Chapter 4 Classification

A supervised learning technique where the goal is to categorize new, unseen data into predefined classes or categories.

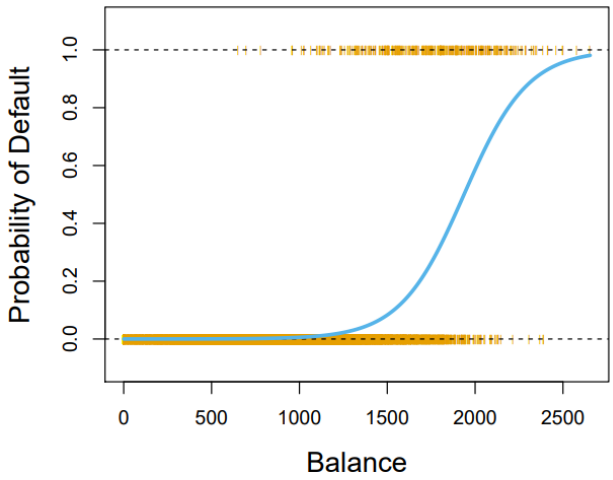
Why not linear regression?

- Encoding can produce different linear models
- For two outcomes we could but results might be out of range [0,1], which makes no sense when looking for the probability of an event.

Logistic Regression

- Model the probability that Y belongs to a particular category.

$$p(x) = \frac{e^{(\beta_0+\beta_1 x)}}{1 + e^{(\beta_0+\beta_1 x)}}$$



- To fit the model, we use maximum likelihood
- Odds in (0, infinity):

$$\frac{p(x)}{1 - p(x)} = e^{\beta_0+\beta_1 \cdot x}$$

- Log odds or logit

$$\log\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x$$

- Increase in X by one unit changes the loγ odds by β_1

Estimating coefficients

- Use maximize likelihood method
- Read the book for the math
- Or use Python
- Z-static same role as t-static
- $H_0: \beta_1 = 0$

Making Predictions

- Once we have coefficients, we can predict the probability of an outcome
- We can also fit a model to dummy variables

Multiple Logistic Regression

- Multiple predictors
- Like in linear regression, same idea
- $$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p,$$
- $$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$
- Use python to fit it
- Results might be counterintuitive or seem paradoxical so it's important to understand conditional probability
- See pages 143-144 for example

Multinomial Logistic Regression

- For more than two classes
- We select Kth class to serve as the baseline

$$\log[\mathrm{P}(\mathrm{Y}=\ \mathrm{k}|\mathrm{X}=\mathrm{x})/\mathrm{P}(\mathrm{Y}=\ \mathrm{K}|\mathrm{X}=\mathrm{x})\ ]]$$

- Interpretation of coefficients depends on the baseline

Generative models for classification

Why and when?

- When there is substantial separation between the classes, the parameter estimates for the logistic regression model are instable. Generative models don't have that problem

- If the distribution of the predictors of X is approximately normal in each of the classes and the sample is small, then the approaches here may be more accurate
- The method in this section can be naturally extended to the case of more than two classes

Baye's Theorem

The probability that observation belongs to the k_th class, given the predictor value of that observation.

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

- π_κ: prior probability
- $f_k(x)$: density function
- $p_k(x)$: posterior probability
- We can compute the π_κ by computing the fraction of the observation that belongs to the kth class
- $f_k(x)$ is harder and we have different methods

Linear Discriminant Analysis (LDA) for p=1

- p=1: only one predictor
- first assumption: $f_k(x)$is a normal or gaussian
- Bayes Precision Boundary: $f_1(x) = f_2(x)$

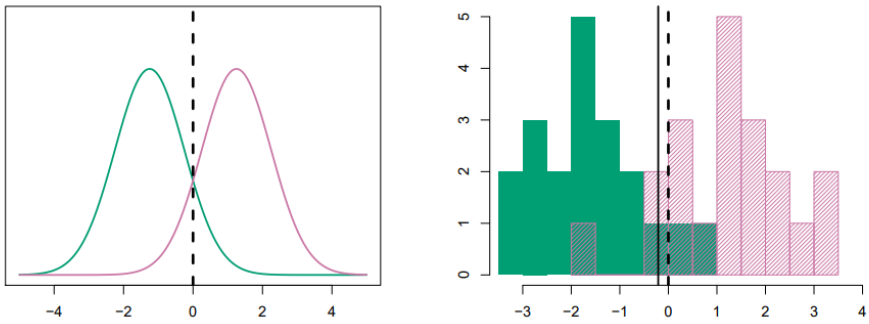


FIGURE 4.4. Left: *Two one-dimensional normal density functions are shown. The dashed vertical line represents the Bayes decision boundary.* Right: *20 observations were drawn from each of the two classes, and are shown as histograms. The Bayes decision boundary is again shown as a dashed vertical line. The solid vertical line represents the LDA decision boundary estimated from the training data.*

