12

Data Mining

Overview. *Support vector machines are often used in data mining. In this case, SVMs are only one part of a complex process. This chapter describes a general data mining strategy and explains which role SVMs can play within this process. Some competitors of SVMs and software tools for data mining are briefly described.*

Prerequisites. *Basic knowledge on SVMs from Chapter 1.*

Support vector machines and other kernel-based techniques treated in this monograph have two main areas of application: risk minimization in machine learning and data mining. In this chapter, we describe the data mining process and explain the role of SVMs in this process.

The goal in standard areas of statistical machine learning is to minimize the empirical risk. Here it is not of primary importance to extract knowledge about the internal structure of the data set as long as the empirical risk is minimized. A typical example is the automatic classification of incoming emails as “no spam” or “spam”. Other examples are detection of credit card fraud and the automatic recognition of hand-written digits. SVMs are often successfully applied in these cases, although the data analyst usually has no or only vague prior information about the probability distribution P that generated the data set. In the former chapters of this monograph, we gave a theoretical foundation of why SVMs based on appropriate choices of the loss function and the kernel and using suitable hyperparameters are able to learn, which makes SVMs especially valuable for these standard areas.

Another field where SVMs and related kernel methods are successfully applied is data mining projects, where these methods are one—but only one—cornerstone of the whole process. The main goal of data mining projects is generally to extract and model formerly unknown information contained usually in large and complex data sets. It seems unrealistic to hope that the application of SVMs can really be successful in data mining without some knowledge about the general data mining process. Therefore, this chapter gives a short overview of data mining and describes the main phases in such projects.

In Section 12.1, we give a definition of data mining and explain why data mining is important. In Section 12.2, we describe a general data mining strategy called CRISP-DM, which was proposed by Chapman *et al.* (2000). CRISP-DM is the abbreviation of CRoss-Industry Standard Process for Data

Mining. In Section 12.3, we explain the role of SVMs as one part in the whole data mining process. Section 12.4 mentions a few software tools for data mining. Section 12.5 contains information about further literature, and Section 12.6 gives a summary of this chapter.

12.1 Introduction

Hand *et al.* (2001, p. 1) define data mining in the following way.

> *Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.*

This definition contains several keywords. Data mining deals with *observational data sets*, which means that the data set is generally not collected only for the purpose of the data mining project. Furthermore, observational data are generally not random samples but often quite large and a number of cases between 10^5 and 10^7 is not unusual. The owner of such a data set sometimes has prior information about the data or the data-generating process such that the goal is to *extract new information* from the data. The kind of information desired grossly differs between data mining projects. Examples are a model with low prediction error, the identification, of high-risk subgroups or the detection of dependencies between attributes. The result of a data mining project is not only the extraction of new information but also to make the result applicable in practice. The information obtained should therefore be summarized in a way that offers high interpretability both from a business and a mathematical point of view. Thus, one can argue that data mining is more than the application of modeling techniques either from parametric statistics, semi-parametric statistics, or non-parametric statistical machine learning theory to large data sets. Some examples of data mining projects are:

- customer relationship management (CRM): customer acquisition, customer assessment, and customer churn analysis
- eCommerce: prediction of sales and detection of associations between customers and products
- text mining and web mining
- credit risk scoring: banking
- insurance tariffs and identification of high-risk subgroups
- analysis of gene expression data: microarray experiments

In the next section, we describe one particular strategy for data mining.

12.2 CRISP-DM Strategy

The CRISP-DM project (Chapman *et al.*, 2000) has developed an industry- and tool-neutral data mining process model. CRISP-DM is the abbreviation of CRoss-Industry Standard Process for Data Mining and was developed by DaimlerChrysler AG (Germany), Teradata being a subdivision of NCR Systems Engineering Copenhagen (USA and Denmark), OHRA Verzekeringen en Bank Groep B.V. (The Netherlands), and the statistical software company SPSS®(USA). The project was partially funded by the European Commission under the ESPRIT program. Starting from the knowledge discovery processes used in industry today and responding directly to user requirements, this project defined and validated a data mining process that is applicable in diverse business sectors. The goal of CRISP-DM is to make large data mining projects faster, cheaper, more reliable, and more manageable.

