Chap 12: Classification

Modern Applied Statistics

November 21, 2022

Seoul National University

Outline

Introduction

Introduction

In the statistical literature the word is used in two distinct senses.

- The sense of cluster analysis discussed in Section 11.2
- The other meaning (Ripley, 1997) of allocating future cases to one of g prespecified classes

It is sometimes helpful to distinguish discriminant analysis in the sense of describing the differences between the g groups from classification, allocating new observations to the groups.

- The first provides some measure of explanation
- The second can be a 'black box' that makes a decision without any explanation.

Discriminant Analysis

Discriminant Analysis

Suppose that we have a set of g classes, and for each case we know the class. We can then use the class information to help reveal the structure of the data.

The sample covariance matrices

$$W = \frac{(X - GM)^T (X - GM)}{n - g}, \quad B = \frac{(GM - \mathbf{1}\bar{x})^T (GM - \mathbf{1}\ \bar{x})}{g - 1}$$

- W : the within-class covariance matrix.
- B : the between-classes covariance matrix.
- M : the $g \times p$ matrix of class means.
- G : the $n \times g$ matrix of class indicator variables.
  - Then the predictions are GM.
- $\bar{x}$: the means of the variables over the whole sample.

Note that B has rank at most $\min(p, g - 1)$.

Discriminant Analysis

Fisher introduced a **linear discriminant analysis** seeking a linear combination xa of the variables that has a maximal ratio of the separation of the class means to the within-class variance.

- Maximizing the ratio $\boldsymbol{a}^T B\boldsymbol{a}/\boldsymbol{a}^T W\boldsymbol{a}$.
  - Choose a sphering xS of the variables.
  - The problem is to maximize $\boldsymbol{a}^T B\boldsymbol{a}$ subject to $\|\boldsymbol{a}\| = 1$.
  - This is solved by taking a to be the eigenvector of B corresponding to the largest eigenvalue.
  - The linear combination a is unique up to a change of sign.

Discriminant Analysis

As for principal components, we can take further linear components corresponding to the next largest eigenvalues.

- **Eigenvalues**: the proportions of the between classes variance explained by the linear combinations.
- The corresponding transformed variables are called the **linear discriminants or canonical variates**.
- The linear discriminants are conventionally centred to have mean zero on dataset.

Discriminant Analysis

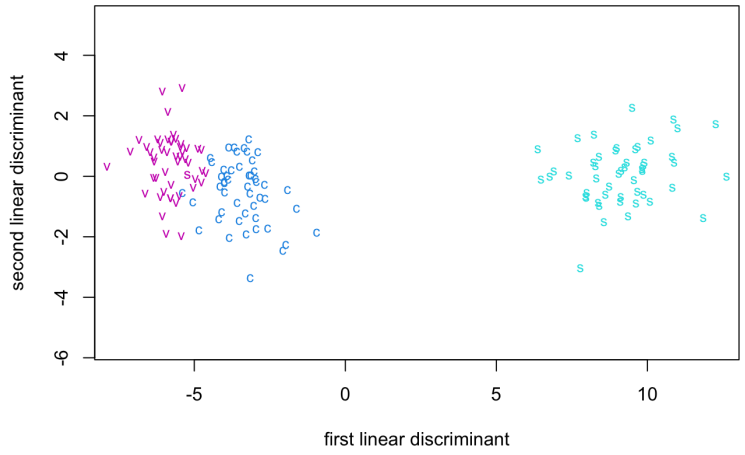


Figure 1: The log iris data on the first two discriminant axes

▶ This shows that 99.65% of the between-group variance is on the first discriminant axis.

- Rao (1948) used the unweighted covariance matrix of the group means.
- Our approach uses a covariance matrix **weighted by the prior probabilities of the classes** if these are specified.

Discrimination for normal populations

An alternative approach to discrimination is via probability models.