LDA for p>1

- Multiple predictors

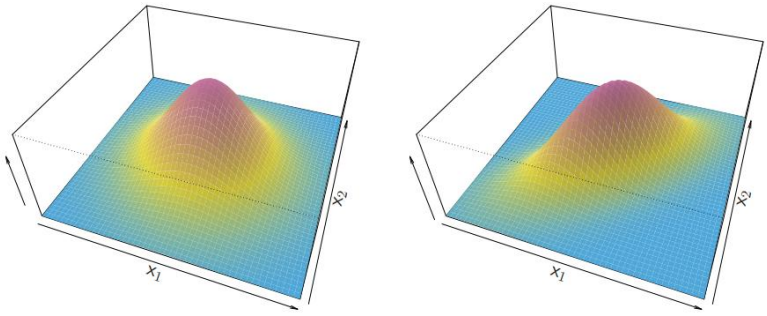


FIGURE 4.5. *Two multivariate Gaussian density functions are shown, with* $p = 2$. Left: *The two predictors are uncorrelated.* Right: *The two variables have a correlation of* 0.7.

-

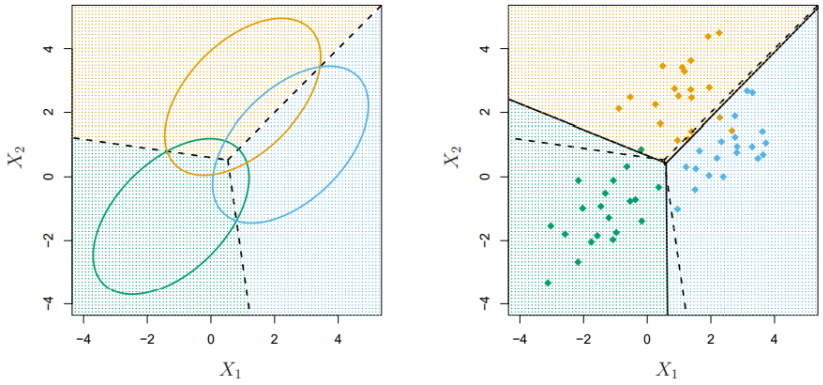


FIGURE 4.6. *An example with three classes. The observations from each class are drawn from a multivariate Gaussian distribution with* $p = 2$, *with a class-specific mean vector and a common covariance matrix.* Left: *Ellipses that contain* 95 % *of the probability for each of the three classes are shown. The dashed lines are the Bayes decision boundaries.* Right: 20 *observations were generated from each class, and the corresponding LDA decision boundaries are indicated using solid black lines. The Bayes decision boundaries are once again shown as dashed lines.*

-
- Bayes classifier is not sensitive. It will yield the smallest possible total number of is classified observations regardless of the class from which they came. For example, it might be able to classify correctly who will NOT default on credit card debt, but fully miss who will default, making the classifier useless.
- We can modify that by adjusting the threshold.

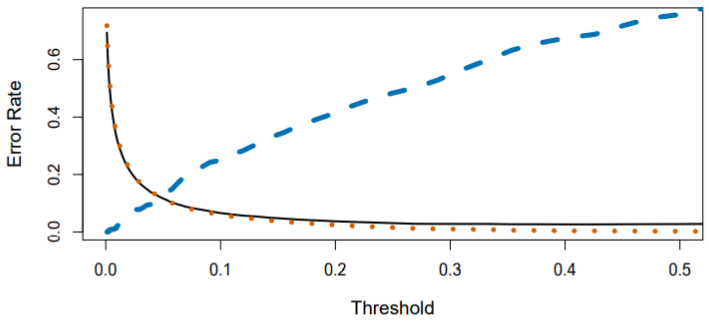
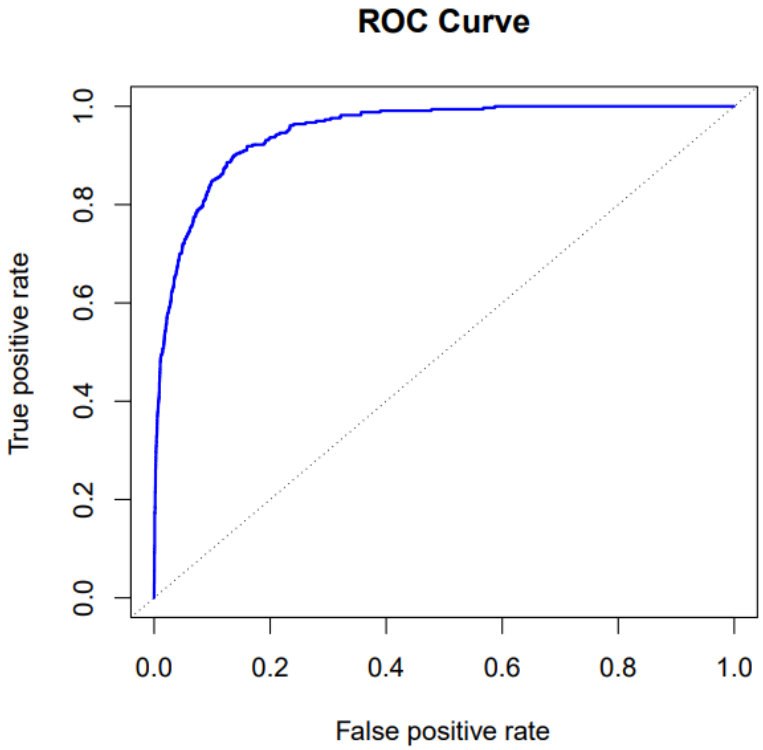