The CRISP-DM strategy consists of the six main phases shown in Figure 12.1 and described in the rest of this section.

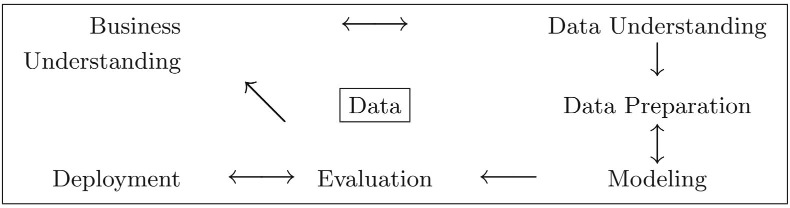

Fig. 12.1. Main phases of data mining according to the CRISP-DM strategy.

Phase 1: Business Understanding

The aim in this phase is the determination of the objectives from a *business perspective*. Often the customer has several competing objectives and constraints that must be balanced. The analyst has the goal of uncovering important but unknown factors that can influence the final outcome. A description is usually given of any solution currently in use for the problem together with a list of their advantages and disadvantages. Further, the *business success criteria* for a successful or at least useful outcome to the project are determined from an applied point of view. Some examples for such criteria are the following:

- Estimate the probability of an event (e.g., a genomic defect) given a list of inputs.
- Determine an insurance tariff that minimizes the risk (i.e., the expectation of the losses).

- Give useful insights as to which customers produce the most expensive claims for an insurance company.
- Improve the response rate in a direct marketing campaign by at least 5 percent.
- Identify the subgroup of patients having a certain type of cancer who will have the highest benefit from a new drug.

At this early stage of a data mining project, one assesses the current situation by considering all resources, constraints, and assumptions that should be considered in the determination of the data analysis goal and by the project plan. One should take into account the available or necessary personnel, type of data files, computing resources, available data mining software tools, and other relevant software. All requirements of the project, including scheduled date of completion, comprehensibility, and quality or precision of the results are listed. Of course, one should make sure that access to the data files is allowed and possible from a technical point of view. All the project-specific assumptions should be listed, no matter whether they can be checked during the data mining process or not. Assumptions are made for the precision of the estimates. The constraints made on the project are listed (e.g., lack of resources to carry out some tasks within the timescale or legal or ethical constraints). An additional aspect is the determination of lower bounds on the required sample size such that the conclusions can be made with a desired precision. All assumptions on external factors such as competitive products or technical advances should be listed. A decision should be made whether the final model should be interpretable in business terminology or not because this can easily influence the choice of the modeling technique. The results of SVMs for example are often harder to interpret from a business point of view than for parametric models such as generalized linear models; see Table 12.2. Additionally, the starting point and the endpoint of the project are listed together with possible risks depending on the size of the data or the data quality.

A cost versus benefit analysis for the data mining project is prepared that compares the cost of the data mining project with the potential benefit to the business. Of course, a data mining project will in general only be done if the potential benefit dominates the cost.

Then the *data mining success criteria* are determined in statistical terms. All business questions are translated into *data mining goals*. For example, a direct-marketing campaign requires segmentation of customers in order to decide who to approach in the campaign, and one should also specify the size of the segments. During this phase, one specifies the type of data mining problem; e.g., classification (see Chapters 5 and 8) or regression (see Chapter 9). Additional criteria for model assessment are specified such as model accuracy, performance, and complexity. Furthermore, benchmarks for the evaluation criteria are determined. An example is the level of the predictive accuracy.

Then a *project plan* is made that lists all stages of the data mining project. The project plan should include duration, required resources, inputs, outputs, and dependencies. Dependencies between the time schedule and risks are described. Recommendations for actions are also given for the case where risks appear. The project plan also contains a list of people, specifying who is responsible for which steps. From our point of view, it is essential for a successful data mining project to take the following rule of thumb into account during the construction of the project plan:

50%–70% of the time and effort for the data preparation phase;
15%–25% for the data understanding phase;
10%–20% for the business understanding phase, modeling, and evaluation;
5%–10% for the deployment phase.