- Posterior distribution of the classes after observing x is

$$p(c \mid \boldsymbol{x}) = \frac{\pi_c p(\boldsymbol{x} \mid c)}{p(\boldsymbol{x})} \propto \pi_c p(\boldsymbol{x} \mid c)$$

- π_c: The prior probabilities of the classes
- $p(\boldsymbol{x} \mid c)$: The densities of distributions of the observations for each class
- **Bayes rule**: The allocation rule which makes the smallest expected number of errors chooses the class with maximal $p(c \mid x)$.

Discriminant Analysis

Suppose the distribution for class c is multivariate normal with mean $\boldsymbol{\mu}_c$ and covariance Σ_c. Then the Bayes rule minimizes

$$\begin{aligned} Q_c &= -2 \log p(\boldsymbol{x} \mid c) - 2 \log \pi_c \\ &= (\boldsymbol{x} - \boldsymbol{\mu}_c) \Sigma_c^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_c)^T + \log |\Sigma_c| - 2 \log \pi_c \end{aligned}$$

- The first term is the squared Mahalanobis distance to the class centre.
- The difference between the Q_c for two classes is a quadratic function of x.
  - **Quadratic discriminant analysis.**
- The boundaries of the decision regions are quadratic surfaces in x space.

Discriminant Analysis

Suppose that the classes have a common covariance matrix Σ.

- Differences in the Q_c are then linear functions of $\boldsymbol{x}$.

We can maximize $-Q_c/2$ or

$$L_c = \boldsymbol{x}\Sigma^{-1}\boldsymbol{\mu}_c^T - \boldsymbol{\mu}_c\Sigma^{-1}\boldsymbol{\mu}_c^T/2 + \log \pi_c$$

To use Q_c or L_c we have to estimate μ_c and Σ_c or Σ.

- Using obvious estimates, estimate μ_c as the sample mean, Σ_c as covariance matrix, and Σ as W.

Discriminant Analysis

How does this relate to Fisher's linear discrimination?

▶ L_c gives new variables, the linear discriminants, with unit within-class sample variance.

- On these variables the Mahalanobis distance is

$$\|x - \mu_c\|^2$$

  - ▶ Only the first r components of the vector depend on c.

$$L_c = \boldsymbol{x}\boldsymbol{\mu}_c^T - \|\boldsymbol{\mu}_c\|^2/2 + \log \pi_c$$

- We can work in r dimensions.

$$L_2 - L_1 = \boldsymbol{x}\left(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\right)^T + \text{ const}$$

  - ▶ An affine function of the linear discriminant.

Crabs dataset

Construct a rule to predict the sex of a future Leptograpsus crab of unknown colour form.

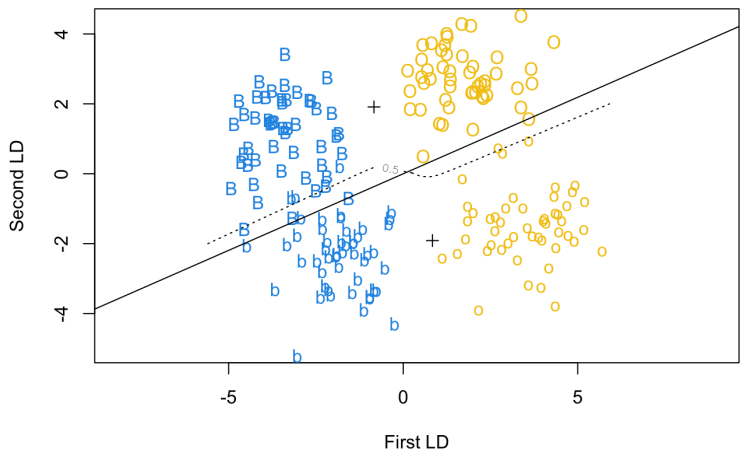


Figure 2: Linear discrirninants for the crabs data

- The first two linear discriminants dominate the between-group variation.
- Using the first two linear discriminants as the data will provide a very good approximation.

