(PART) Part: Classification

Overview on classification problems

Some examples

Banks and credit

- A financial institution such as a bank lends money to profit from it: the interest paid by the debtor in exchange for the unavailability of the money for the duration of the loan.
- The creditor also assumes the risk that the debtor will not return (in bad faith or because he is not able) the loan sum.
- When a person asks for a loan, it is necessary to have a **criterion to decide** whether it is convenient to grant it or not.
- This criterion will **ideally refer to the probability** that the subject will return The loan, probability **that can depend on various factors** related to the future cash availability of the subject itself.

- We are assuming that the bank would prefer objective rules to case-by-case assessment made by credit line managers, this or
  - because the requests are too many to be examined in detail,
  - or to provide a guideline to managers
  - and leave no margin of arbitrariness
  - or to keep under check the overall risk.

- We distinguish **two classes** of (aspiring) debtors: worthy of credit and not worthy of credit
- We also have **data** on the membership to each of the classes for n subjects, for each of which we also know other characteristics (variables, in statistical language) such as employment, age, current account balance etc.
- The class has the role of dependent variable or **response**, the other variables are independent or **explanatory** variables.
- The bank is interested
  - in **predicting** the class of subjects for whom only the explanatory variables are known
  - the **interpretation** of the model, i.e. the reading of what are the variables that most influence the membership of one or another class .

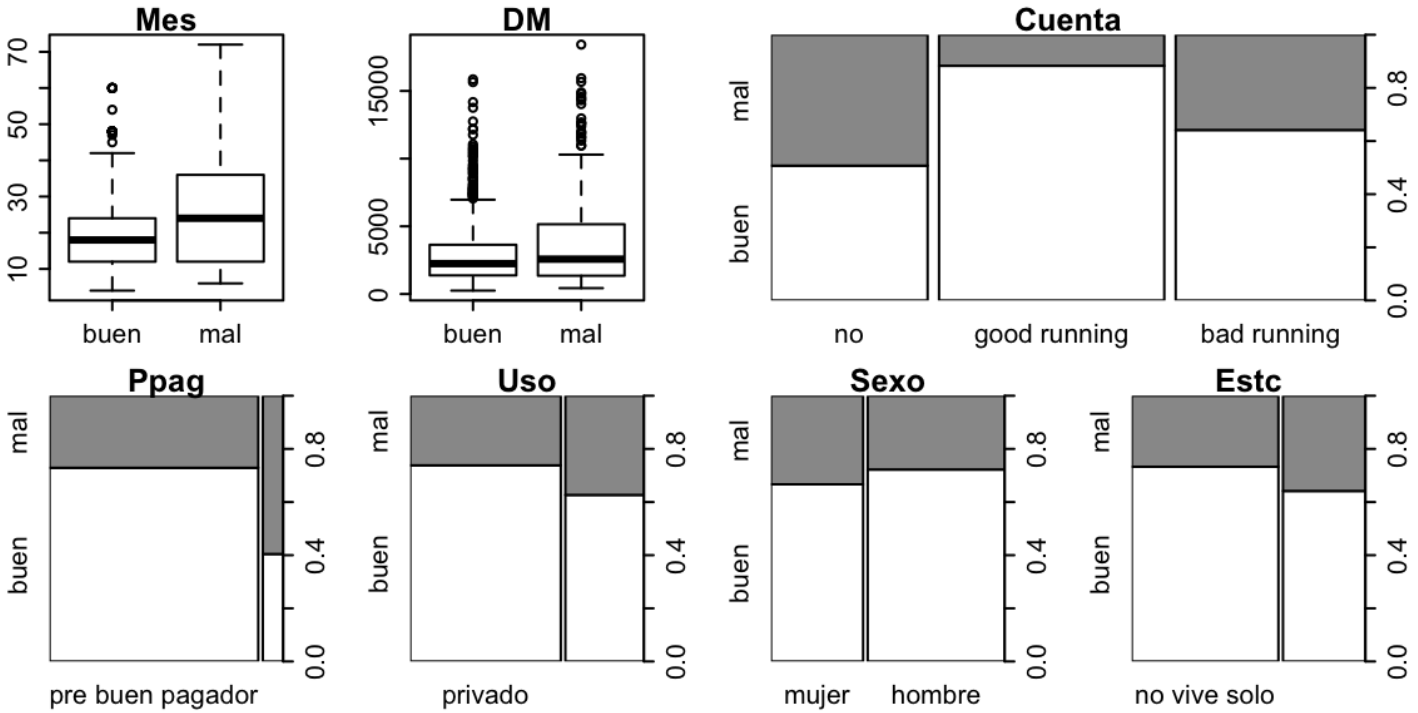
- This type of application is a typical example of **asymmetry** in the importance of the errors that can be made:
  - assigning to credit-worthy class someone who then turns out to be non-solvent (it's more expensive and therefore a more serious error than)
  - assigning to bad class someone who deserves credit (you lose a profit opportunity and maybe damage the credit institution's image)