Phase 2: Data Understanding

The first task in this phase is the *collection of initial data* listed in the project resources. Therefore, a list of necessary data sets or databases and their types is constructed. Further, the software tools and the methods to acquire them are listed. If problems are encountered, they should also be listed.

A *data description report* is made that describes the main properties of the data, including

- quantity of data (d : number of variables or attributes, n : number of cases or records);
- format of the data;
- coding, percentage, and patterns of missing values;
- identifier variables needed for merging data from different databases or tables;
- time period when the data were collected.

Then a *data exploration report* is made that describes the distribution of the key attribute(s); e.g., the main response variables (or target attribute) of a prediction problem. A list of possible values and a contingency table are given for categorical variables. For continuous variables, some descriptive statistics are listed; e.g., minimum and maximum values, mean and standard deviation, or their robust pendants, such as median and median absolute deviation (MAD). Furthermore, low-dimensional relationships and dependencies between pairs or a small number of attributes are computed taking the scales (nominal, ordinal, continuous) of the attributes into account. This report also describes properties of interesting subpopulations for further examination; e.g., stratification by gender, age, or geographical region.

Additionally, the *data quality report* lists the results of the data quality verification. It also mentions possible solutions for the case of quality problems. Such solutions often depend on deep knowledge of the business and of the data itself. Many authors argue that the data quality is essential for success in data mining projects; see, e.g., Hipp *et al.* (2001).

Phase 3: Data Preparation

The goal of this phase is to obtain clean data sets or databases that can be used in the modeling phase. First, *data selection* is done by deciding which attributes will be included or excluded in the next phases. The selection covers the selection of attributes (columns) and the choice of cases (rows). Possible criteria for the decisions are the relevance to the data mining goals, the percentage of missing values, and the data quality.

Then a *data cleaning report* is made that describes the actions to increase the data quality. The report describes which actions were taken to overcome the data quality problems. If missing values are replaced by imputation methods or other strategies (see Rubin, 1987), the report should describe which methods were used and how many data points were modified.

The *construction of the clean data sets* includes data preparation operations such as correction of typing errors and the transformation of existing attributes (by using logarithms, square roots, Box-Cox transformations, indicator variables, etc.) and by defining derived attributes. As an example, we mention the definition of the body mass index (BMI) for adults, which is calculated by the following metric formula:

$$\mathrm{BMI} = \frac{\text{height (in meters)}}{\text{weight}^2\text{(in kilograms)}} \,.$$

These BMI values are usually classified into groups such as “underweight”: BMI below 18.5; “normal”: BMI between 18.5 and 24.9; etc. There exist modified formulas to compute the BMI for children and teens, taking gender and age into account. The goals of derived attributes such as the body mass index are twofold: a reduction of dimensionality and ease of interpretation.

The next step of the data preparation phase is the *integration of data.* This is generally needed because data from multiple tables or data sets have to be merged together or new cases must be created about the same object or person. Sometimes merged data also cover aggregations that are operations to summarize information from multiple cases or tables. As an example, we mention a data mining project for analyzing insurance data. Assume that there are three tables partially due to data security:

- Table A, with a unique index variable, say ID, and personal and demographic information about the customers;
- Table B, containing the ID variable, the number of claims, the claim amount for the current year, and possible explanatory variables (inputs);
- Table C, containing the ID variable and the claim history covering the last decade.

Here the ID variable is needed to merge the data belonging to a single customer together into one large table. A useful aggregation step in this example is the construction of new variables for the sum of years without a claim, the total number of claims per year, and the average claim amount per year for each customer.

The final step in this phase is *formatting the data*, which is generally required by the modeling tool and to increase the readability of the results. It is often useful to format the values of categorical attributes such that the preferred reference class is the first class or the last class. It can be helpful to convert text attributes into uppercase after trimming blanks.