Robust estimation of multivariate location and scale

- Apply a robust location estimator to each component of a multivariate mean.
- Consider the estimation of mean and variance simultaneously.
- Two methods for robust covariance estimation
  - ▶ Our function `cov.rob`
  - ▶ The S-PLUS functions `cov.mve` and `cov.mcd` and `covRobin` in library section robust.

  - MVE(minimum volume ellipsoid method): seeks an ellipsoid containing $h = \lfloor (n+p+1)/2 \rfloor$ points that is of minimum volume.

  - MCD(minimum covariance determinant method): eeks h points whose covariance has minimum determinant.

Robust estimation of multivariate location and scale

- The search for an MVE or MCD
  - ▶ provides h points whose mean and variance matrix give an initial estimate.
  - ▶ This is refined by selecting those points whose Mahalanobis distance from the initial mean using the initial covariance is not too large
  - ▶ and returning their mean and variance matrix.
- Alternative approach
  - ▶ Fitting a multivariate t_ν distribution for a small number ν of degrees of freedom.
  - ▶ This is implemented in our function `cov.trob`
- Normally `cov.trob` is faster than `cov.rob`, but it lacks the latter's extreme resistance.

Robust estimation of multivariate location and scale

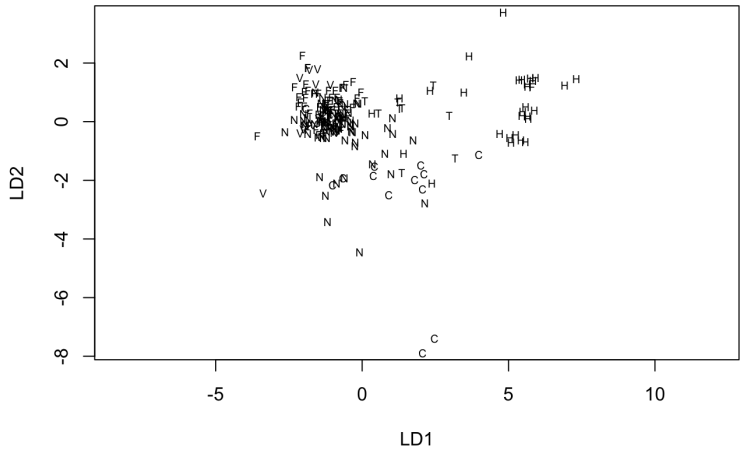


Figure 3: The fgl data on the first two discriminant axes

- We can use linear discriminant analysis on more than two classes, and illustrate this with the forensic glass dataset `fgl`.

Classification Theory

Classification Theory

- In the terminology of pattern recognition
  - ▶ Training set: The given examples together with their classifications.
  - ▶ Test set: Future cases form.
  - ▶ **Our primary measure of success is the error rate.**
- The type of errors
  - ▶ A confusion matrix gives the number of cases with true class i classified as of class j.
  - ▶ We assign costs L_{ij} to allocating a case of class i to class j.

Classification Theory

The average error cost is minimized by the Bayes rule, which is to allocate to the class c minimizing $\sum_i L_{ic} p(i \mid \boldsymbol{x})$.

- $p(i \mid \boldsymbol{x})$ is the posterior distribution of the classes after observing $\boldsymbol{x}$.
- If the costs of all errors are the same, this rule amounts to choosing the class c with the largest posterior probability $p(c \mid \boldsymbol{x})$.

The minimum average cost is known as the Bayes risk.

- Estimate a lower bound for it by the method of Ripley (1996, pp. 196-7).

Classification Theory

In Section 12.1,

- How $p(c \mid \boldsymbol{x})$ can be computed for normal populations.
- How estimating the Bayes rule with equal error costs leads to linear and quadratic discriminant analysis.

As our functions `predict.lda` and `predict.qda` return posterior probabilities.

