similar to the one for QDA, except that we compute the pooled covariance matrix

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{N} \sum_{c=1^C} \sum_{i:\, y_i = c} (x_i - \widehat{\mu}_c)(x_i - \widehat{\mu}_c)^T,$$

for $c = 1, \ldots, C$.

Comparison to logistic regression

We can write the linear discriminant function as

$$\begin{aligned}\delta_c(\boldsymbol{x}) &= \log \pi_c + \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_c - \frac{1}{2} \boldsymbol{\mu}_c^{\mathrm{T}} \boldsymbol{\Sigma}_c^{-1} \boldsymbol{\mu}_c \\ &= \alpha_{0c} + \boldsymbol{\alpha}_c^T \boldsymbol{x},\end{aligned}$$

where

$$\alpha_{0c} = \log \pi_c - \frac{1}{2} \boldsymbol{\mu}_c^{\mathrm{T}} \boldsymbol{\Sigma}_c^{-1} \boldsymbol{\mu}_c$$

and

$$\boldsymbol{\alpha}_c = \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_c.$$

Comparison to logistic regression

Therefore, the conditional probability under the LDA is

$$p(y=c|\boldsymbol{x}) = \frac{\exp\left[\delta_c(\boldsymbol{x})\right]}{\sum_{c'=1}^{C}\exp\left[\delta_{c'}(\boldsymbol{x})\right]}$$

$$= \frac{\exp\left(\alpha_{0c} + \boldsymbol{\alpha}_c^T\boldsymbol{x}\right)}{\sum_{c'=1}^{C}\exp\left(\alpha_{0c'} + \boldsymbol{\alpha}_{c'}^T\boldsymbol{x}\right)}.$$

This is the same specification for the conditional probability as for logistic regression. The only difference between the two methods is how they estimate the coefficients.

Regularised Gaussian discriminant analysis

We can achieve a compromise between QDA and LDA by computing the covariance matrix

$$\widehat{\boldsymbol{\Sigma}}_c(\alpha) = \alpha \widehat{\boldsymbol{\Sigma}}_c + (1-\alpha)\widehat{\boldsymbol{\Sigma}},$$

for $\alpha \in [0, 1]$ selected by cross validation.

Another useful extension is to shrink $\boldsymbol{\Sigma}_c$ or $\boldsymbol{\Sigma}$ toward diagonal matrices, achieving a compromise between QDA or LDA and the Naive Bayes method.

Extensions

Alternative distributions. We can use the multivariate t distribution instead of the normal distribution for better robustness to outliers.

Mixture discriminant analysis. In mixture modelling, we assume that that the class conditional distributions are mixtures of distributions

$$p(\boldsymbol{x}|y=c) = \sum_{r=1}^{R_c} \pi_{cr} p(\boldsymbol{x}; \boldsymbol{\theta}_{cr}),$$

leading to highly flexible models that can approximate complex decision boundaries.

Empirical example

Customer churn data

- We randomly split the data to allocate 70% of customers (350 observations) to the training set.
- The customer attrition rate in the training data is 45%.
- Among the six predictors, three are continuous, two are count variables, and one is binary. With the exception of revenue, the continuous predictors are pronouncedly positively skewed.

