Business Statistics

Tommaso Proietti

DEF - Università di Roma 'Tor Vergata'

Classification

Introduction

Classification (discrimination) is the second class of supervised learning problems that we consider.

Our task is to classify an individual into one of several categories on the basis of a set of measurements on that individual.

More formally, given an output variable, denoted by G, taking values in a discrete index set, $\mathcal{G}$, with K classes or categories, we aim at establishing a classification rule which allocates cases to the categories according to the value of X.

A **classifier** is a **prediction rule** that, based on the X's, assigns a response category: we denote it by $\hat{G}(X)$

Example

Consider two response categories: $\mathcal{G}_0 = \texttt{solvent}$, $\mathcal{G}_1 = \texttt{insolvent}$.

We estimate

$$p_k(X) = P(G = k|X), k = 0, 1,$$

on the basis of the training sample and construct the prediction rule

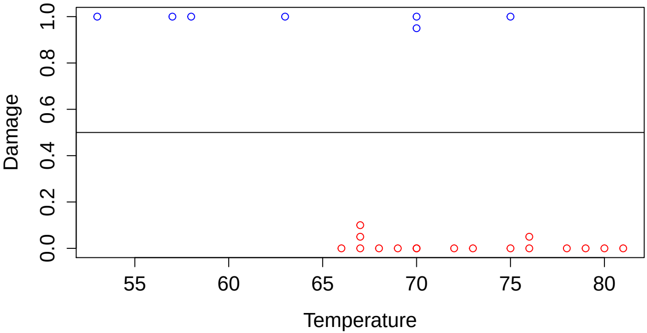
$$\hat{G}(X) = \mathsf{argmax}_k \left\{ \hat{p}_k(X) \right\}.$$

(argmax_k stands for the value k that maximises the function in curly brackets).

The Challenger Disaster

- January 28, 1986: the space shuttle Challenger exploded after take off.
- This was due to a failure of an O-ring seal in the right solid rocket booster (SRB).
- For the previous 24 launches the SRB had been recovered from the ocean and inspected. 7 had incidents of damage to the joints, 16 had no incidents of damage.
- Is 'joint damage' related to the temperature at the time of the launch?
- Temperature on the day of the launch was very low: 29 F.

Figure: The Challenger Disaster data.



Loss functions for Classification

In the linear regression problem for a continuous output we focused on the mean square error (quadratic loss) and derived the optimal predictor $\hat{Y} = \hat{\mathsf{E}}(Y|X)$.

In the classification case, an important LF is the 0-1 Loss:

$$L(G, \hat{G}(X)) = I(G \neq \hat{G}) = \begin{cases} 1, & G \neq \hat{G}, \\ 0, & G = \hat{G}, \end{cases}$$

(i.e. a unit loss is incurred in the case of missclassification).
For a population with two groups, $\mathcal{G} = \{0, 1\}$, the loss function $L(G, \hat{G}(X))$ behaves as follows:

	$\hat{G}(X)$	
G	0	1
0	0	1
1	1	0

Bayes classification rule

The following Bayes classifier is optimal under the 0-1 loss function:

$$\boxed{\hat{G}(X) = \mathcal{G}_k \text{ if } P(G = k|X) \text{ is a maximum for all } k}$$

[a unit should be allocated to the group for which the a posteriori probability is a maximum]

When there are only two classes, $\mathcal{G} = \{0, 1\}$, the Bayes classifier is defined as follows:

$$\hat{G}(x) = \begin{cases} 1, & P(G = 1|X = x) > P(G = 0|X = x) \\ 0, & P(G = 1|X = x) < P(G = 0|X = x) \end{cases}$$

The set of x values for which $P(G = 1|X = x) = P(G = 0|X = x)$ is the *decision boundary*.

Definitions: how good is a classification?

Consider the *confusion* matrix:

	$\hat{G}(X)$ (prediction outcome)	
G (actual value)	0	1
0	True negative (TN)	False positive (FP)
1	False negative (FN)	True positive (TP)

The **true positive rate** (TPR) is defined as

$$P(\hat{G}(X) = 1 | G = 1) = TPR = \frac{TP}{TP + FN}$$

this is also referred to as the **sensitivity** rate.
The **false positive rate** (FPR) is defined as

$$P(\hat{G}(X) = 1 | G = 0) = FPR = \frac{FP}{TN + FP}$$

The **specificity rate** is $P(\hat{G}(X) = 0 | G = 0) = \frac{TN}{TN+FP}$.