Credit Score Data From a South German Bank

Data for 1000 clients of a south german bank, 700 good payers and 300 bad payers. They are used to construct a credit scoring method (from *Multivariate Statistical Modelling Based on Generalized Linear Models* by L. Fahrmeir and G. Tutz).

```
library(Fahrmeir)
data(credit)
```

	Y	Cuenta	Mes	Ppag	Uso	DM	Sexo	Estc
1	buen	no	18	pre buen pagador	privado	1049	mujer	vive solo
2	buen	no	9	pre buen pagador	profesional	2799	hombre	no vive solo
3	buen	bad running	12	pre buen pagador	profesional	841	mujer	vive solo
998	mal	good running	21	pre buen pagador	profesional	12680	hombre	no vive solo
999	mal	bad running	12	pre buen pagador	privado	6468	hombre	no vive solo
1000	mal	no	30	pre buen pagador	privado	6350	hombre	no vive solo


Mes
10
30
50
70
buen
mal
DM
0
5000
15000
buen
mal
Cuenta
mal
buen
0.0
0.4
0.8
no
good running
bad running
Ppag
mal
buen
0.0
0.4
0.8
pre buen pagador
Uso
mal
buen
0.0
0.4
0.8
privado
Sexo
mal
buen
0.0
0.4
0.8
mujer
hombre
Estc
mal
buen
0.0
0.4
0.8
no vive solo

Other examples

- Spam or non-spam: build an automatic rule that allows you to classify messages like spam or non-spamt, based on mail features such as frequence of special characters, of capital letters, of certain words, the source address, etc.

- Character recognition: reading characters, digits, punctuation–without the need for human intervention– on the basis of the number and the position of pixels occupied by the printed character.

- Diagnostic programs: build rules to decide whether or not a patient suffers from a particular disease or, more generally, to decide which illness he sugffers from, based on the symptoms and / or other physiological information (heartbeat, pressure, etc.) of the patient.

- Marketing: given a number of brands of products (cars, drinks, etc.), investigate whether similar groups can be recognized among these for product (physical, organoleptic. etc) characteristics.

This case is different from the previous ones because there are no objectively predetermined classes, indeed, it is not even said that it makes sense to divide the product set into classes.

Classification

Classification: what do we mean?

- The general task of a classification procedure is to **divide data into classes (groups)**.
- More in detail, the task is, given a set of individuals (statistical units) for which certain characteristics (variables) are known, to build a criterion for assigning each individual to a group **on the basis of known features**.

- We distinguish two cases according to the type of groups:
  - whether the fact that individuals can be subdivided into groups is pre-experimental, we refer to as **supervised classification**.
  - if the existence of a 'sensible' partition of individuals into groups is not obvious, but, indeed, one of the purposes of the investigation is to decide whether such groups exist, we refer to as unsupervised classification or grouping or **cluster analysis**.

Data type: supervised case

We have, for each of n individuals,

- the observations for p variables,
- the class of membership exactly determined by one **supervisor**. (We suppose class of an individual can be observed directly.)

$$(\mathbf{y}, \mathbf{X}) = \begin{bmatrix} y_1 & x_{11} & x_{12} & \dots & x_{1p} \\ y_2 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ y_n & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} = (\mathbf{y}, \mathbf{x_1}, \mathbf{x_2}, \dots, \mathbf{x_p})$$

- x_{ij} is the observation for the j-th variable on the i-th individual
- $y_i \in \mathcal{G} = (G_1, \dots, G_g)$

Example

Consider a sample of four individuals for which we observe age, sex, and title of study.