- They can be used for classification with error costs.

Classification Theory

The posterior probabilities $p(c \mid \boldsymbol{x})$ may also be estimated directly.

- For two classes: we can model $p(1 \mid \boldsymbol{x})$ using a logistic regression, fitted by glm.
- For more than two classes: we need a multiple logistic model.
  - ▶ Using a surrogate log-linear Poisson GLM model.
  - ▶ Using the multinom function in library section `nnet`.

Classification trees model the $p(c \mid \boldsymbol{x})$ directly.

▶ Since the posterior probabilities are given by the `predict` method it is easy to estimate the Bayes rule for unequal error costs.

Predictive and 'plug-in' rules

To find the Bayes rule we need to know the posterior probabilities $p(c \mid \boldsymbol{x})$.

- Since these are unknown, we use an parametric family $p(c \mid \boldsymbol{x}; \theta)$.
- We act as if $p(c \mid x; \hat{\theta})$ were the actual posterior probabilities.
  - $\hat{\theta}$ is an estimate computed from the training set $\mathcal{T}$.
  - **'plug-in' rule.**

The 'correct' estimate of $p(c \mid \boldsymbol{x})$ is (Ripley, 1996, §2.4) to use the predictive estimates

$$\tilde{p}(c \mid \boldsymbol{x}) = P(c = c \mid \boldsymbol{X} = \boldsymbol{x}, \mathcal{T}) = \int p(c \mid \boldsymbol{x}; \theta) p(\theta \mid \mathcal{T}) \mathrm{d}\theta$$

Predictive and 'plug-in' rules

$$\tilde{p}(c \mid \boldsymbol{x}) = P(c = c \mid \boldsymbol{X} = \boldsymbol{x}, \mathcal{T}) = \int p(c \mid \boldsymbol{x}; \theta) p(\theta \mid \mathcal{T}) \mathrm{d}\theta$$

▶ It is possible for linear and quadratic discrimination with appropriate 'vague' priors on θ.

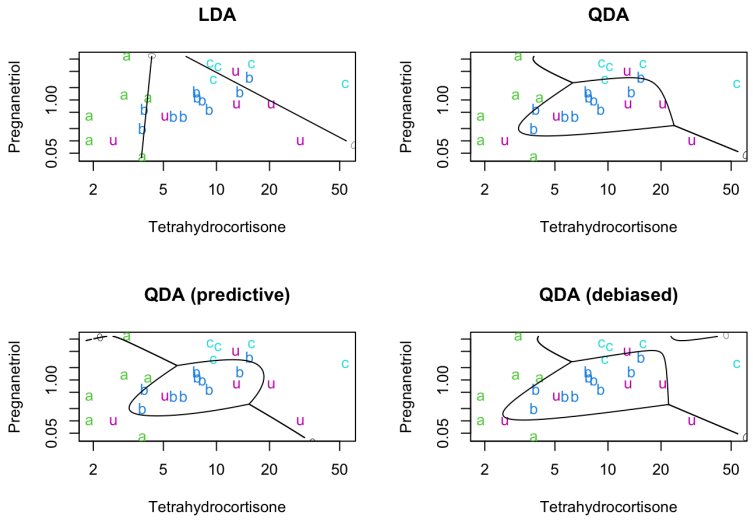
▶ This estimate is implemented by `method = "predictive"` of the predict methods for our functions `lda` and `qda`.

▶ When there are not enough data for a good estimate of the variance matrices, Considerable improvement can be obtained by using an unbiased estimator of $\log p(\boldsymbol{x} \mid c)$, implemented by the argument `method = "debiased"`.

A simple example: Cushing's syndrome

- The data are on diagnostic tests on patients with Cushing's syndrome.
- This dataset has three recognized types of the syndrome represented as a, b, c.
- The observations are urinary excretion rates, and are considered on log scale.