The **empirical error rate in the training sample** of size N is

$$\bar{\mathrm{err}} = \frac{1}{N}\sum_{i=1}^{N} I(G_i \neq \hat{G}_i) = \frac{1}{N}(FP + FN)$$

(proportion of missclassified units - **missclassification rate** or error). Our objective is to select the model for which the test sample missclassification error is a minimum.

Overview: methods for classification

There are methods that estimate directly $P(G = k|X)$ (logistic regression).

Others exploit Bayes theorem (discriminant analysis).

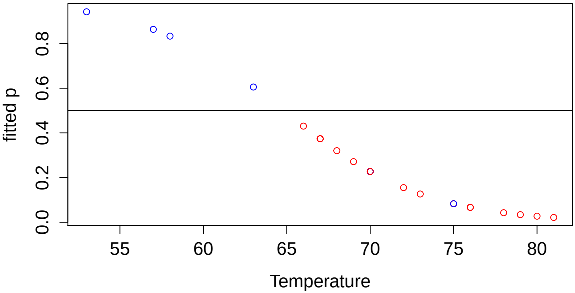
Let

- π_k: prior probability of group k, $\sum_k \pi_k = 1$.
- $f_k(x)$: multivariate density of X in group k.

The posterior probability (Bayes theorem) is

$$P(G = k|X = x) = \frac{P(G = k)f(x|G = k)}{\sum_{j=1}^{K} P(G = j)f(x|G = j)} = \frac{\pi_k f_k(x)}{\sum_{j=1}^{K} \pi_j f_j(x)}$$

Figure: The Challenger Disaster data. The probability of joint damage, $P(G = 1|X)$, is estimated as a function of temperature by a logistic regression model.



Discriminant analysis

We are going to assume that $f_k(\mathbf{x})$ is Gaussian. This is a strong parametric assumption, but it leads to considerable insight and simplification in the form of the decision boundary.

Quadratic Discriminant Analysis

Assume $X|G = k \sim \mathsf{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ so that

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_k|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right\}$$

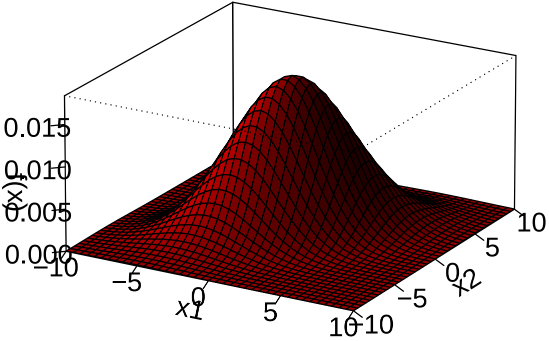
A unit with feature vector $\mathbf{x}$ is allocated to the class for which $P(G = k|\mathbf{x}) \propto \pi_k f_k(\mathbf{x})$, or equivalently its logarithm

$$\ln(\pi_k f_k(\mathbf{x})) = \ln \pi_k - \frac{1}{2}\ln|\boldsymbol{\Sigma}_k| - \frac{1}{2}d(\mathbf{x}, \boldsymbol{\mu}_k; \boldsymbol{\Sigma}_k)$$

is highest.

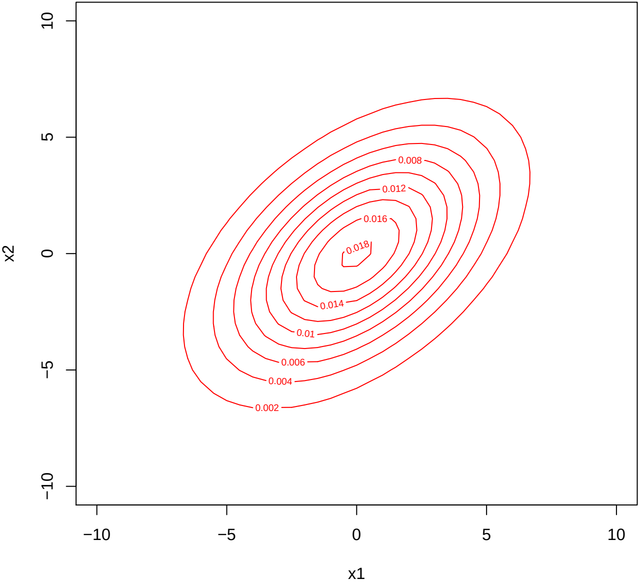
Bivariate Normal Distribution

$\mu_1 = 0, \mu_2 = 0, \sigma_{11} = 10, \sigma_{22} = 10, \sigma_{12} = 15, \rho = 0.5$