Individual	Employed	Gender	Age	Education
1	employed	M	27	degree
2	employed	W	31	degree
3	employed	W	21	secondary
4	employed	M	19	secondary

What is $\mathbf{y}$? $\mathbf{x_2}$? X? And $\mathcal{G}$?

Motivation: classification

- Supervised classification is applied when the *supervision* is an inadequate method of classifying new units: then a decision criterion based on the other observable characteristics is sought. .
- The inadequacy of 'supervision' may refer to different circumstances:
  - in the diagnosis of illness when supervision requires an examination post mortem;
  - in the optical recognition of the characters because you just want to cut off the supervisor;
  - in credit risk wherever you want to determin risk before lending money.
- In many cases, besides the **purpose of prediction**—of replacing the supervisor, there may be a **descriptive purpose** (the classification results indicate which variables separate the groups more).

Motivation: clustering

- Unsupervised classification is more ambitious. In fact, it tries to **determine a partition* of individuals into groups based on the observed characteristics
  - **without knowing whether a partition into groups is meaningful**,
  - hence, without knowing the number of groups.
- For example, companies use it for market **segmentation**, to distinguish between different types of customers (eg, to send targeted ads).

Methods

- Methods used to build a classification criterion are of different type and it is not easy to give a description. In simple terms, they try to summarize the explanatory variables so as to enhance those features which differentiate more the membership to a group rather than to another.
- As for the supervised classification
  - widely used: **logistic regression**, linear discriminant analysis (**LDA**), **K-nearest neighbours** more computer intensive: **GAM**, **trees**, **random forests**, **boosting**, **SVM**
- As for the unsupervised classification
  - **cluster analysis**

Output

- classification analysis
  - a **partition of the space** generated by the p explanatory variables
  - a **rule that associates each constituent with a class**
  - this can be not necessarily true when
    - we consider classes with non null intersection
    - we look for a hierarchy (taxonomy)
    - we decide to provide information on not only whether or not you belong to a class, but also the degree of membership
- cluster analysis
  - a **partition of the space** generated by the p explanatory variables (which, in the case of non-supervised classification, is the sample space)

In synthesis, in case of a supervised classification, if the classes are $(G_1, \ldots, G_g)$, a classification rule can be expressed by

$$R_1, \ldots, R_g$$

where

- $R_1, \ldots, R_g$ partition of $\mathcal{X}$
- $R_i = \{\mathbf{x} \in \mathcal{X} \text{ such that } \mathbf{x} \to G_i\}$

Assessment of classification

As we did for linear models, we aim at selecting the best subset of variables (which need be included by an opportune transformation in the regression model).

- LIFT and ROC curves
- validation
- cross-validation

A naif method

We can reformulate the classification problem by saying that

- we have a variable y to explain in function of X,
- the target is to predict the value of y `by making a wrong prediction as little as possible '.

So we are in the reference frame of the **linear models**:

$$y = X\beta + \epsilon$$

We assume that

- all x variables have been transformed into quantitative variables,
- y has been coded as $0/1$.

We predict

$$\hat{y} = X\hat{\beta}$$

We can bild a rule as follows:

