**Preface:**

**Convergence of communication and** computing has produced a society that feeds on information, yet most of that information is in raw form. If data is characterized as recorded facts, then information is the set of patterns or expectations that underlie the data. There is a huge amount of information locked up in databases, **our mission** is to bring it forth.

**What is data mining?** Data mining is the extraction of implicit, previously unknown, and potentially useful information from data. Build algorithms that sift through databases automatically, seeking patterns or regularity in the data. Strong patterns, if found will likely generalize to make accurate predictions on future data. Many patterns will be banal or uninteresting, others will be spurious or accidental.

**Real data** is imperfect, some parts will be mixed up, others missing. Anything discovered will be inexact, with exceptions to every rule and cases not covered by a rule. Algorithms need to be robust enough to cope with imperfect data and extract regularities that are inexact but useful.

-------------------------

**Machine learning** provides the technical basis of data mining. It is used to extract information from the raw data in databases, ideally expressed in a comprehensible form and useable for a variety of purposes.

**The process is one of abstraction**, inferring structure underlying the data. This course is about tools and techniques of machine learning used for finding and/or describing structural patterns in data. There is no magic, hidden power, or alchemy in machine learning, instead, simple and practical techniques can often extract useful information from data.

**Machine learning enables the acquisition of structural descriptions from examples.** These descriptions can be used for prediction, explanation, and understanding.

**Some applications focus on prediction**, forecasting what will happen in new situations from data that describe what happened in the past, often by guessing the classification of new examples.

We are equally if not more interested in applications where the result of learning is a **description of a structure** that can be used to classify examples, supporting explanation and understanding as well as prediction. Insights gained by the user are of most interest in the majority of practical data mining applications.

The objective of this course is to introduce the tools and techniques for machine learning that are used in data mining. You will learn what these techniques are, and appreciate their strengths and applicability.

**To apply machine learning productively, you need to understand how the algorithms work**, you cannot apply blindly and expect to get good results. Different problems require different techniques, and it is rarely obvious which technique is suitable for a given situation, you need to know about the range of possible solutions. The datasets used in this course are chosen not to illustrate actual large scale practical problems, but to help understand what the different techniques do, how they work, and their range of application.

-------------------------------------------------------------------------------------

**Overview Part 1:**

**Machine learning methods will be covered at successive levels of detail.**

**Chapter 1** (What’s it all about) - We will first learn what machine learning is, where it can be used, and see some examples of practical applications.

**Chapter 2** (Input: concepts, instances, attributes) and

**Chapter 3** (output: knowledge representation) cover different kinds of input and output or knowledge representation involved.

**Chapter 4** (Algorithms: the basic methods) describes the basic methods of machine learning, simplified to make them easy to comprehend.

**Chapter 5** (Credibility: evaluating what’s been learned) equips us to evaluate the results obtained from machine learning - to make progress in machine learning its essential to be able to measure how well you are doing.

----------------------------------------------------------------------------------------

**Overview Part 2 – the advanced techniques of machine learning for data mining**.

We will see at the lowest and most detailed level:

**Chapter 6 (Trees and rules), Chapter 7 (extending instance-based and linear models**), the low-level issues of implementing a spectrum of machine-learning algorithms, including the complexities necessary to make them work well in practice. We cannot ignore this detailed information, because it is at this low level that full working implementations of machine learning schemes are written.

**Chapter 8 (data transformations)** describes practical topics involved with engineering the input/output to machine learning (selecting and discretizing attributes).

Chapter 9 (Probabilistic methods) and chapter 10 (deep learning) provide rigorous accounts of probabilistic methods for machine learning, and deep learning respectively. Chapter 11 (beyond supervised and unsupervised) looks at semi-supervised and multi-instance learning.

**Chapter 12 (Ensemble learning)** covers techniques of ensemble learning, combining the output from different machine learning techniques.

Part 1: introduction to data mining

**Chapter 1: What’s it all about?**

**This course is about machine learning techniques for data mining.** we’ll start by explaining the meaning of machine learning and data mining, and give some simple examples of machine learning problems, such as classification and numeric prediction tasks, to illustrate the kinds of input and output involved. A brief review of several fielded (practical) applications will follow, from a diverse set of areas.

**Machine learning is closely linked to statistics**, and we will discuss the connections between those fields. We also discuss the connection of machine learning to AI and search. We conclude with a discussion of ethics.