Logistic regression

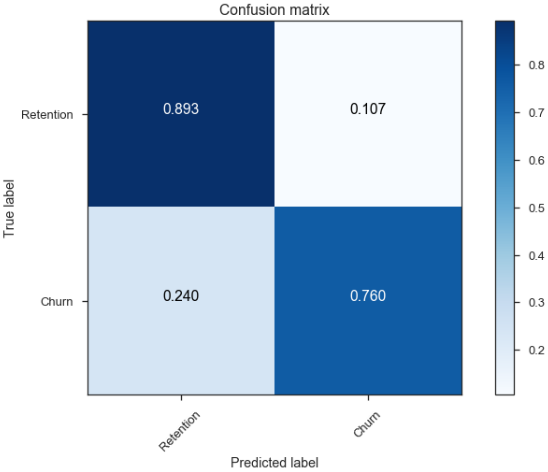
```
                           Logit Regression Results
==============================================================================
Dep. Variable:                  Churn   No. Observations:                  350
Model:                          Logit   Df Residuals:                      343
Method:                           MLE   Df Model:                            6
Date:                                   Pseudo R-squ.:                  0.4722
Time:                                   Log-Likelihood:                -127.06
converged:                       True   LL-Null:                       -240.75
                                        LLR p-value:                 2.783e-46
==================================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------
const              4.5768      0.897      5.101      0.000       2.818       6.335
Avg_Ret_Exp        0.0584      0.008      7.596      0.000       0.043       0.073
Revenue           -0.0258      0.010     -2.537      0.011      -0.046      -0.006
Employees         -0.0039      0.000     -7.846      0.000      -0.005      -0.003
Total_Crossbuy    -0.8610      0.128     -6.705      0.000      -1.113      -0.609
Total_Freq        -0.0547      0.027     -2.002      0.045      -0.108      -0.001
Industry           0.2758      0.335      0.823      0.410      -0.381       0.933
==================================================================================
```

Test results

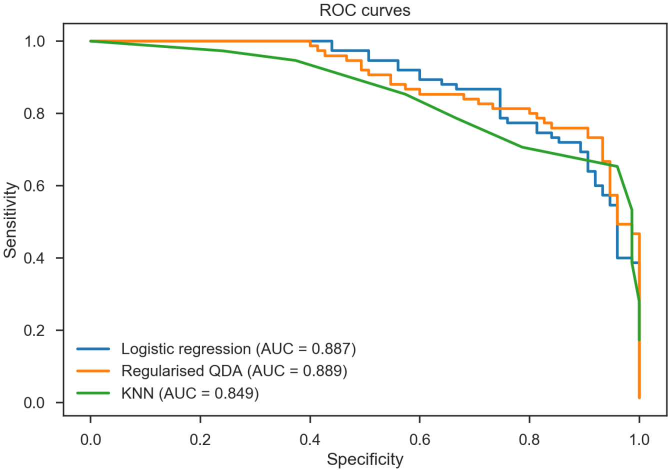
Classification results

	Error rate	Sensitivity	Specificity	AUC	Precision
Logistic regression	0.213	0.747	0.827	0.887	0.812
ℓ_1 regularised	0.213	0.747	0.827	0.888	0.812
ℓ_2 regularised	0.207	0.720	0.867	0.887	0.844
LDA	0.187	0.760	0.867	0.887	0.851
QDA	0.187	0.760	0.867	0.889	0.851
Regularised QDA	0.173	0.760	0.893	0.889	0.877
KNN	0.193	0.653	0.960	0.849	0.942

Confusion matrix (regularised QDA)



ROC curves


ROC curves
Sensitivity
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Logistic regression (AUC = 0.887)
Regularised QDA (AUC = 0.889)
KNN (AUC = 0.849)

Comparison of classification methods

Generative vs discriminative classifiers

If the assumptions of Gaussian discriminant analysis are correct, the model will need less training data than logistic regression to achieve a given level of performance (since the additional assumptions will improve estimation efficiency).

In the same way, if the assumptions are incorrect, logistic regression will perform better.

Generative vs discriminative classifiers

Some other advantages and disadvantages are as follows.

Generative classifiers. Easy to fit, do not need to be retrained if we new classes, can easily handle missing features, can be used in semi-supervised learning (partially labelled data).

Discriminative classifiers. Can easily handle constructed predictors, generally better calibrated probability estimates.

Review questions

- What is the logistic regression model?
- What is Gaussian discriminant analysis?
- What is the difference between LDA and QDA?
- Identify whether each classification method that we studied so far is generative or discriminative.
- What are the relative advantages and disadvantages of generative vs discriminative classifiers?
- In what situations do each of the classification methods that we have studied so far tend to be most useful?