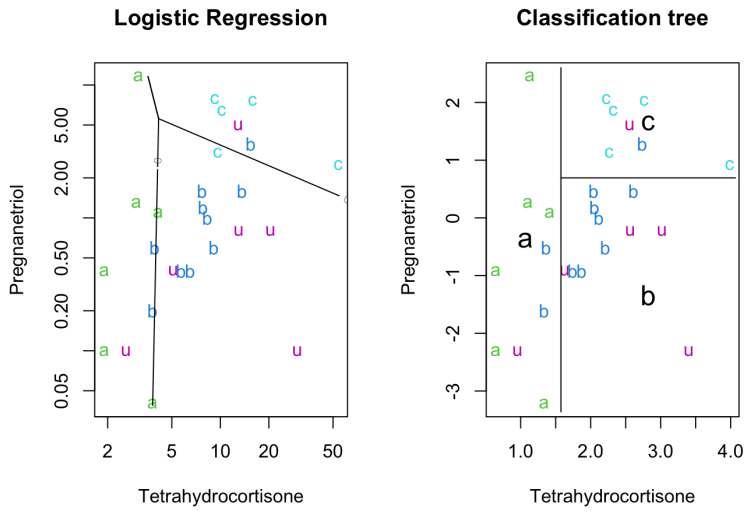
Classification Theory



▶ There are six patients of unknown type (marked *u*), one of whom was later found to be of a fourth type, and another was measured faultily.

▶ Figure 12.4 shows the classifications produced by 1 da and the various options of quadratic discriminant analysis.

Classification Theory



- When, as here, the classes have quite different variance matrices, linear and logistic discrimination can give quite different answers.

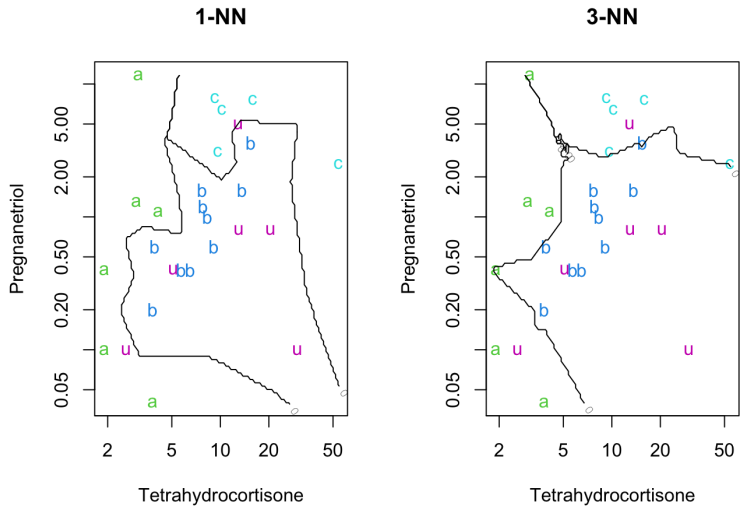
Non-parametric Rules

Non-parametric Rules

- Library section `class`: The k-nearest neighbour classifier and related methods and learning vector quantization.
- These are all based on
  - ▶ finding the k **nearest examples** in some reference set.
  - ▶ taking a **majority vote** among the classes of these k examples.
  - ▶ estimating the **posterior probabilities** $p(c \mid \boldsymbol{x})$ by the proportions of the classes among the k examples.
- These methods almost always measure 'nearest' by **Euclidean distance**.

Non-parametric Rules

For the **Cushing's syndrome data** we use Euclidean distance on the logged covariates, rather arbitrarily scaling them equally.