FIGURE 4.7. *For the* `Default` *data set, error rates are shown as a function of the threshold value for the posterior probability that is used to perform the assignment. The black solid line displays the overall error rate. The blue dashed line represents the fraction of defaulting customers that are incorrectly classified, and the orange dotted line indicates the fraction of errors among the non-defaulting customers.*

-
- How? Domain knowledge such detailed information about the problem/project
- Use ROC method. The are under the curve gives **Performance**



Summary of Terms and important measures

		True class		
		− or Null	+ or Non-null	Total
Predicted class	− or Null	True Neg. (TN)	False Neg. (FN)	N^*
	+ or Non-null	False Pos. (FP)	True Pos. (TP)	P^*
	Total	N	P	

TABLE 4.6. *Possible results when applying a classifier or diagnostic test to a population.*

Name	Definition	Synonyms
False Pos. rate	FP/N	Type I error, 1−Specificity
True Pos. rate	TP/P	1−Type II error, power, sensitivity, recall
Pos. Pred. value	TP/P^*	Precision, 1−false discovery proportion
Neg. Pred. value	TN/N^*	

TABLE 4.7. *Important measures for classification and diagnostic testing, derived from quantities in Table 4.6.*

Quadratic Discriminant Analysis (QDA)

- Same as LDA but!
- Assumes each of the observations has its own covariance matrix
- An observation from the kth class is of the form

$$x \sim N(\mu_k, \Sigma_k)$$

Why does this matter? Why LDA or QDA?

- LDA: flexible, less variance => potentially can improve prediction performance BUT can suffer from high bias
    - If fewer training observations, then LDA is better

Naïve Bayes

- Different approach at $f_k(x)$
- **Within the kth class, the p predictors are independent**
- We don't really believe that, but it can be pretty good, especially when n is not large enough relative to p.
- Introduces some bias, but reduces variance

Estimate the one-dimensional density function f_{kj}

To estimate the one-dimensional density function f_{kj} using training data x1j ,…,xnj , we have a few options.

- If X_j is quantitative, then QDA but assume that the class-specific covariance matrix is diagonal.
- OR use non-parametric estimate. Use kernel density estimator. Alternatively make histograms for the observations of the j_th predictor, then $f_{kj}(x_j)$is the fraction of the training observation in the k_th class that belongs to the same bin as x_j
- OR if X_j is qualitative, count proportions of training observations corresponding to each class

$$\hat{f}_{kj}(x_j) = \begin{cases} 0.32 & \text{if } x_j = 1 \\ 0.55 & \text{if } x_j = 2 \\ 0.13 & \text{if } x_j = 3. \end{cases}$$

- Naïve Bayes will be better than LDA or QDA if p is large or n small, so reducing variance is important.

A comparison of classification methods

- See 4.5.1 for a full analytical comparison
- None of LDA, QDA, or Naïve Bayes uniformly dominates over the other: it depends on the true distribution of the predictors in each of the K classes, as well as other considerations, such as values of n and p
- We expect LDA to outperform log. Reg. when the normality assumption (approximately) holds, and se expect logistic regression to outperform otherwise.
- Recall KNNs from Chapter 2. KNN dominates log. Reg. and LDA when the decision boundary s non-linear, provided that n is very large and p is small
- KNN requires a lot of observation; n >>> p
- If n is modest or p not very small, then QDA might be better
- Unlike log. Reg., we don't get a table of coefficients
- Read 4.5.2 for an empirical comparison with a use case

Generalized Linear Models

- What if the response is not quant or qual? (bikeshare data)
- Linear regression will give us numbers out of logical range, variance doesn't change, and can give non-integers when it makes no sense to do. NOT A GOOD IDEA

Poisson Regression

- Interpretation: increase in X_j by one unit is associated with a change in $E(y) = \lambda$ by a factor of e^{β_j}

- Mean-variance relationship: $Var(y) = E(y) = \lambda$
- Non negative fitted values: none

In greater generality, anything from the exponential family of distribution can be used.