Phase 4: Modeling

First, one has to *select the modeling technique(s)*, taking the data mining goals, the properties of the data, and the plausibility of model assumptions into account. One advantage of kernel methods including SVMs is that these non-parametric methods only need rather weak assumptions in comparison with parametric methods. Support vector machines are of course only one class of modeling techniques used in data mining projects, and we would like to mention three strong competitors: generalized linear models, generalized additive models, and tree-based methods. These methods are implemented in several data mining tools.

Generalized Linear Models

A generalized linear model (GLIM) is a parametric regression model having three components: a stochastic component, a linear predictor, and a link function; see Nelder and Wedderburn (1972), McCullagh and Nelder (1989), and Fahrmeir and Kaufmann (1985). The *stochastic component* assumes that the $(d+1)$-dimensional random variables (X_i, Y_i), $i = 1, \dots, n$, are stochastically independent and that the conditional distribution of Y_i given $X_i = x_i$ is an element of an exponential family (see (A.17)). The *linear predictor* describes the assumption that the vector x influences Y only via the linear combination $\eta = x^\mathsf{T}\theta$. The bijective *link function* g connects the linear predictor to the conditional expectation $\mu(\theta) := \mathbb{E}_{\mathrm{P}_\theta}(Y|x)$ via $g(\mu(\theta)) = \eta$. GLIMs make implicitly a hard additional assumption, namely that there is a functional relationship specified by some fixed variance function $v : \mathbb{R} \to [0, \infty)$ between the conditional expectation and the conditional variance of Y given x. Table 12.1 lists the conditional distribution of Y given x, the link function, and the variance function of special GLIMs. Note that the variance function is a polynomial in these cases. This simple relationship between expectation and variance of the response variable can be grossly violated in practice, see Christmann (2005) for a data set from insurance companies.

The classical estimator of $\theta \in \mathbb{R}^d$ is the maximum likelihood (ML) estimator, which has nice properties (consistency in probability or almost sure, convergence rate of $n^{-1/2}$, asymptotic efficiency, and asymptotic normality) if the assumptions of the generalized linear model are satisfied; see Fahrmeir and Kaufmann (1985, 1986). These asymptotic properties of the ML estimator

Table 12.1. Important special cases of GLIMs. Here Λ and Φ denote the cumulative distribution functions of the standard logistic and standard Gaussian distributions, respectively.

Model	Distribution of $Y\|x$	Link Function	Variance Function
Linear regression	Normal	identity: $\eta = \mu$	$v(\mu) = 1$
Logistic regression	Binomial	logit: $\eta = \Lambda^{-1}(\mu)$	$v(\mu) = \mu(1-\mu)$
Probit regression	Binomial	probit: $\eta = \Phi^{-1}(\mu)$	$v(\mu) = \mu(1-\mu)$
Poisson regression	Poisson	$\eta = \ln(\mu)$	$v(\mu) = \mu$
Gamma regression	Gamma	$\eta = 1/\mu$	$v(\mu) = \mu^2$
Gamma regression	Gamma	$\eta = \ln(\mu)$	$v(\mu) = \mu^2$
Inverse Gaussian regression	Inverse Gaussian	$\eta = \mu^{-2}$	$v(\mu) = \mu^3$
Negative Binomial regression	Negative Binomial	$\eta = \ln(\mu)$	$v(\mu) = \mu + k\mu^2$

allow the construction of asymptotically optimal hypothesis tests and confidence regions for β. However, this estimator can have two serious drawbacks. It may not exist for some data sets; see Albert and Anderson (1984) and Santner and Duffy (1986). Furthermore, the maximum likelihood estimator is non-robust in several special cases, including linear regression and logistic regression. In a rather informal way, one can describe robust methods by the property that small violations of the model assumptions should have only a small and bounded impact on the result; see Chapter 10 for details. Robust alternatives to the maximum likelihood estimator were proposed for example by Rousseeuw (1984) and Rousseeuw and Yohai (1984) for linear regression, and Künsch *et al.* (1989) and Christmann (1994, 1998) for generalized linear models; see Chapter 10.