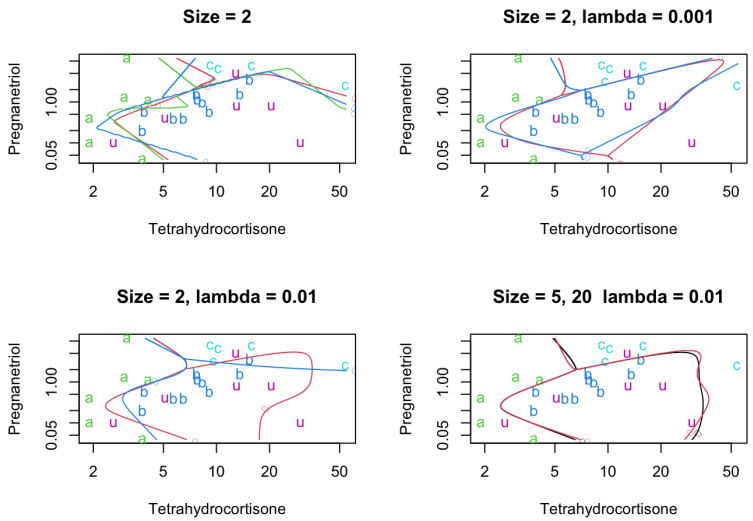


▶ This dataset is too small to try the editing and LVQ methods in library section class.

Neural Networks

Neural Networks

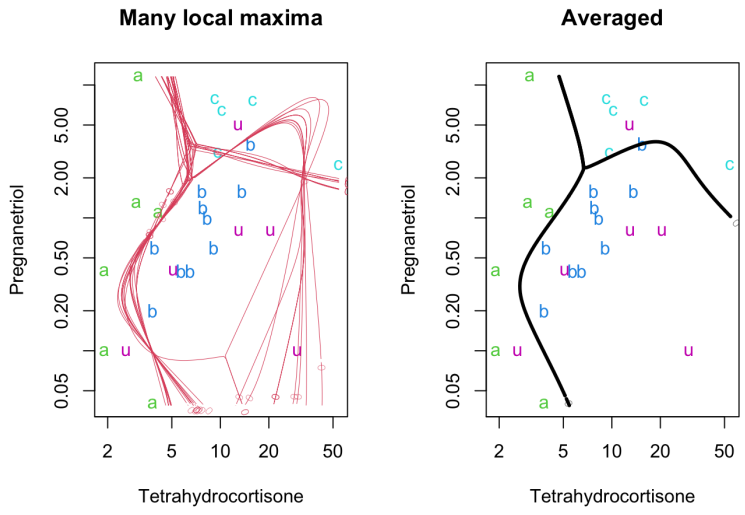
Neural networks provide a flexible non-linear extension of multiple logistic regression, as we saw in Section 8.10.



▶ We see that in all cases there are multiple local maxima of the likelihood, since different runs gave different classifiers.

Neural Networks

Once we have a **penalty**, the choice of the number of hidden units is often not critical (see Figure 12.7). The spirit of the predictive approach is to average the predicted $p(c \mid \boldsymbol{x})$ over the local maxima.



▶ Note that there are two quite different types of local maxima occurring here, and some local maxima occur several times (up to convergence tolerances). An average does better than either type of classifier.

Support Vector Machine

Support Vector Machine

Support vector machines (SVMs)

The method for $g = 2$ classes

- **Logistic regression** will fit exactly in separable cases where there is a hyperplane that has all class-one points on one side and all class-two points on the other.
- **Support vector methods** attempt directly to find a hyperplane in the middle of the gap, that is with maximal margin.
  - ▶ Such a hyperplane has support vectors, data points that are exactly the margin distance away from the hyperplane.

Support Vector Machine

The problem: Separating hyperplane usually doesn't exist. This difficulty is tackled in two ways.

- We can allow some points to be on the wrong side of their margin subject to a constraint on the total of the 'mis-fit' distances being less than some constant, with Lagrange multiplier $C > 0$.
- The set of variables is expanded greatly by taking non-linear functions of the original set of variables.
  - Thus we seek $f(x) = h(\boldsymbol{x})^T\boldsymbol{\beta} + \beta_0 = 0$ for a vector of $M \gg p$ functions h_i.
  - Then finding a optimal separating hyperplane is equivalent to solving

$$\min_{\beta_0,\boldsymbol{\beta}} \sum_{i=1}^{n} [1 - y_i f(\boldsymbol{x}_i)]_+ + \frac{1}{2C}\|\boldsymbol{\beta}\|^2$$

where $y_i = \pm 1$ for the two classes.