**Human IVF** involves collecting several eggs from the woman’s ovaries, which produce embryos after fertilization. Some of the embryos are selected and transferred to the uterus. The problem is to select the best embryo, or the ones most likely to survive. Selection is based on ~60 recorded features of the embryo, characterizing their morphology, oocyte, follicle, and sperm sample. This large number of features makes it difficult for an embryologist to perform a comprehensive assessment, and compare by hand with historical data. A research project in England has investigated machine learning as a technique for making this selection, using historical records of embryos and their outcome as training data.

**Every year in new Zealand dairy farmers make a tough business** decision – which cows to retain, and which cows to butcher, typically 1/5 cows per year are culled. Each cow’s breeding, milk production history, age, history of health problems, calving difficulty, temperament, and fertility influences this decision. ~700 attributes of several million cows have been recorded over the years, and machine learning has been used as a way of ascertaining what factors are taken into account by successful farmers, to propagate their skills and experience to others.

1.1 Data mining and machine learning

**Patterns in data**

**This course is about looking for patterns in data.** People have been seeking patterns in data ever since human life began. Hunters seek patterns in animal migratory behavior, farmers seek patterns in crop growth, politicians seek patterns in voter opinion, and lovers seek patterns in their partners’ responses.

**A data scientist’s job** is to make sense of the data, discover patterns that govern the process being measured and encapsulate them into rules or theories that can be used for predicting what will happen in new situations.

The unbridled growth of databases in recent years, data on everyday activities such as customer choice, brings data mining to the forefront of new business technologies. It has been estimated that the amount of data stored in the world’s databases doubles every 20 months, as amount of data increases opportunities for data mining also increases.

**Data mining is about solving problems by analyzing data already present in databases.** Take the example of customer loyalty in a competitive marketplace. A database of customer choices along with customer profiles holds the key to this problem. Patterns of behavior recorded from former customers can be analyzed to identify distinguishing characteristics of those likely to switch products and those likely to remain loyal. Once these characteristics or features have been found, they can be used to identify customers likely to switch brands, and this group can be targeted for special treatment that is too expensive to apply to the entire customer base. In today’s highly competitive, customer centered, service-oriented economy, data is the raw material that fuels business growth.

**Data mining is defined as the process of discovering patterns in data**. The process must be automatic or semi-automatic, the patterns must be meaningful in that they lead to some advantage or understanding. Useful patterns allow us to make nontrivial predictions on new data. Patterns can be expressed either as a black box, whose innards are effectively incomprehensible, or a transparent box, revealing the structure of the pattern. Both make good predictions, the difference is whether the patterns that are mined are represented in terms of a structure that can be examined, reasoned about, and used to inform future decisions. We call these patterns structural because they capture the decision structure in an explicit way. Most of the course is about techniques for finding and describing structural patterns in the data, but there are applications where black box methods are more appropriate because they yield significantly greater predictive accuracy and we will cover those as well.

------------------------------------------------------------------------------------------------

**Describing structural patterns:**

**what is meant by a structural pattern? How do you describe them? What form does the input take?** We will answer these questions by way of example (Contact lens data, Table 1.1). this table gives the conditions under which an optician might want to prescribe soft contact lenses, hard contact lenses, or no contact lenses. Part of a structural description of this information might be as follows:

If tear production rate = reduced, then recommendation = none

Otherwise, if age=young and astigmatic=no, then recommendation=soft

Structural descriptions need not be expressed as a set of rules such as these, decision trees, which specify the sequence of decisions that need to be made along with the resulting recommendation, are another popular means of expression. This example is a very simplistic one, first, all combinations of possible values are represented in the table, there are 24 rows, representing 3 possible values of age and 2 values each for spectacle prescription, astigmatism, and tear production rate (3\*2\*2\*2=24).

**The rules do not generalize from the data, they merely summarize it**. In most learning situations, the set of examples is far from complete, and part of the job is to generalize to new examples. Imagine omitting some of the rows in the table for which tear production is reduced and still come up with the rule:

If tear production rate = reduced then recommendation = none

Which would generalize to the missing rows and fill them in correctly. Second, values are specified for all the features in the examples, real-life datasets often contain examples in which the values of some features, for some reason or another are unknown (measurements not take or lost). Third, the preceding rules characterize all the examples correctly, whereas often, because of errors or noise or incomplete information, misclassification occurs, even on examples used to create the classifier.