Bivariate Normal Distribution

$\mu_1 = 0, \mu_2 = 0, \sigma_{11} = 10, \sigma_{22} = 10, \sigma_{12} = 15, \rho = 0.5$



The component $d(\mathbf{x}, \boldsymbol{\mu}_k; \boldsymbol{\Sigma}_k) = (\mathbf{x} - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)$ is the Mahalanobis distance from the centroid (vector of group means) of the k-th group.
We define $\delta_k(\mathbf{x}) = \ln(\pi_k f_k(\mathbf{x}))$ a *quadratic discriminant function*. The terminology alludes to the fact that the decision boundary between groups k and l, $\{\mathbf{x} : \delta_k(\mathbf{x}) = \delta_l(\mathbf{x})\}$, is a quadratic function of $\mathbf{x}$.

Estimation

From the training sample we compute the variable means in that group, $\hat{\mathbf{x}}_k$, the proportion of cases in group k, and the within group covariance matrix:

$$\hat{\pi}_k = \frac{1}{N}\sum_i I(G_i = k) = \frac{N_k}{N}, \qquad \hat{\mathbf{S}}_k = \frac{1}{N_k}\sum_{i:(G=k)} (\mathbf{x}_i - \hat{\mathbf{x}}_k)(\mathbf{x}_i - \hat{\mathbf{x}}_k)'$$

Hence, the classifier $\hat{G}(X) = \text{argmax}_k\{\delta_k(\mathbf{x})\}$ depends on the prior probabilities, π_k, and the within group covariance. When π_k does not vary with k, $\mathbf{x}$ is allocated to the group to which it is closest, i.e. the *Mahalanobis distance* is a minimum.

Linear Discriminant Analysis

A simplification occurs if $\mathbf{\Sigma}_k = \mathbf{\Sigma}$ for all k. In this case the discriminant function depends on $\mathbf{x}$ only via a linear term:

$$\delta_k(\mathbf{x}) = c + \ln \pi_k + \boldsymbol{\mu}_k' \mathbf{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_k' \mathbf{\Sigma}^{-1} \boldsymbol{\mu}_k$$

The decision boundary between groups k and l is linear in $\mathbf{x}$.

Estimation:

$$\hat{\mathbf{\Sigma}} = \sum_k \frac{N_k}{N} \hat{\mathbf{S}}_k$$

a weighted average of the within group covariance matrices, with weights equal to relative group size.

Example 1: in the single input and 2 groups case, assume $\pi_0 = \pi_1 = 0.5$, and that $X|G = 0 \sim \mathsf{N}(80, 4)$, $X|G = 1 \sim \mathsf{N}(70, 4)$.

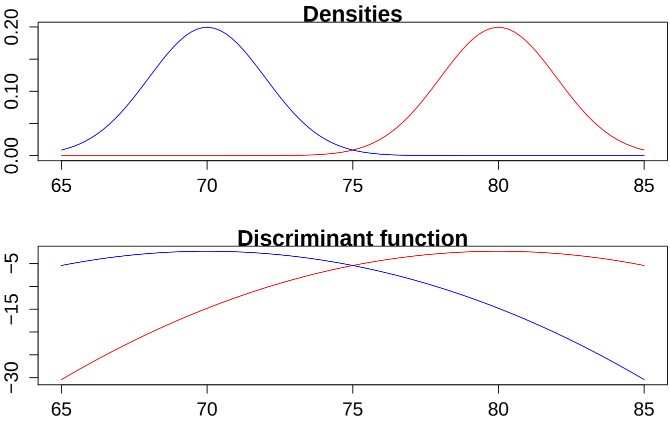
The decision boundary is the point at which $f_1(x) = f_2(x)$, that is $x = 75 = \frac{\mu_0+\mu_1}{2}$. The probability of missclassification is $1 - \Phi\left(\frac{75-70}{2}\right) + \Phi\left(\frac{75-80}{2}\right) = 0.0124$.