Generalized Additive Models

A generalized additive model (GAM) is a semi-parametric regression model having three components: a stochastic component, an additive predictor, and a link function; see Hastie and Tibshirani (1990). In contrast to the linear predictor $\eta = x^{\mathsf{T}}\beta$ used by GLIMs, a GAM allows the explanatory variables $x = (x_1, \ldots, x_\ell, \ldots, x_d) \in \mathbb{R}^d$ to influence the conditional distribution $Y|x$ via the more flexible way

$$\eta = \alpha + (x_1, \ldots, x_\ell)^{\mathsf{T}}\beta + \sum_{j=\ell+1}^{d} f_j(x_j)\,, \tag{12.1}$$

where the intercept term $\alpha \in \mathbb{R}$, the slope parameters $\beta \in \mathbb{R}^\ell$, and the functions $f_j : \mathbb{R} \to \mathbb{R}$, $j = \ell+1, \ldots, d$, $0 \le \ell \le d$, are unknown and must be estimated from the data. Linear regression methods, including parametric splines and Fourier fits or univariate regression smoothers such as local polynomial regression, are used to estimate the unknown functions f_j, $j = \ell + 1, \ldots, d$. The backfitting algorithm is often used to fit the unknown functions in an iterative manner; see Hastie *et al.* (2001, p. 260). Second-order interactions (i.e., $f_{j_1,j_2}(x_{j_1}, x_{j_2})$ with $\ell < j_1 < j_2 \leq d$) can be included by using surface smoothers, but such smoothers increase the computational effort. It can be quite hard to model higher-order interactions using such smoothers in a nonparametric way with generalized additive models due to the curse of high dimensionality. SVMs can be useful here to fit the non-parametric part of GAMs, including higher-order interactions.

Tree-Based Methods

Finally, we mention tree-based methods as important alternatives to support vector machines; see Breiman *et al.* (1984) and Hastie *et al.* (2001). Special cases are classification trees and regression trees. The basic principle of tree-based methods is quite different from those of GLIMs and GAMs. Trees partition the space X defined by the input variables in a recursive manner to maximize a score of class purity, which means that the majority of the data points in each cell of the partition belong to one class for the case of classification trees. The construction of the tree begins with the whole data set (i.e., all data points are in the same cell). A tree-based method then computes the best split or partition consisting of ℓ cells of the whole data set such that the data points in each cell are significantly more homogeneous than data points of different cells. The cells are called *nodes.* A binary splitting rule ($\ell = 2$) is often preferred to a multi-way splitting rule ($\ell > 2$), which may fragment the data too quickly, leaving insufficient data at the next stage. Then the procedure is repeated: each node is again split into ℓ subnodes. This recursive technique is repeated until a user-defined stopping rule is satisfied. Finally, a tree constructed in this way is usually truncated to the most important splits and the corresponding nodes. This step is called *pruning.* Tree-based methods differ with respect to the splitting, stopping, and pruning rules. Trees often perform well for low-dimensional data sets and the computation time is often short, but trees can have problems in modeling complex high-dimensional dependency structures.

In phase 4 of CRISP-DM, one has to decide which modeling techniques are appropriate for the project. This decision is often non-trivial, because many aspects should be taken into account. A—partially subjective—comparison of advantages and disadvantages of GLIMs, GAMs, trees, and SVMs is given in Table 12.2. Data quality can play a major role in data mining projects, as was explained above. Furthermore, poor data quality makes it hard to justify hard parametric model assumptions, and large databases often contain a small amount of typing errors, different types of outliers, and data points measured in an imprecise manner. Hence, knowledge from the first three data mining phases is also useful to decide whether robustness aspects are important for the specific data mining project.

Table 12.2. Properties of several modeling techniques (mod.: moderate; ML: maximum likelihood estimation; robust: robust estimation).

