Statistical Learning

Tommaso Proietti

DEF Tor Vergata

Classification

Introduction

- *Classification* (discrimination) is the second class of supervised learning problems that we consider.
- Our task is to classify an individual (unit) 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

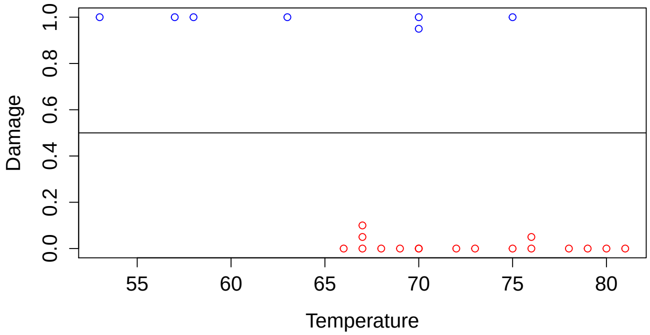
$$\hat{G}(X) = \text{argmax}_k \left\{\hat{p}_k(X)\right\}.$$

(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 the indicator variable of 'joint damage' related to the temperature at the time of the launch?
- Temperature on the day of the launch was very low: 29 F.

The Challenger Disaster data. Plot of the indicator variable of a joint damage vs temperature.



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 $\hat{Y} = \hat{\mathrm{E}}(Y|X)$.

In the classification case, an important loss function 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

- What is the optimal classification rule if we face a constant loss for a missclassfication?
- 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*.

Overview: methods for classification

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

Others exploit Bayes theorem (discriminant analysis).

Let

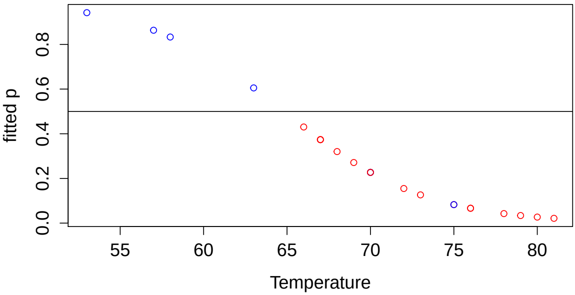
- π_k: prior probability of group k, $\sum_k \pi_k = 1$.
- $f_k(x)$: multivariate density of X in group k.

The posterior probability (Bayes theorem) is

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

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.

Figure



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

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

(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.

Discriminant analysis

Recall that the posterior probability, by Bayes theorem, is

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

We are going to assume that π_k is given and $f_k(\boldsymbol{x})$ 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

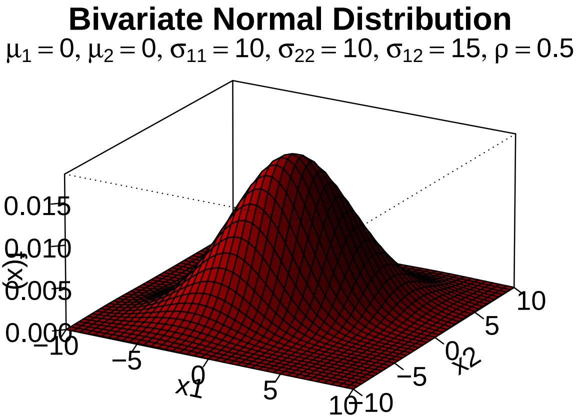
Assume $X|G=k\sim \mathrm{N}(\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$ so that

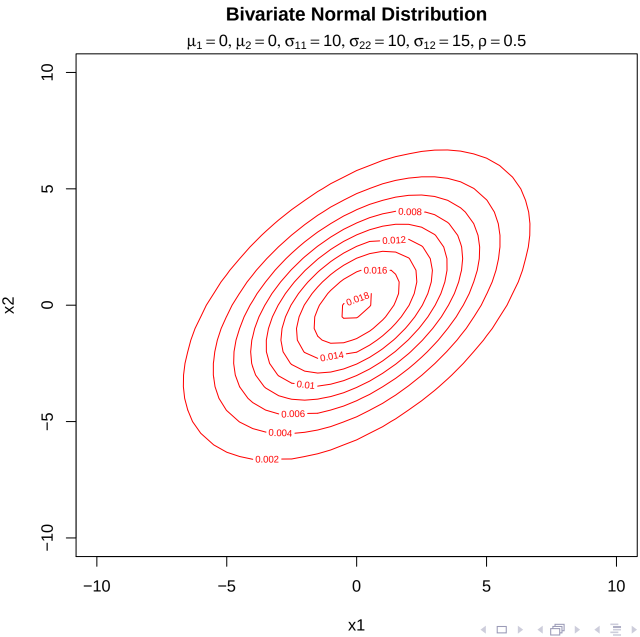
$$f_k(\boldsymbol{x})=\frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_k|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right\}$$

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

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

is highest.