Example 2: let $\mathbf{x} = (x_1, x_2)' \in \mathbb{R}^2$, $\mathcal{G} = A, B$.
The decision boundary is the set of points for which $\delta_A(\mathbf{x}) = \delta_B(\mathbf{x})$.
This is the straight line $ax_1 + bx_2 = c$, where

$$c = 0.5(\boldsymbol{\mu}_B'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_B - \boldsymbol{\mu}_A'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_A) + \ln \pi_A - \ln \pi_B$$

$$(a, b) = (\boldsymbol{\mu}_B - \boldsymbol{\mu}_A)'\boldsymbol{\Sigma}^{-1}$$

Figure: Gaussian densities and discriminant function for $\pi_0 = \pi_1 = 0.5$ and $X|G = 0 \sim \mathsf{N}(80, 4)$, $X|G = 1 \sim \mathsf{N}(70, 4)$.



Canonical analysis

Fisher's linear discriminant analysis aims at determining one or more linear combinations of the X variables which maximise the separation between the groups.

These are referred to as *canonical variables*.

Discrimination is useful for classification of a new unit and allocation to one of the groups.

Let $\mathbf{X}$ denote a $N \times p$ matrix. We form a linear combination (a one-dimensional synthesis) of the p variables

$$\mathbf{z} = \mathbf{X}\mathbf{a}, \quad z_i = \mathbf{a}'\mathbf{x}_i$$

where $\mathbf{a}' = (a_1, \ldots, a_p)$.

Denoting by $\bar{z}_k, k = 1 \ldots, K$, the group means of the variable $\mathbf{z}$ and by $\bar{z}$ the overall mean, the deviance of $\mathbf{z}$ can be decomposed as follows:

$$\begin{array}{rcl} T_z(\mathbf{a}) & = & \sum_i (z_i - \bar{z})^2 \\ & = & \sum_{k=1}^K \sum_{i \in \mathcal{G}_k} (z_i - \bar{z}_k)^2 + \sum_{k=1}^K N_k (\bar{z}_k - \bar{z})^2 \\ & = & W_z(\mathbf{a}) + B_z(\mathbf{a}) \end{array}$$

- $W_z(\mathbf{a})$ is the within-group deviance
- $B_z(\mathbf{a})$ is the between-group deviance

The notation stresses the dependence on $\mathbf{a}$.

The vector $\mathbf{a}$ is chosen so as to maximise the between-group deviance (separation between groups), subject to the normalisation constraint $W_z(\mathbf{a}) = 1$.

Classifier. A unit with canonical score z_i is allocated to the group k for which the distance $(z_i - \bar{z}_k)^2$ is a minimum.

(The following algorithmic details can be ignored)
In general, the total deviance matrix $\mathbf{T} = \sum_i (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' = N\mathbf{S}$, can be decomposed as $\mathbf{T} = \mathbf{W} + \mathbf{B}$,

- $\mathbf{W} = \sum_k N_k \mathbf{S}_k$ is the within-group deviance matrix
- $\mathbf{B} = \sum_k N_k (\bar{\mathbf{x}}_k - \bar{\mathbf{x}})(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})'$ is the between-group deviance matrix

The vector $\mathbf{a}$ is determined by the following algorithm:

- Determine the spectral decomposition $\mathbf{W} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}'$
- Form the matrix $\mathbf{W}^{-1/2} = \mathbf{V}\boldsymbol{\Lambda}^{-1/2}$ and the new unit profiles $\mathbf{u}_i = \mathbf{W}^{-1/2}\mathbf{x}_i$, $(\mathbf{U} = \mathbf{X}\mathbf{W}^{-1/2'})$. The new variables are orthogonal and have unit deviance within the groups. The deviance of the $\mathbf{u}_i$ values is $\mathbf{W}^{-1/2}\mathbf{T}\mathbf{W}^{-1/2'} = \mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2'} + \mathbf{I}$.
- Compute the eigenvalues and the eigenvectors of $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2'}$ (the between group deviance of the $\mathbf{U}$ variables).
- Set $\mathbf{a}$ equal to the eigenvector corresponding to its largest eigenvalue.