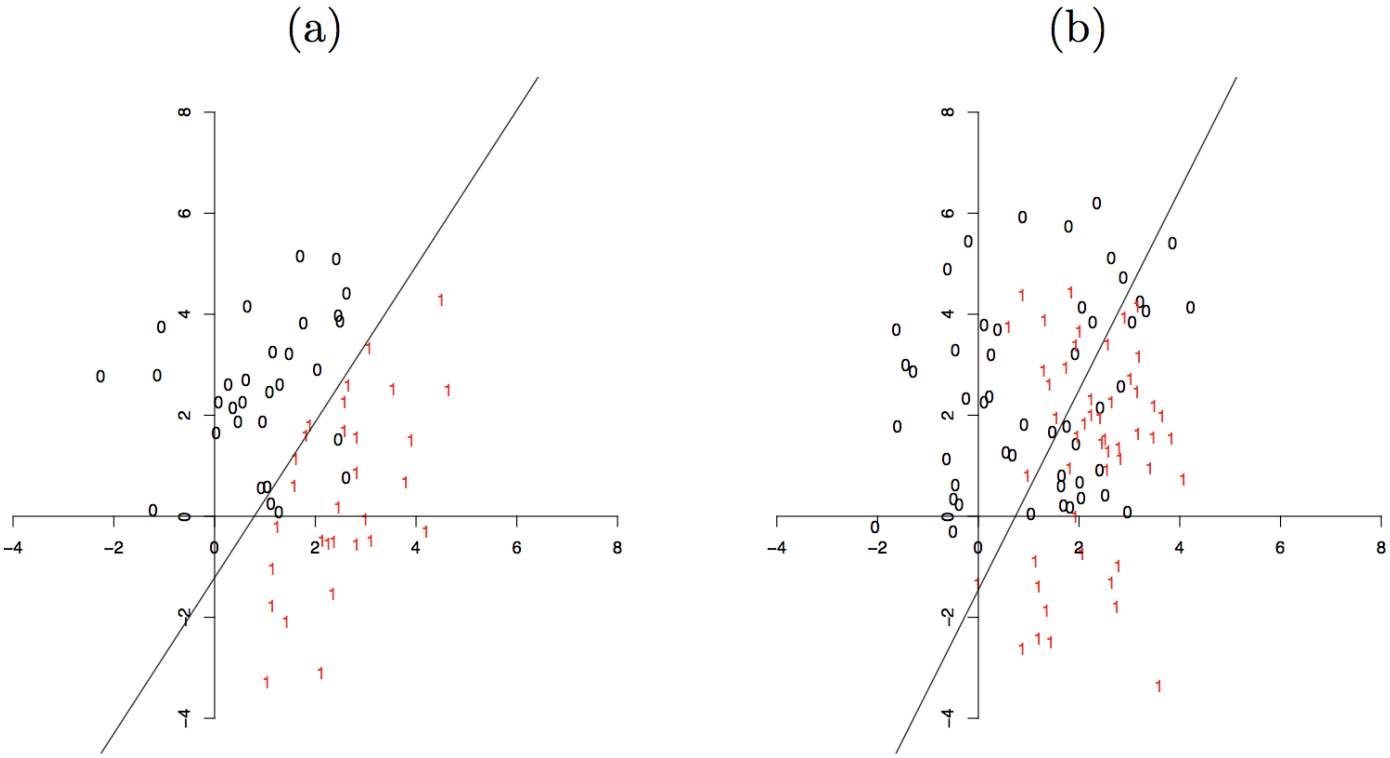
$$\tilde{y}_i = \begin{cases} 0 & \text{if} \quad \hat{y}_i < 0.5 \\ 1 & \text{if} \quad \hat{y}_i \geq 0.5 \end{cases}$$

Or

$$\tilde{y}_i = 1 \ \text{ if and only if } \ 0.5 \leq \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{ip}$$