------------------------------------------------------------------------------------------

**Machine Learning. Now that we have some idea of inputs and outputs, we turn to machine learning. What is learning anyways?** Dictionary defines “to learn” as 1) to get knowledge by study, experience, or being taught 2) to become aware by information or from observation 3) to commit to memory 4) to be informed of, ascertain 5) to receive instruction. These meanings have shortcomings when applied to computers. Cannot quantify 1&2, others are too passive and trivial for computers.

**What we are really interested in is improvements in performance.** Earlier we defined data mining as the process of discovering useful patterns, automatically or semiautomatically, in large quantities of data. We can formulate a similar operational definition for learning: things learn when they change their behavior in a way that makes them perform better in the future (skinner behaviorism). This ties learning to performance rather than knowledge. can test learning by observing the behavior and comparing it with past behavior, which is a more objective kind of definition.

**Training vs learning:** a vine grows around a trellis in a vineyard, has it learned or been trained? Learning without purpose is merely training. In learning, the purpose is the learner’s while in training it is the teacher’s. to decide whether something has actually learned you need to see whether it intended to, whether or not there was purpose involved.

**fortunately, these conceptual problems do not arise in learning techniques explained in this course,** we call them ‘machine learning’ without presupposing any particular philosophical stance about what learning actually is. We are interested in techniques for finding patterns in data that provide insight or enable fast and accurate decision making.

**Input Data will take the form of a set of examples,** (customers who have switched loyalty, or situations in which certain kinds of contact lens can be prescribed).

**Output takes the form of predictions on new examples** (whether or not a customer will switch, or what type of contact lens will be prescribed).

Many learning techniques look for structural descriptions of what has been learned, these can become fairly complex and are typically expressed as sets of rules such as those described previously. These descriptions serve to explain what has been learned, in other words, the basis for new predictions.

**In many applications of machine learning to data mining, the structural description is at least as important as the ability to perform well on new examples**. People frequently use data mining to gain knowledge, not just make predictions.

------------------------------------------------------------------------------------------

1.2 Simple examples: weather problem and others:

**There are several standard datasets that we will come back to repeatedly**, different datasets tend to expose new issues and challenges, and it is instructive to have in mind a variety of problems when considering learning methods. The illustrations in this section are unrealistically simple, serious machine learning applications involve thousands or millions of examples, but when explaining how algorithms work we need simple examples that enable us to capture the essence of the problem, yet small enough to be comprehensible in every detail. Another problem with real datasets is they are often proprietary, companies will not share their customer and product choice datasets, corporate data is a valuable asset, whose value has increased tremendously with the development of machine learning techniques described in this course. The examples considered here are simple but not simplistic, they exhibit the features of real datasets.

**The weather problem:** a tiny dataset that we will use repeatedly to illustrate machine learning methods. Concerns conditions suitable for playing some unspecified game. In general, examples in a dataset are characterized by values of features or ‘attributes’, that measure different aspects of the example. For the weather dataset, there are four attributes: outlook, temperature, humidity, and windy. The outcome is whether or not to play. Table 1.2 shows the weather dataset in its simplest form, where all attributes have values that are symbolic categories rather than numbers. Outlook can be sunny, overcast, or rainy, temperature can be hot, mild, or cool, humidity can be high or normal, and windy can be true or false. This creates 36 possible combinations, of which 14 are present in the set of input examples.

**We can create a set of rules based on the weather data.** The rules are meant to be interpreted in order, check the first, if it doesn’t apply move to the second, etc. a set of rules intended to be interpreted in sequence is called a decision list, interpreted correctly in sequence this decision list classifies all the cases correctly, however taken out of context some of the rules are incorrect. For example, if humidity=normal then play=yes gets an example wrong, the meaning of a set of rules depends on how it is interpreted.

------------------------------------------------------------------------------------------------------

**In a slightly more complex form (table 1.3) two** of the attributes (Temperature and humidity) have numeric values, this means any learning scheme must create inequalities involving these attributes, rather than a simple equality test as in the previous case. This is called a numeric attribute problem, and in this case a mixed-attribute problem because not all attributes are numeric. A slightly more complex process is required to come up with rules that involve numeric tests.