- The solution amounts to choosing $\mathbf{a}$ so as to maximise $\mathbf{a}'\mathbf{B}\mathbf{a}$ subject to $\mathbf{a}'\mathbf{W}\mathbf{a} = 1$.
- If the number of groups is greater than 2, we can determine other linear combinations, with zero within group correlation, that maximise the separation between the groups. Their coefficients are obtained from the eigenvectors corresponding to the remaining eigenvalues, which are in decreasing order.
- Note: canonical analysis is the same as linear discriminant analysis when all the canonical variates are considered. The Mahalanobis distance in $\delta_k(\mathbf{x})$ becomes an Euclidean distance: $d(\mathbf{x}, \boldsymbol{\mu}_k; \boldsymbol{\Sigma}) = (\mathbf{x} - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k) = (\mathbf{u} - \bar{\mathbf{u}}_k)'(\mathbf{u} - \bar{\mathbf{u}}_k)$.

Iris dataset

150 observations on 4 variables concerning the length and width of sepal and petal for three northern american species of iris: iris setosa, iris versicolor, iris verginica ($N_1 = N_2 = N_3 = 50$).

```
Group means:
           Sepal.Length Sepal.Width Petal.Length Petal.Width
setosa              5.0         3.4          1.5        0.25
versicolor          5.9         2.8          4.3        1.33
virginica           6.6         3.0          5.6        2.03

Coefficients of linear discriminants:
               LD1    LD2
Sepal.Length  0.83  0.024
Sepal.Width   1.53  2.165
Petal.Length -2.20 -0.932
Petal.Width  -2.81  2.839

Proportion of trace:
   LD1    LD2
0.9912 0.0088
```

Figure: Iris dataset.

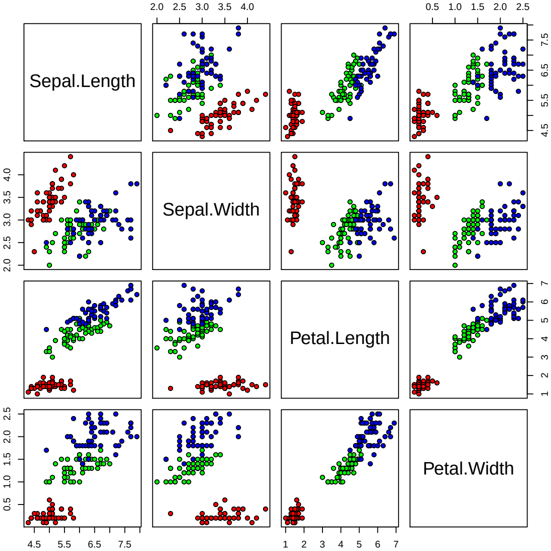
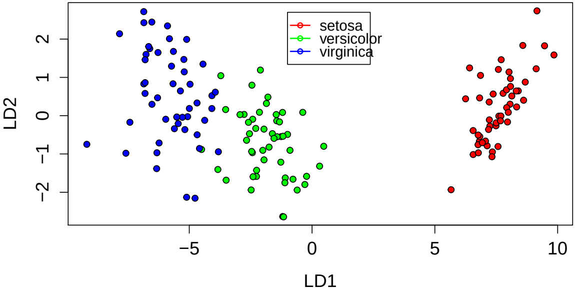


Figure: Plot of canonical variables for Iris data.



Linear Regression Methods for Classification

Consider a categorical response variable G with K categories and let $\mathbf{X}$ an $N \times (p+1)$ data matrix collecting the values of p covariates for N individuals and the intercept in the first column.

We define a set of K indicator variables, Y_k, one for each response category or groups, taking value 1 if $G = k$ and 0 otherwise.

We can form an indicator response matrix $\mathbf{Y}$ $(N \times K)$, such that each row contains the values of the indicator variables $(Y_1, \ldots, Y_K)$ for the i-th unit.

Note that

$$\mathbf{Y}'\mathbf{Y} = \mathbf{N} = \text{diag}(N_1, N_2, \ldots, N_K)$$

where N_k is the number of units in group k (total n. on 1's in column k of matrix $\mathbf{Y}$). We denote the k-th column of $\mathbf{Y}$ by $\mathbf{y}_k$.