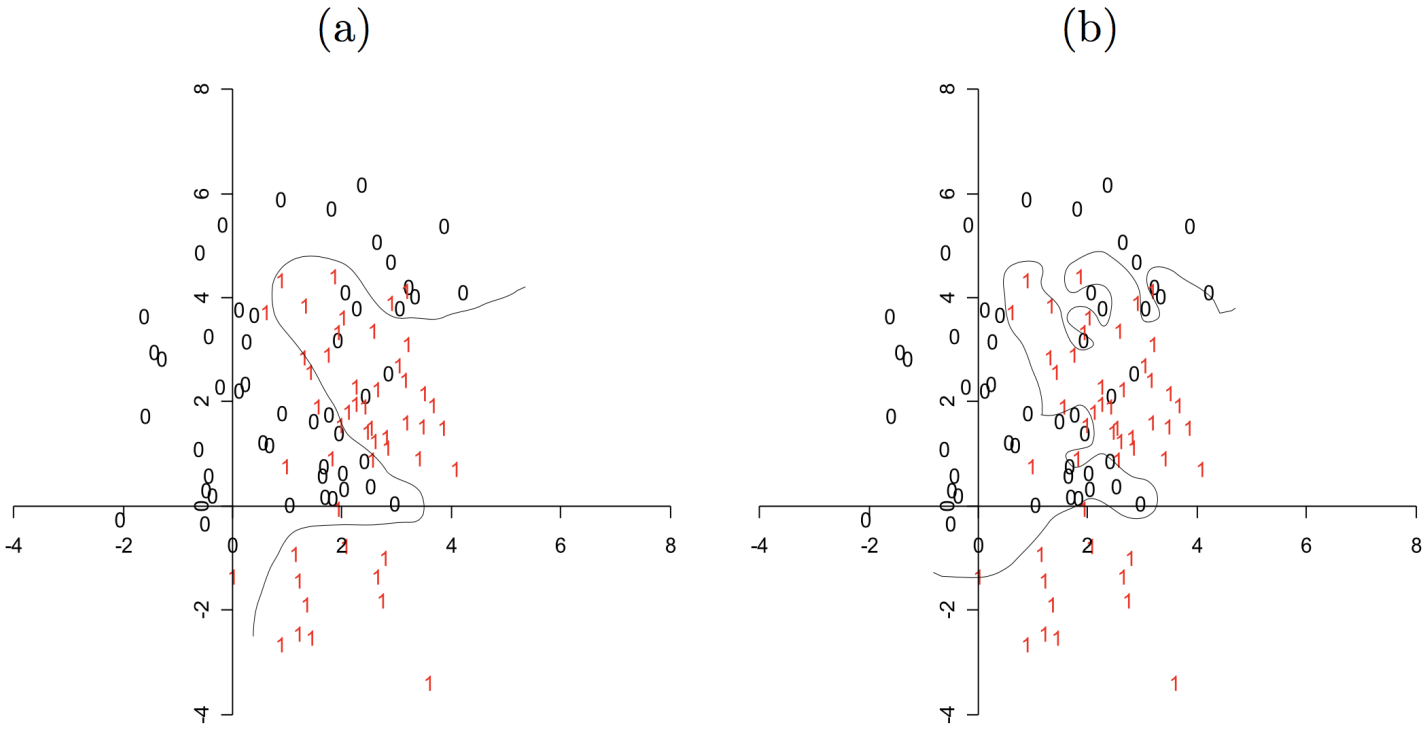
i.e. $\Re^p$ space is partitioned into two half-spaces, the common frontier is the hyperplane $\hat{y}_i = \mathbf{x}_i\hat{\beta}$.

Two examples with $p = 2$



Output of classification analysis by a naif method based on a linear model.

Example (b)



Two more classifiers for the second example (at right).

Generalizability and precision

- Regularity of the frontier *vs* precision
- Generalizability of the rule *vs* empirical precicion

Empirical estimate of classification error

		Classe prevista (G_j)				
	$i \backslash j$	1	2	$\dots$	g	
	1	n_{11}	n_{12}		n_{1g}	$n_{1.}$
Classe reale (G_i)	2	n_{21}	n_{22}		$\vdots$	$\vdots$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	g	n_{g1}	n_{g2}		n_{gg}	$n_{g.}$
		$n_{.1}$	$n_{.2}$		$n_{.g}$	$n_{..}$

Confusion matrix.

Probability of misclassifying an individual of class i:

$$\hat{e}(G_i) = \sum_{j \neq i} \hat{P}(\hat{G}_j | G_i) = \sum_{j \neq i} \frac{n_{ij}}{n_{i.}} = 1 - \frac{n_{ii}}{n_{i.}}$$

Thus, the empirical classification error is:

$$\hat{e}_{\text{app}} = \sum_{i=1}^{g} \hat{e}(G_i)\hat{P}(G_i) = \sum_{i=1}^{g} \hat{e}(G_i)\frac{n_{i.}}{n} = \frac{1}{n}\sum_{i=1}^{g}(n_{i.} - n_{ii})$$

that is an underestimate of the classification error.