Bivariate Normal Distribution
$\mu_1 = 0, \mu_2 = 0, \sigma_{11} = 10, \sigma_{22} = 10, \sigma_{12} = 15, \rho = 0.5$
0.015
0.010
0.005
0.000
−10
−5
0
5
10
x1
x2


Bivariate Normal Distribution
μ1 = 0, μ2 = 0, σ11 = 10, σ22 = 10, σ12 = 15, ρ = 0.5
x1
x2
−10
−5
0
5
10
0.002
0.004
0.006
0.008
0.01
0.012
0.014
0.016
0.018

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

Estimation

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

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

Hence, the classifier $\hat{G}(X) = \text{argmax}_k \{\delta_k(\boldsymbol{x})\}$ depends on the prior probabilities, π_k, and the within group covariance. When π_k does not vary with k, $\boldsymbol{x}$ 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 $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$ for all k. In this case the discriminant function depends on $\boldsymbol{x}$ only via a linear term.
- The decision boundary between groups k and l is linear in $\boldsymbol{x}$.

Example 1: in the single input and 2 groups case, assume that the prior probabilities are

$$\pi_0 = \pi_1 = 0.5,$$

(we call this a diffuse prior) and that

$$X|G = 0 \sim \mathrm{N}(80, 4), \qquad X|G = 1 \sim \mathrm{N}(70, 4).$$

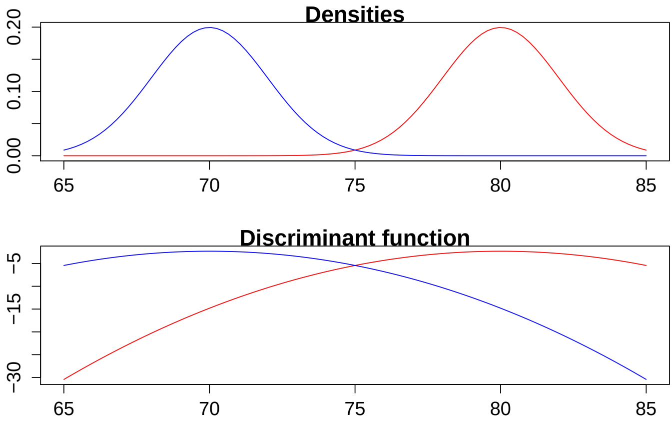
- The decision boundary is the point at which $f_1(x) = f_2(x)$, that is $x = 75 = \frac{\mu_0 + \mu_1}{2}$.
- The probability of missclassification is

$$\begin{aligned} P(X < 75|G = 0) + P(X > 75|G = 1) &= \Phi\left(\tfrac{75-80}{2}\right) + \left(1 - \Phi\left(\tfrac{75-70}{2}\right)\right) \\ &= 0.0124, \end{aligned}$$

where $\Phi(z) = P(Z < z)$ for $Z \sim \mathrm{N}(0, 1)$, the c.d.f. of a standard normal r.v.

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

Figure



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 $\mathrm{E}(\epsilon|X) = 0$ and $\mathrm{Var}(\epsilon|X) = \sigma^2$.
- 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:

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

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

The specification of the model is completed by the assumption that

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

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

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

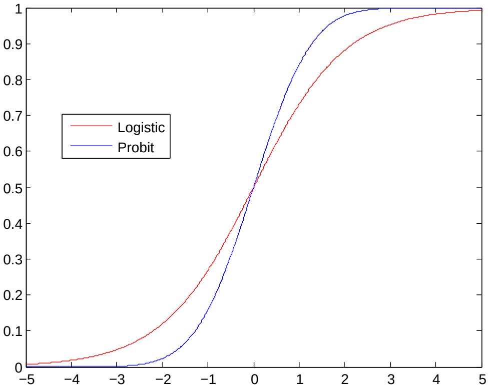
$$p(\boldsymbol{x};\boldsymbol{\beta}) = \frac{\exp(\boldsymbol{\beta}'\boldsymbol{x})}{1+\exp(\boldsymbol{\beta}'\boldsymbol{x})}.$$

- 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:

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

(the logit transformation transforms probabilities 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 $\{(y_i, \boldsymbol{x}_i), i = 1, \ldots, N\}$.

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

Estimation

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

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

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

Writing $P(y_i = k|\boldsymbol{x}_i) = p_i^{y_i}(1 - p_i)^{1-y_i}$, 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 maximum likelihood estimator of $\boldsymbol{\beta}$ is the value of $\boldsymbol{\beta}$ that maximises $\ell(\boldsymbol{\beta})$ (or equivalently $L(\boldsymbol{\beta})$).