$\mathcal{G}_k$	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$
Happy	1	0	0
Happy	1	0	0
Don'k know	0	0	1
Happy	1	0	0
Unhappy	0	1	0
Happy	1	0	0
Unhappy	0	1	0
Unhappy	0	1	0
Unhappy	0	1	0
Don'k know	0	0	1

Linear Regression of an indicator matrix

- Regress $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_K]$ on $\mathbf{X}$ (whose 1st column is the vector $\mathbf{i}_N$) by LS: $\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{Y}$ is the regression matrix. The fitted values are $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{B}$.
- To classify a new observation with features $\mathbf{x}$, compute the $K \times 1$ vector $\hat{y}(\mathbf{x}) = \hat{\mathbf{B}}'\mathbf{x}$ and select the label k for which $\hat{y}_k(\mathbf{x})$ is a maximum.
- This complies with Bayes classification rule, as $\hat{y}_k(\mathbf{x}) = \hat{P}(G = k|X = \mathbf{x})$
- It is true that $\sum_k \hat{y}_k(\mathbf{x}) = 1$, but there is no guarantee that $0 \leq \hat{y}_k(\mathbf{x}) \leq 1$.
- Recall that LS is optimal under squared loss.

If there are only two classes, we just need to regress the indicator for one of the classes, e.g. the 1st, $\mathbf{y}$, on $\mathbf{X}$ and assign an individual to that class if the predicted value is larger than 0.5. $R : \mathbf{x}'\hat{\boldsymbol{\beta}} > 0.5$. The decision boundary is linear.
Notice also that in the case $K = 2$ we can build the variable $\mathbf{y}_1 - \mathbf{y}_2$ or $\mathbf{Y}(1, -1)'$. The decision boundary is then $\mathbf{x}'\hat{\boldsymbol{\beta}} = 0$. LDA yields the same solution.

Logistic Regression

We focus on the case in which G has only two response categories (binary, or dichotomous, variable).

The linear regression model does not make the most efficient use of the information available.

In fact, we know that LS is optimal for a regression model in which the errors ϵ are such that $\mathsf{E}(\epsilon|X) = 0$ and $\mathsf{Var}(\epsilon|X) = \sigma^2$.

It can be shown that when Y is binary the error term is heteroscedastic. Moreover, the predictor $f(X)$ could be outside the theoretical range [0,1].

Specification

We assume that conditional on X, G has a Bernoulli distribution:

$$G = \begin{cases} 0, & \text{with probability } P(G = 0|X = \mathbf{x}) = 1 - p(\mathbf{x}; \boldsymbol{\beta}) \\ 1, & \text{with probability } P(G = 1|X = \mathbf{x}) = p(\mathbf{x}; \boldsymbol{\beta}) \end{cases}$$

so that $\mathsf{E}(G|X) = p(\mathbf{x}; \boldsymbol{\beta})$ and $\mathsf{Var}(G|X) = p(\mathbf{x}; \boldsymbol{\beta})(1 - p(\mathbf{x}; \boldsymbol{\beta}))$ where β is a vector of unknown parameters.

The specification of the model is completed by the assumption that

$$p(\mathbf{x}; \boldsymbol{\beta}) = F(\boldsymbol{\beta}'\mathbf{x})$$

where $F(\cdot)$ is a function taking values in $[0, 1]$.

- The logistic regression model chooses the logistic function for $F(\cdot)$:

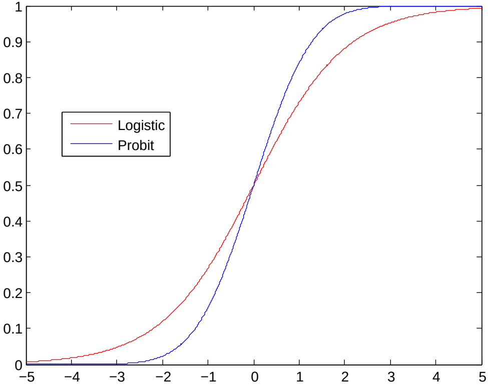
$$p(\mathbf{x};\boldsymbol{\beta}) = \frac{\exp(\boldsymbol{\beta}'\mathbf{x})}{1+\exp(\boldsymbol{\beta}'\mathbf{x})}.$$

- Other choices for F are possible: the Probit model uses the standard normal cumulative distribution function.
- The logistic model is easier to interpret. In particular, the specification implies that the log-odds (logit) is linear:

$$\ln \frac{P(G=1|X=\mathbf{x})}{P(G=0|X=\mathbf{x})} = \ln\left[\frac{p(\mathbf{x};\boldsymbol{\beta})}{1-p(\mathbf{x};\boldsymbol{\beta})}\right] = \boldsymbol{\beta}'\mathbf{x}.$$

(the logit transformation transforms probabilitities in [0,1] into logit scores in $\mathbb{R}$).

Figure: Logistic and Probit link functions.