Support Vector Machine

There is an implementation of SVMs for R in function `svm` in package `e1071`. Here `cost` is C and `gamma` is a coefficient of the kernel used to form h.

- We can try a 10-fold cross-validation
- The extension to $g > 2$ classes: The `svm` function uses one attributed to Knerr *e* al. (1990) in which classifiers are built comparing each pair of classes, and the majority vote amongst the resulting $g(g-1)/2$ classifiers determines the predicted class.

Forensic Glass Example

Forensic Glass Example

The forensic glass dataset `fgl` has 214 points from six classes with nine measurements, and provides a fairly stiff test of classification methods.

- The types of glass do not form compact well-separated groupings, and the marginal distributions are far from normal.
- There are some small classes, so we cannot use quadratic discriminant analysis.

Forensic Glass Example

- **Performance assessment**: 10 -fold cross-validation, using the same random partition for all the methods.
- **Logistic regression** provides a suitable **benchmark**, and in this example linear discriminant analysis does equally well.
- We can use nearest-neighbour methods to estimate the lower bound on the Bayes risk as about 10%
- We need to cross-validate over the choice of tree size, which does vary by group from four to seven.

Forensic Glass Example

Neural Networks

- For testing neural network models by V-fold cross-validation, we rescale the dataset so the inputs have range $[0, 1]$.
- We want to average across several fits and to choose the number of hidden units and the amount of weight decay by an inner cross-validation.

Forensic Glass Example

Neural Networks

- This fits a neural network 1000 times, and so is fairly slow
- An alternative is to use logarithmic scoring.
- Rather than count 0 if the predicted class is correct and 1 otherwise, we count $-\log p(c \mid x)$ for the true class c.

```
> sum(-log(res[cbind(seq(along = truth),
as.numeric(truth))]/nreps))
```

Forensic Glass Example

Learning vector quantization

- For LVQ as for k-nearest neighbour methods we have to select a suitable metric.(The rescaled variables or Mahalanobis distance) code
- Our initialization code in `lvqinit` follows Kohonen's in selecting the number of representatives.

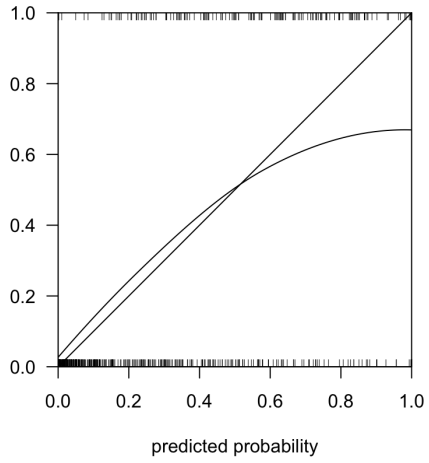
Calibration Plots

Calibration Plots

One measure that a suitable model for $p(c \mid \boldsymbol{x})$ has been found is that the predicted probabilities are well calibrated.

- For **the forensic glass example** we are making six probability forecasts for each case, one for each class. To ensure that they are genuine forecasts, we should use the **cross-validation procedure**.

Calibration Plots



▶ This plot does show substantial over-confidence in the predictions, especially at probabilities dose to one. Indeed, only 22/64 of the events predicted with probability greater than 0.9 occurred.

Calibration Plots

Where calibration plots are **not straight**, the best solution is to find a better model.

- The over-confidence is minor, and mainly attributable to the use of plug-in rather than predictive estimates.
- Then the plot can be used to adjust the probabilities.