**Rules we have seen so far are classification rules, they predict the classification of the example in terms of whether to play or not. We can also disregard the classification and just look for any rules that strongly associate different attribute values, we call these association rules.** Many association rules can be derived from the weather data, some examples: if temperature=cool then humidity=normal, if humidity=normal and windy=false then play=true, if outlook=sunny and play=no then humidity=high, if windy=false and play=no then outlook=sunny and humidity=high. These rules are 100% correct, and apply to 4, 4, 3, and 2 examples respectively. Nearly 60 association rules can be found that apply to two or more examples of the weather data, and are correct.

Contact lenses: an idealized problem.

Data introduced earlier tells you the kind of lens to prescribe, given certain information about the patient. Figure 1.2 shows a structural description for the contact lens data in the form of a decision tree, which in many cases is more concise and can be visualized more easily. Tree calls for a test on tear production rate, first two branches correspond to two possible outcomes. If tear production rate is reduced, outcome is none, if it is normal, a second test is made. Eventually, whatever the outcome a leaf of the tree is reached that dictates the contact lens recommendation for that case. The question of what is the most natural and easily understood format for the output from a machine learning scheme is one that we will return to in chapter 3: knowledge representation.

Irises: a classic numeric dataset

The iris dataset is arguably the most famous dataset used in machine learning, contains 50 examples each of 3 types of plant iris setosa, iris versicolor, iris virginica. Has four attributes sepal length, sepal width, petal length, petal width (cm). unlike previous datasets, all attributes have numeric values. The following rules may be learned from this dataset (rules). These rules are very cumbersome, and we will see in chapter 3: output: knowledge representation how more compact rules can be expressed that convey the same information.

CPU Performance: introducing numeric prediction

Although the iris dataset involves numeric attributes, the outcome (type of iris) is a category, not a numeric value. Table 1.5 shows some data for which the outcome and the attributes are numeric. Concerns the relative performance of computer processing power on the basis of a number of relevant attributes, each row represents 1 of 209 different computer configurations. The classic way of dealing with continuous prediction is write the outcome as a linear sum of attribute values with appropriate weights (equation) this is called a linear regression equation and the process of determining the weights is called linear regression, a well known procedure from statistics we will review in chapter 4: Algorithms, the basic methods.

Labor negotiations (not covered)

Soybean classification (not covered)

Fielded Applications: most of the illustrations above are toy problems, deliberately chosen to be small so we can use them to work through algorithms later in the course. Here are some applications of machine-learning that have actually been put to use. Being fielded applications, the emphasis is on the ability to perform well on new examples, and the fact that the decision structure is comprehensible is a key feature in the successful adoption of the application.

Web mining: search engines mine the content of the web, they also mine the content of your queries, to select advertisements you might be interested in. they have a strong incentive to do this accurately because they only get paid by advertisers when users click on their links. Search engine companies mine your clicks, because knowledge of which results you click on can be used to improve the search next time. Online booksellers mine the purchasing database to come up with recommendations such as “users who bought this book also bought these ones”, movie sites recommend movies based on your previous choices and other people’s choices, they want you to keep coming back to the website.

Decisions involving judgement:

When you apply for a loan, you have to fill out a questionnaire asking for relevant financial and personal information. This information is used by the loan company as the basis for its decision whether to lend you money. Such decisions are often made in two steps, first, statistical methods are used to determine clear accept and reject cases, remaining borderline cases are more difficult and call for human judgement. Credit industry professionals point out that borderline cases tend to be active customers of credit institutions because their finances remain in chronically volatile condition. Enter machine learning, the input was 1000 training examples of borderline cases for which a loan had been made, specifying whether the borrower had paid off or defaulted. For each training example 20 attributes were extracted (age, years with current employer, years at current address, etc), and classification rules were derived that made correct predictions on 2/3 of borderline cases in an independently chosen test set. The rules improved the success rate of the loan decisions and were also attractive because they could explain to the applicant the reason behind the decision.