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

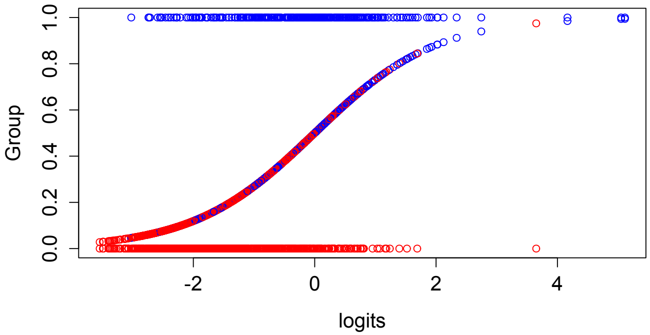
`Duration`	Duration in months (quantitative)
`CreditAmount`	Amount of credit in DM (quantitative)
`StatusCAccount`	Balance of current account (categorical)
`CreditHistory`	Payment of previous credits (categorical)

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 $\hat{\beta}' \boldsymbol{x}_i'$



Diagnostic checking, hypothesis testing and goodness of fit

The ML estimate of the k-th coefficient, scaled by its standard error, $z_k = \hat{\beta}_k / st.err(\hat{\beta}_k)$ (the z-value), provides a 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 by the Pearson residuals

$$r_i = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}, \quad i = 1, \ldots, N.$$

The Pearson Statistic

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

can be used to assess the goodness of fit.

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

A measure of the training error is $\overline{\text{err}} = -\frac{2}{N}\ell(\hat{\boldsymbol{\beta}}) = D/N$.

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 $\hat{G}(\boldsymbol{x}) = 1$ if $\hat{\boldsymbol{\beta}}'\boldsymbol{x} > 0$, because this implies $P(G=1|\boldsymbol{x}) > 0.5$.

Model selection criteria

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

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

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

Bayes' Theorem

Let $\{A_1, A_2, \ldots, A_m\}$ be a collection of events performing a *partition* of Ω (the events A_i are disjoint, $A_i \cap A_j = \emptyset$ for $i \neq j$, and their union is the entire sample space, $\cup_{i=1}^{m} A_i = \Omega$). Let B denote an event with $P(B) > 0$.

According to the **law of total probability** (LTP) the probability of B can be computed as follows:

$$P(B) = \sum_{i=1}^{m} P(B \cap A_i) = \sum_{i=1}^{m} P(B|A_i)P(A_i)$$

Proof: write

$$\begin{aligned} B &= B \cap \Omega \\ &= B \cap (A_1 \cup A_2 \cup \cdots \cup A_m) \\ &= (B \cap A_1) \cup (B \cap A_2) \cup \cdots \cup (B \cap A_m) \end{aligned}$$

The events $(B \cap A_i), i = 1, \ldots, m$, are disjoint. Applying the third axiom and the multiplication rule $P(B \cap A_i) = P(B|A_i)P(A_i)$, the result follows.

- Bayes' theorem is a fundamental result for statistical learning.
- Consider a particular event A_j of the partition $\Omega = \cup_{i=1}^{m} A_i$.
- Let $P(A_j)$ be its prior probability (i.e., its marginal probability regardless of B occurring or not). Then, the posterior probability of the event A_j, given knowledge that another event B has occurred, is proportional to the product of the prior probability and $P(B|A_j)$, the likelihood of the event B if A_j had occurred:
$$P(A_j|B) = \frac{P(A_j)P(B|A_j)}{\sum_i P(A_i)P(B|A_i)}$$
- The proof is simple: recall $P(A_j|B) = P(A_j \cup B)/P(B)$. Replace $P(A_j \cup B) = P(B|A_j)P(A_j)$ in the numerator and $P(B) = \sum_{i=1}^{m} P(B|A_i)P(A_i)$ by the LTP.

Figure: Reverend Thomas Bayes (1702-1761)

Example: credit scoring

Let

$$
\begin{array}{lcl}
A & = & \{ \text{ the client is credit-worthy} \} \\
\bar{A} & = & \{ \text{ the client is not credit-worthy } \} \\
B & = & \{ \text{ the financial situation is good} \} \\
\bar{B} & = & \{ \text{ the financial situation is not good } \}
\end{array}
$$

Prior probability: $P(A) = 0.70$ (elicitable from our previous records)
Likelihoods: $P(B|A) = 0.95$, $P(B|\bar{A}) = 0.10$.
By the LTP,

$$P(B) = P(A)P(B|A) + P(\bar{A})P(B|\bar{A}) = 0.70 \times 0.95 + 0.30 \times 0.10 = 0.695.$$

By Bayes's theorem:

$$
\begin{array}{lcl}
P(A|B) & = & \frac{P(A)P(B|A)}{P(A)P(B|A)+P(\bar{A})P(B|\bar{A})} \\
 & = & \frac{0.70\times 0.95}{0.70\times 0.95+0.30\times 0.10} \\
 & = & 0.957
\end{array}
$$