Characteristic	GLIM (ML/robust)	GAM	Trees	SVMs
Type of model	parametric	semi-parametric	non-parametric	non-parametric
Predictive power	bad/mod.	mod.	bad	good
Ability to extract linear combinations of features	good	good	bad	good
Ability to detect complex dependencies	bad	mod.	good	good
Natural handling of data of mixed type*	mod.	mod.	good	mod.
Handling of missing values	bad	bad	good	mod.
Interpretability	good	good	mod.	low–mod.
Dependency on hyper-parameters	no/yes	yes	yes	yes
Robustness w.r.t. outliers in inputs	bad/good	bad	mod.–good	mod.–good
Robustness w.r.t. outliers in outputs	bad/good	bad	mod.–good	mod.–good
Insensitive to monotone transformations of inputs	bad	mod.	mod.–good	mod.
Computation time	good	mod.	mod.	mod.
Computational scalability (large n)	good	mod.	mod.	mod.
Availability in data mining tools	good	mod.	good	mod. (increasing)

* Nominal, ordinal, continuous.

Generating a test design is the next step in phase 4 of CRISP-DM. A description of the intended plan should be given to ensure that the criteria to measure the empirical risk or the goodness of the predictions are fair and that the data set was not overfitted. A common practice in data mining projects is to split the data set randomly into three parts called the training data set, validation data set, and test data set. The training data set is usually modeled several times with the same modeling technique but with different sets of so-called hyperparameters. As an example, we mention the support vector regression: the hyperparameters are $(\epsilon, \gamma, \lambda) \in (0, \infty)^3$ if the ϵ-insensitive loss function $L_{\epsilon\text{-insens}}$, the Gaussian RBF kernel $k_{\text{RBF}}(x, x') = \exp(-\gamma^{-2}\|x-x'\|_2^2)$, $x, x' \in \mathbb{R}^d$, and the regularizing constant λ are used; see Chapters 4 and 9. The validation data set is used for fine-tuning purposes. The models learned from the training data set are applied to the validation data set to obtain an optimal setting of the hyperparameters yielding the best properties of the modeling technique for the validation data set. The test data set is necessary to obtain a fair measure of how well the modeling technique actually works for *unseen* data points never used before.

Additionally, the procedure should be described as how to divide the whole data into these disjoint parts (e.g., by simple random sampling without replacement or proportional stratified random sampling based on a certain stratification attribute). Proportional stratified random sampling involves dividing the whole data set into homogeneous and disjoint subgroups defined by the values of the stratification attribute and then taking a simple random sample in each subgroup. Which attribute is appropriate for stratification purposes depends on the data mining project. A common rule in stratified random sampling is that data points in the same stratum should be more similar to each other with respect to the response variable than data points of different strata. Often a stratification by gender, age group, geographical region, or a combination of these variables is useful.

There are several reasons why a stratified sampling may be preferable over simple random sampling; see Cochran (1977) or Levy and Lemeshow (1999). It assures that results can be obtained not only for the overall data set but also for interesting subgroups of the data set. This can be quite important, for example, to assure that the training, validation, and the test data sets all contain a reasonable number of people for interesting minority groups. If one data mining goal is to obtain new knowledge about subgroups, this may be the only way to effectively assure that this goal can be achieved. Finally, if one of the subgroups is extremely small, one can use different sampling fractions within the different strata to randomly oversample the small group, which results in non-proportional stratified random sampling. It is of course necessary in this situation to weight the within-strata estimates depending on the sampling fraction whenever overall population estimates are needed.

Cross-validation can be an alternative to the splitting approach described above especially for data sets of only moderate size.

In the *model building step*, the modeling methods are run on the prepared and carefully cleaned data sets to obtain the following outputs:

- The best choice of the *hyperparameters* that was detected for the validation data set is listed. This is done for each modeling method that depends on such parameters.
- The *fitted models* are given using these hyperparameters together with predictions $\hat{y}_i$ for all data points in the test data set.

- A *model description* of the fitted models is also given. The description contains, for example, the risk evaluated for the test data set and a corresponding list of significant explanatory variables.