Screening images: environmental scientists want to detect oil slicks from satellite images to give early warning of ecological disasters and deter illegal dumping. Radar satellites provide an opportunity for monitoring coastal waters day and night, regardless of weather conditions, oil slicks appear as dark regions in the image whose size and shape evolve depending on weather and sea conditions. However, other look-alike dark regions can be caused by local weather conditions such as high wind. Detecting oil slicks is an expensive manual process requiring highly trained personnel who assess each region in the image. Machine learning allows a hazard detection system to be trained on examples of spills and nonspills, the input is a set of raw pixel images from a satellite and the output is a smaller set of images with potential oil slicks marked by a colored border.

Diagnosis: preventative maintenance of electromechanical devices such as motors and generators can forestall failures that disrupt industrial processes. Technicians inspect each device, measuring vibrations at various points to determine whether the device needs servicing, typical faults include shaft misalignment, mechanical loosenings, faulty bearings, and unbalanced pumps. A particular chemical plant uses more than 1000 different devices, ranging from small pumps to large turbo-alternators, which used to be diagnosed by a human expert with 20 years of experience. 600 faults, each comprising a set of measurements along with the expert’s diagnosis were available, representing 20 years of experience. The goal was to diagnose the kind of fault, given one existed. The derived attributes were run through a rule induction algorithm, to produce a set of diagnostic rules. Although the rules were quite complex, the expert liked them because he could justify them in light of his mechanical knowledge, and subsequent performance tests indicated that the learned rules were superior to the handcrafted ones previously elicited by the expert, confirmed by subsequent use in the chemical factory. It is interesting to note that the system was put into place not because of its good performance, but because the domain expert approved of the rules that had been learned.

Market basket analysis: the use of association techniques to find groups of items that tend to occur together in transactions, such as supermarket checkout. For example, customers that buy beer may also buy chips, a discovery that may be significant from supermarket’s point of view. Or, on thursdays customers often purchase beer and diapers together, an initially surprising result that makes sense in light of the fact as young parents stock up for a weekend at home. Such information can be used for many purposes, planning store layouts, offering coupons for a matching product when one is sold alone, etc.

1.4 the data mining process: figure 1.4. 1) understand what you want to achieve by implementing a data mining scheme, this is the business understanding phase, investigating the business objectives and requirements, deciding whether data mining can be applied to meet them, and what kind of data needs to be collected to build a deployable model. 2) data understanding phase an initial dataset is established and studied to see whether it is suitable for further processing, if the data quality is poor it may be necessary to collect new data based on more stringent criteria. The next 3 steps (data preparation, modeling, and evaluation) are what the book deals with. Preparation involves pre-processing the raw data so the machine learning algorithms can produce a model, ideally a structural description of information implicit in the data. Preprocessing may involve model building as well, because many preprocessing tools build an internal model of the data to transform it. Data preparation and modeling usually go hand in hand, and it is almost always necessary to iterate, results obtained using modeling provide new insights that affect the choice of preprocessing techniques. Evaluation involves checking if the structural descriptions inferred from the data have any predictive value, or simply reflect spurious regularities. If the model’s accuracy is sufficiently high, the next step is to deploy it in practice, integrating it into a larger software system.

1.5 Machine learning and statistics – what is the difference between machine learning and statistics? Cynics equate data mining to statistics + marketing hype. There is no dividing line, instead a continuum of data mining techniques, some that derive from skills taught in standard statistics courses, others more closely associated with the kind of machine learning that has arisen out of computer science. If forced to point to a single difference to emphasize, it is that statistics is more concerned with testing hypotheses, whereas machine learning is more concerned with formulating the process of generalization as a search through possible hypotheses. However this is a gross oversimplification, statistics is far more than just hypothesis testing, and many machine learning techniques do not involve any searching at all.

1.6 Generalization as search – one way of visualizing the problem of learning – and one that distinguishes it from statistical approaches – is to imagine a search through a space of possible concept descriptions for one that fits the data. Suppose that concept descriptions – the result of learning – are expressed as rules such as those for the weather problem. Suppose we list all the possible sets of rules and then look for ones that satisfy a given set of examples. A big job, but not an infinite job. So the process of generalization with rule sets can be regarded as a search through an enormous, but finite, search space. In principle the problem can be solved by enumerating descriptions and striking out those that don’t fit the examples presented.

Enumerating the concept space – regarding it as search is a good way of looking at the learning process. However the search space though finite is extremely large, and it is impractical to enumerate all possible descriptions and see which ones fit. In the weather problem, there are 4\*4\*3\*3\*2 = 288 possibilities for each rule.