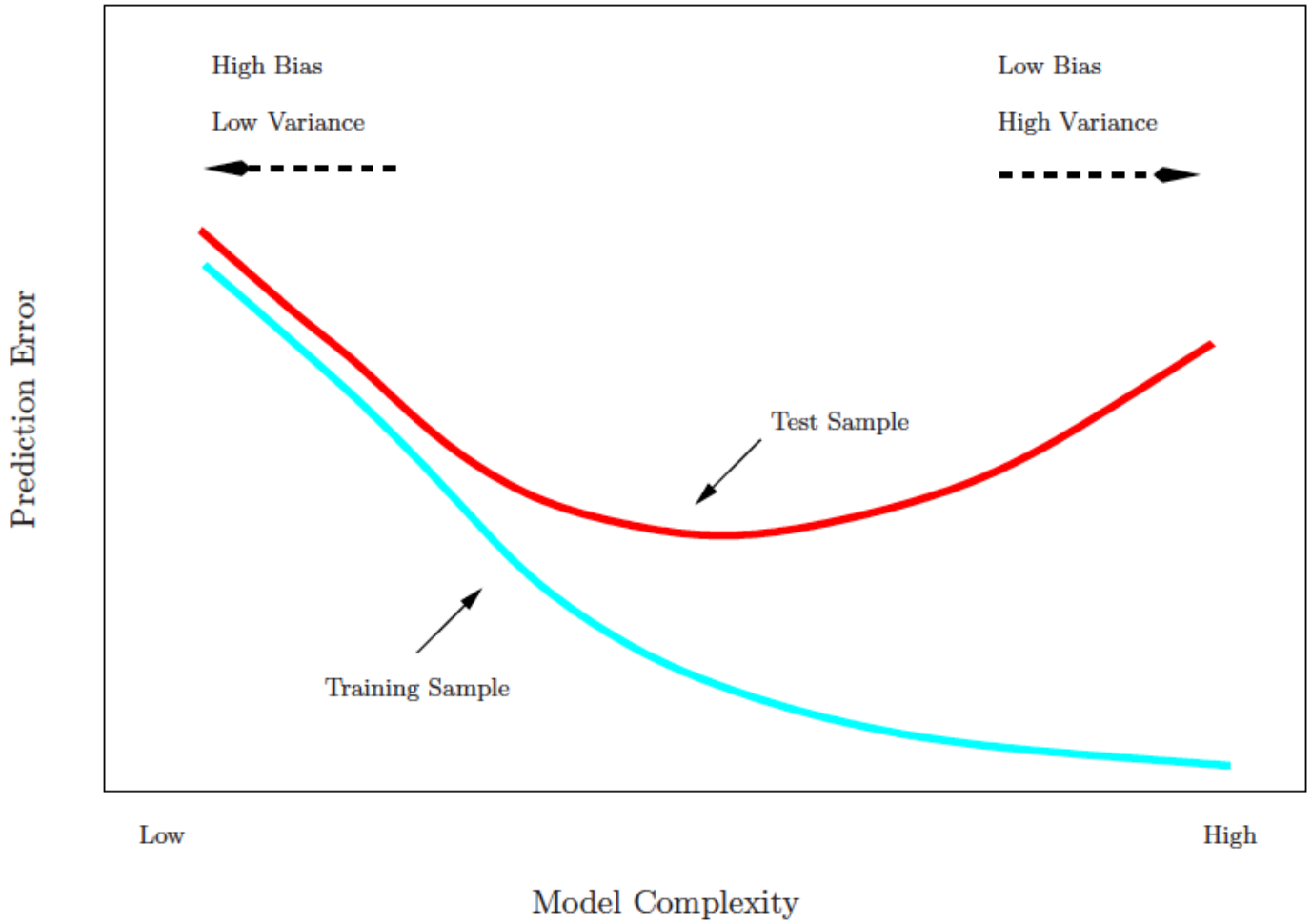
- Remark: it is sensible to assume that an estimate of the prior probability of G_i is $\frac{n_{i.}}{n}$.

Validation

- validation-set approach
- k-fold CV

The result can be shown by a **confusion matrix** and misclassification rates can be estimated by using the formulas above, but, this time, using the sole *test set*.

Classification error: training versus test set



Training- versus Test-Set Performance.