The final step within this CRISP-DM phase is the *model assessment*. The statistician has to interpret the models, taking into account the data mining success criteria, domain knowledge, and the desired test design. A ranking of the models is useful according to the formerly specified evaluation criteria such as accuracy or generality of the model if different models were fitted.

It is not unusual for the hyperparameters to be revised several times due to fine-tuning of the model. This results in an iteration of the model building step and the model assessment step until no essential improvement can be made. It is recommended to document all such revisions and assessments for future phases and for future data mining projects of similar type.

Phase 5: Evaluation

This phase starts with the *evaluation of the results*. The statistician generally discusses the results with business analysts and domain experts to ensure that not only the narrower technical success criteria treated in phase 4 are (hopefully) met but also the business success criteria. One goal is to determine whether there exists a business-relevant reason why the "optimal" model obtained during the former phase is deficient. Furthermore, the model can be evaluated to check whether the test application meets all budgets or time constraints.

The *review process* summarizes the whole data mining process so far and describes the unexpected findings, missed activities, and overlooked potential risk factors and lists actions that should be repeated.

Finally a decision is made as how to proceed. Are the data mining results sufficiently worthwhile to move to the deployment phase, or should the project stop?

Phase 6: Deployment

If the data mining results were strong enough to justify deployment into the business, a *deployment plan* is made. It describes a strategy for how the relevant findings of the data mining project can become part of the business solution used in practice. A *monitoring plan* and a *maintenance plan* can help to avoid unnecessarily long periods of wrong or suboptimal usage of the data mining results. Of course, *final reports* and a *final presentation* are made at the end of the data mining project. A final *project review* is also useful to document experience that can be important for further data mining projects, such as pitfalls, misleading or encouraging approaches, and problems with certain software tools.

12.3 Role of SVMs in Data Mining

It is obvious from the description of the typical phases in the data mining process given in Section 12.2 that SVMs and related kernel methods are of interest mostly in the modeling phase (i.e., in phase 4 of CRISP-DM). We refer to Chapter 8 for classification, and Chapter 9 for regression problems. However, there is ongoing research on how to use SVMs in other phases, too.

In the data preparation phase (phase 3), kernel feature extraction by kernel PCA can be quite useful to reduce the dimensionality of the inputs and to construct derived attributes. We refer to Schölkopf and Smola (2002), Shawe-Taylor and Cristianini (2004), and the references cited in these books for details on kernel PCA.

Pelckmans *et al.* (2005) proposed kernel-based classifiers using a worst-case analysis of a finite set of observations, including missing values of the inputs. The approach is based on a component-wise SVM and an empirical measure of maximal variation of the components to bound the influence of the components that cannot be evaluated due to missing values.

12.4 Software Tools for Data Mining

There are many commercial and non-commercial software tools for data mining. No attempt is made to mention all of them.

`CART`® and `TreeNet`®

The commercial software tools `CART`® and `TreeNet` are distributed by Salford Systems. `CART`® is a well-known classification and regression tree package. `TreeNet`® is based on boosted decision trees. It is an accurate model builder and can also serve as a powerful initial data exploration tool. Both software tools are more specialized to only a few methods than the other software products listed below but have proven to give good results in several competitions on knowledge discovery and data mining.

`IBM`® `DB2 Intelligent Miner for Data`

The `IBM`® `DB2 Intelligent Miner` is based on a client-server architecture. The server executes mining and processing functions and can host mining results. The client is powered with administrative and visualization tools. Hence, the client can be used to visually build a data mining operation, execute it on the server, and have the results returned for visualization and further analysis. In addition, the application programming interface (API) provides `C++` classes and methods as well as `C` structures and functions for application programmers. This allows the user to define and use new methods. The Intelligent Miner can read data stored by `Oracle`, `SAS`®, and `SPSS`® and contains scalable algorithms.