Training sample

A training sample consisting of N observations, drawn independently from the same population, is available.

We code the two classes by the dichotomous variable Y, taking values 0, if $G = 0$ and 1, if $G = 1$.

The sample is thus $\{(y_i, \mathbf{x}_i), i = 1, \ldots, N\}$.

In the sequel we will denote $p_i = p(\mathbf{x}_i; \boldsymbol{\beta})$.

Estimation

Suppose that the observed sample is $\{(y_1 = 0, \mathbf{x}_1), (y_2 = 1, \mathbf{x}_2), \ldots, (y_N = 0, \mathbf{x}_N)\}$.

The probability of observing this sample (likelihood) implied by our model and by our sampling mechanism (units are drawn independently) is

$$P(y_1 = 0|\mathbf{x}_1)P(y_2 = 1|\mathbf{x}_2) \cdots P(y_N = 0|\mathbf{x}_N) = (1 - p_1)p_2 \cdots (1 - p_N)$$

Writing $P(y_i = k|\mathbf{x}_i) = p_i^{y_i}(1 - p_i)^{1-y_i}$, which is a handy notation for saying that when $y_i = 1$ then we should have p_i, whereas when $y_i = 0$ then we should have $1 - p_i$, the likelihood is defined as the joint probability associated with the observed sample

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{N} p_i^{y_i}(1 - p_i)^{1-y_i}.$$

This is a function of $\boldsymbol{\beta}$.

The log-likelihood is

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{N} [y_i \ln p_i + (1 - y_i) \ln(1 - p_i)]$$

The estimator is computed by the following algorithm (*skip these details*):

1. Start with trial value $\hat{\boldsymbol{\beta}}^{(0)}$ (e.g. $\mathbf{0}$)
2. At iteration $k = 0, 1, \ldots,$ compute $\hat{\boldsymbol{\beta}}^{(k)'}\mathbf{x}_i$ and $\hat{p}_i = \exp(\hat{\boldsymbol{\beta}}^{(k)'}\mathbf{x}_i)/[1 + \exp(\hat{\boldsymbol{\beta}}^{(k)'}\mathbf{x}_i)]$
3. Compute the Pearson residuals

$$r_i = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1 - \hat{p}_i)}}$$

   and rescale the inputs:

$$\mathbf{x}_i^* = \mathbf{x}_i\sqrt{\hat{p}_i(1 - \hat{p}_i)}$$

4. Regress r_i on $\mathbf{x}_i^*$. Let $\boldsymbol{\delta}^{(k)}$ denote the LS estimates.
5. Update $\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} + \boldsymbol{\delta}^{(k)}$.
6. Iterate until convergence, i.e. until $\boldsymbol{\delta}^{(k)} = 0$.

At convergence, r_i is orthogonal to $\mathbf{x}_i^*$ and all the relevant information contained in the inputs has been successfully incorporated.
For an interpretation as iteratively reweighted least squares (IRWL) see e.g. Hastie, Tibshirani and Freedman (2007).

Example: German Credit Data

The `German Credit` data set consists of $N = 1000$ consumers' credits from a southern German bank (source: Fahrmeir and Tutz and http://www.stat.uni-muenchen.de/service/datenarchiv/kredit/kredit_e.html

The output variable is Creditability (Group), (0: credit-worthy, 1: not credit-worthy). 20 inputs were collected. A forward stepwise procedure selects the following inputs

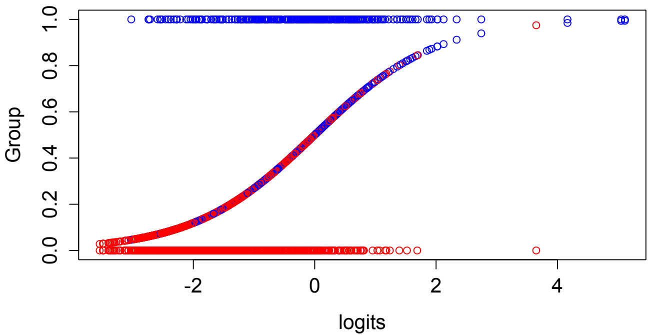
`Duration`	Duration in months (quantitative)
`CreditAmount`	Amount of credit in DM (quantitative)
`StatusCAccount`	Balance of current account (categorical)
`CreditHistory`	Payment of previous credits (categorical)