`SAS`® `Enterprise Miner`™

The `SAS Enterprise Miner` is one of the leading commercial software tools for data mining. SAS propagates the so-called SEMMA strategy which is similar to the CRISP-DM strategy described in Section 12.2. SEMMA is the abbreviation of Sample, Explore, Modify, Model, and Assess. The `SAS`® `Enterprise Miner`™ has a user-friendly graphical user interface (GUI) and contains tools for all necessary steps such as data pre-processing, sampling, classification, regression, trees, clustering, association rules, visualization, assessment, and scoring. Additionally, this software allows user-defined models and ensemble techniques, making it a flexible data mining tool. The `SAS`® `Enterprise Miner`™ runs stably and is capable of handling huge data sets efficiently because it is delivered as a distributed client-server system. To our knowledge there (currently) exists only a non-productive version of an SVM implementation in the `SAS Enterprise Miner`™.

`SPSS`® `Clementine`

The commercial data mining tool `SPSS`®`Clementine` has a structure similar to the `SAS`®`Enterprise Miner`™ and has a user-friendly GUI. New versions of `Clementine` also use a client-server architecture. However, `Clementine` is—from our point of view— less flexible and offers fewer analytical methods than `SAS`® `Enterprise Miner`™.

`Weka`

`Weka` is a collection of machine learning algorithms for data mining tasks; see Witten and Frank (2005). The algorithms can either be applied directly to a data set or called from user-defined code. `Weka` contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. `Weka` is open source software issued under the GNU General Public License. There exists an interface from the statistical software system `R` to `Weka`.

12.5 Further Reading and Advanced Topics

There are a lot of well-written textbooks on data mining. Besides the CRISP-DM approach developed by Chapman *et al.* (2000), we would like to mention the following books for a detailed description of data mining. Hand *et al.* (2001) treat data mining from a broad view covering all fundamental topics. The textbook by Berry and Linoff (1997) is aimed especially for business users, and Mitchell (1997) and Witten and Frank (2005) emphasize the machine learning viewpoint of data mining. An overview and a description of modeling techniques from a more statistical point of view is given by Hastie *et al.* (2001).

McCullagh and Nelder (1989), Cox and Snell (1989), and Agresti (1996) treat generalized linear models, maximum likelihood estimation and hypothesis testing in detail. Robust parameter estimation is treated, for example, by Hampel *et al.* (1986), Rousseeuw (1984), Künsch *et al.* (1989), Bickel *et al.* (1993), Christmann (1994, 1998), and Bianco and Yohai (1996) for the case of a generalized linear model.

Cochran (1977) and Levy and Lemeshow (1999) investigate sampling techniques. Methods for the determination of sample sizes for simple random sampling, stratified random sampling, and other sampling plans are treated in detail by Desu and Raghavarao (1990) and Lemeshow *et al.* (1990).

Methods dealing with missing values are treated by Rubin (1987) and Little and Rubin (1987). The statistical software package `SAS`® contains the procedures `PROC MI` and `PROC MIANALYZE`, which can be used to detect patterns of missing values and for the imputation for such values. These procedures are based on the EM algorithm or on the Markov Chain Monte Carlo (MCMC) method. The statistical software package `SPSS`® offers the module "missing value analysis" for the same task.

12.6 Summary

This chapter briefly described data mining, the main phases of data mining projects, and the role of kernel methods treated in this book within the whole data mining process. Data mining is the analysis of usually large observational data sets to detect and model unsuspected relationships and summarize the data in ways that are both understandable and useful to the data owner; see Hand *et al.* (2001, p. 1). The CRISP-DM project (Chapman *et al.*, 2000) has developed an industry- and tool-neutral data mining process strategy. This data mining strategy consists of six main phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Support vector machines treated in this monograph are mainly used in the modeling phase, but there is ongoing research to use SVMs in other phases also.

12.7 Exercises

12.1. CRISP-DM versus SEMMA ($\star$)
Compare the data mining strategies CRISP-DM proposed by Chapman *et al.* (2000) and SEMMA used by the `SAS`® `Enterprise Miner`™.

12.2. Importance of data preparation and data quality ($\star$)
Explain why data preparation and methods to improve the data quality can be quite time-consuming. Explain why these steps are important even for non-parametric methods that make only minor model assumptions.