as well as the square of `CreditAmount` and the interaction of `Duration` and `CreditAmount`

```
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                -6.575e-02  4.588e-01  -0.143 0.886062
Duration                    9.020e-02  1.562e-02   5.776 7.63e-09 ***
CreditAmount               -1.963e-04  9.569e-05  -2.051 0.040247 *
factor(StatusCAccount)A12  -5.438e-01  1.889e-01  -2.879 0.003988 **
factor(StatusCAccount)A13  -1.064e+00  3.394e-01  -3.135 0.001717 **
factor(StatusCAccount)A14  -1.888e+00  2.084e-01  -9.056  < 2e-16 ***
factor(CreditHistory)A31   -2.021e-01  4.839e-01  -0.418 0.676300
factor(CreditHistory)A32   -1.035e+00  3.815e-01  -2.713 0.006674 **
factor(CreditHistory)A33   -9.962e-01  4.417e-01  -2.255 0.024111 *
factor(CreditHistory)A34   -1.631e+00  4.031e-01  -4.046 5.21e-05 ***
I(CreditAmount^2)           4.279e-08  1.012e-08   4.226 2.38e-05 ***
I(Duration * CreditAmount) -1.076e-05  2.777e-06  -3.876 0.000106 ***
---
    Null deviance: 1221.73  on 999  degrees of freedom
Residual deviance:  996.76  on 988  degrees of freedom
AIC: 1020.8

Confusion matrix
    FALSE TRUE
  0   636   64
  1   176  124
```

Figure: kredit dataset. Plot of y_i and $\hat{p}_i$ versus the logits $\mathbf{x}_i'\hat{\boldsymbol{\beta}}$



Diagnostic checking, hypothesis testing and goodness of fit

The large sample distribution of $\hat{\boldsymbol{\beta}}$ is normal with mean $\boldsymbol{\beta}$ and covariance matrix $(\mathbf{X}^{*'}\mathbf{X}^{*})^{-1}$, where $\mathbf{X}^{*}$ is the matrix of the rescaled inputs (with i-th row $\mathbf{x}_i^{*'}$).

The square root of the diagonal element is the standard error and $z_k = \hat{\beta}_k / st.err(\hat{\beta}_k)$ (the z-value) is the test statistic for the null that the k-th coefficient is 0.

Its square is the Wald test for the same null (chi-squared distribution).

Diagnostic checking is carried out on the Pearson residual r_i.
The Pearson Statistic

$$\chi^2 = \sum_{i=1}^{N} r_i^2$$

is the main g.o.f. statistic.

The deviance residual, d_i, is the signed square root of $-2\left[y_i \ln \hat{p}_i + (1 - y_i) \ln(1 - \hat{p}_i)\right].$

The deviance is

$$D = -2[\ell(\hat{\boldsymbol{\beta}})] = \sum_i d_i^2$$

(the sum of squares of the deviance residuals).

The null deviance $D_0 = -2\ell_0$ is the model with $\beta_1 = \cdots = \beta_p = 0$ (only the intercept is fitted, so that $\hat{p} = N_1/N$ and $\ell_0 = N_1 \ln(N_1/N) + N_0 \ln(N_0/N)$).

HTF define the training error $\bar{\mathrm{err}} = -\frac{2}{N}\ell(\hat{\boldsymbol{\beta}}) = D/N$.

The proportion of units missclassified when the Bayes classifier is adopted is the measure of training error consistent with the 0-1 loss. The classifier is $\hat{G}(\mathbf{x}) = 1$ if $\hat{\boldsymbol{\beta}}'\mathbf{x} > 0$, because this implies $P(G = 1|\mathbf{x}) > 0.5$.

Model selection criteria

$$AIC = -2\frac{1}{N}\ell(\hat{\boldsymbol{\beta}}) + 2\frac{p}{N}$$

$$BIC = -2\frac{1}{N}\ell(\hat{\boldsymbol{\beta}}) + \ln(N)\frac{p}{N}$$

(note: the null model always features the intercept, and thus the d.f. are p)

Relation with LDA

In the case of only two classes the log-posterior odds is linear in $\mathbf{x}$ for both methods.
The difference arise in the way the coefficients are estimated. Logistic regression leaves the distribution of X unrestricted, and bases the estimated coefficients on the likelihood conditional on X, whereas LDA assumes that it is normal.
Note that the latter assumption is untenable if X includes dummy variables.