Chapter 1

Introduction and course overview

What is machine learning and what can we use it for. Fundamental concepts. Different types of learning. Outline of the course.

Note: for details on assignments, class schedules and assessment, please refer to the course page

1.1 What is machine learning

Machine learning is the science of building systems that improve with data. This is a broad concept that includes instances ranging from self-driving cars to sorting images on a database and from recommendation systems for diagnosing diseases to fitting parameters in climate change models. The fundamental idea is that the system can use data to improve its performance at some task. Which immediately points us to the three basic elements of a well-posed machine learning problem:

1. The task that the system must perform.
2. The measure by which its performance can be evaluated.
3. The data that can be used to improve its performance.

For example, suppose we want to automate airline ticket purchasing by phone. The task to perform is thus to identify the requests, such as asking to book a flight, the origin and destination, the required flight and so on. The system's performance at this task can be measured by the frequency of correctly identified expressions. In the work reported in [9], the data the used was a corpus of manually annotated expressions. This is an example of the system parsing spoken sentences and identifying the relevant elements for processing the requests:

```
<book_flight> please book me on </book_flight> <numflt> flight twenty one </numflt>

<i_want_to_go> i would like to fly </i_want_to_go>
<city_from> from philadelphia </city_from> <city_to> to dallas </city_to>

<request1> could you please list the </request1> flights
<city_from> from boston </city_from> <city_to> to denver </city_to>
on <date> july twenty eighth </date>
```

Different tasks will determine different approaches. We may want to predict some continuous value, such as the price of apartments, which is a *Regression* problem. Or we may have a *Classification*, when we want to predict in which category, from a discrete set, each example belongs to. If we do this from a set of data containing the right answers, so we can then extrapolate to new examples, we are doing *Supervised Learning*.

We may want to find *Association Rules*, which are joint or conditional probability distributions of some features. For example, which products customers tend to purchase together, so we can optimize their placement in supermarket aisles. We may want to do *Density Estimation* to understand how feature values are distributed, or perhaps group examples according to some similarity measure, which is called *Clustering*. For example, grouping images together according to how similar they look. These are examples of *Unsupervised Learning*, because in these cases there is no *Ground Truth* in the data that can tell us if we made the right or wrong choice.

Supervised Learning requires that all data be labelled and *Unsupervised Learning* uses only unlabelled data. But it is possible to use data sets in which some data is labelled but the rest, usually most of the data, is not. In this case, we have *Semi-supervised Learning*. This approach has the advantage that, usually, unlabelled data is much easier to find than correctly labelled data. For example, it is possible to obtain from the World Wide Web many examples of English texts but to label correctly each grammatical element of each sentence would be very laborious. By combining clustering and classification it is possible to use unlabelled texts to improve the parsing and classification of elements from a set of labelled texts.

Some tasks involve sequences of decisions, such as when playing a game, and the outcome can only be assessed after all decisions are made (e.g. win or lose). This is an example of a *Reinforcement Learning* problem, where each move is not good or bad by itself but only in the context of the sequence of moves.

With such a diverse range of applications and problems, Machine Learning benefits from contributions of many other disciplines. Computer Science, evidently, from subjects such as artificial intelligence, algorithms, complexity analysis and data management. Statistics is also important for inference, experiment design and data analysis. Mathematics is also crucial, with numerical methods at the base of most machine learning algorithms and probability theory underlying many machine learning approaches. Finally, Machine Learning is strongly inspired by Neuroscience, in particular perception, learning and memory, and Philosophy, especially epistemology and ontology.

1.2 Why machine learning is useful

Machine learning is useful if we cannot, or do not wish to, explicitly program a solution to our problems. For example, humans can easily identify handwritten digits such as those in zip codes of mail addresses. However, it is not easy to find specific rules for programming a computer to automate this task. This is a classic example of a problem where machine learning is a useful solution. For decades now, machine learning algorithms have been applied to automating identification of handwritten digits [7]. Figure 1.1 shows an example of this problem.

Machine learning is also becoming increasingly more useful as the amount, complexity and quality of data increases. Recently, the trend has been towards an exponential growth in data.

Another reason for using machine learning is so that the system can adapt to changing conditions. If we have a static set of data we may figure out some rules for organizing and grouping the examples after careful examination of the data. However, if the data set is continuously changing, as is generally

Figure 1.1: Handwritten zip code digits

the case for most applications, it is not feasible to have programmers dedicated to constantly adapting the code to extract information from the data. In these cases, automated systems that can constantly learn from the new data are crucial. An example of this is the optimization of search engines. The search engine must interpret the user's query, consider how to expand the search by using synonyms or words with overlapping meaning, and, especially, how to rank the results. These systems constantly learn from the users, remembering which links are preferred, which search terms are most used and their associations, and a large amount of information on the user (often even arguably violating privacy rights).

Machine learning raises some important technical challenges, and even some ethical issues regarding the information that is used and the purpose for which it is used, but it is clear that machine learning is an important field in computer science and its importance can only grow as data and computation power keep growing.

1.3 Fundamental concepts

Throughout this course we will constantly rely on some important concepts and it is important that they are clear from the beginning. First is the concept of the *hypothesis class*. This is the space of possibilities in which we will try to optimize the solution to our machine learning problem. Suppose we have the data set represented in Figure 1.2, where each point has two continuous features, represented in the X and Y axes, and is labelled either in the red or blue class.

One possible way of separating them would be to try to find the horizontal line that best splits the two classes. In this case, our *hypothesis class* would be the set of horizontal lines, as represented in Figure 1.3.

Machine learning is closely associated with statistics and so the term *model* is also used to refer to a representation of a *hypothesis class*, typically using some parameters. For example, we could describe this set of all horizontal lines with the parametric model $y = \theta$, where θ is the parameter to adjust in order to instantiate the model into a specific line. This is an hypothesis in the *hypothesis class*. In the literature, and in this course, it is common to find both *model* and *hypothesis class* to refer to the set of possible instances in which we want to find the best solution to our learning problem. An alternative to the horizontal line model would be to consider all circles of radius 1 and try to find which of these circles includes all blue points and excludes red points (Figure 1.4).

This different *hypothesis class* allow us to find different *hypotheses* that cannot be expressed with the $y = \theta$ model. In this case, we would have a model with two parameters, $(x - \theta_1)^2 + (y - \theta_2)^2 = 1$.

We can say that the circle of radius 1 centered at $(-1, -1)$ is an instance of the $(x-\theta_1)^2+(y-\theta_2)^2 = 1$ model, or a hypothesis from this hypothesis class, and the line $y = 0$ is an instance of the $y = \theta$

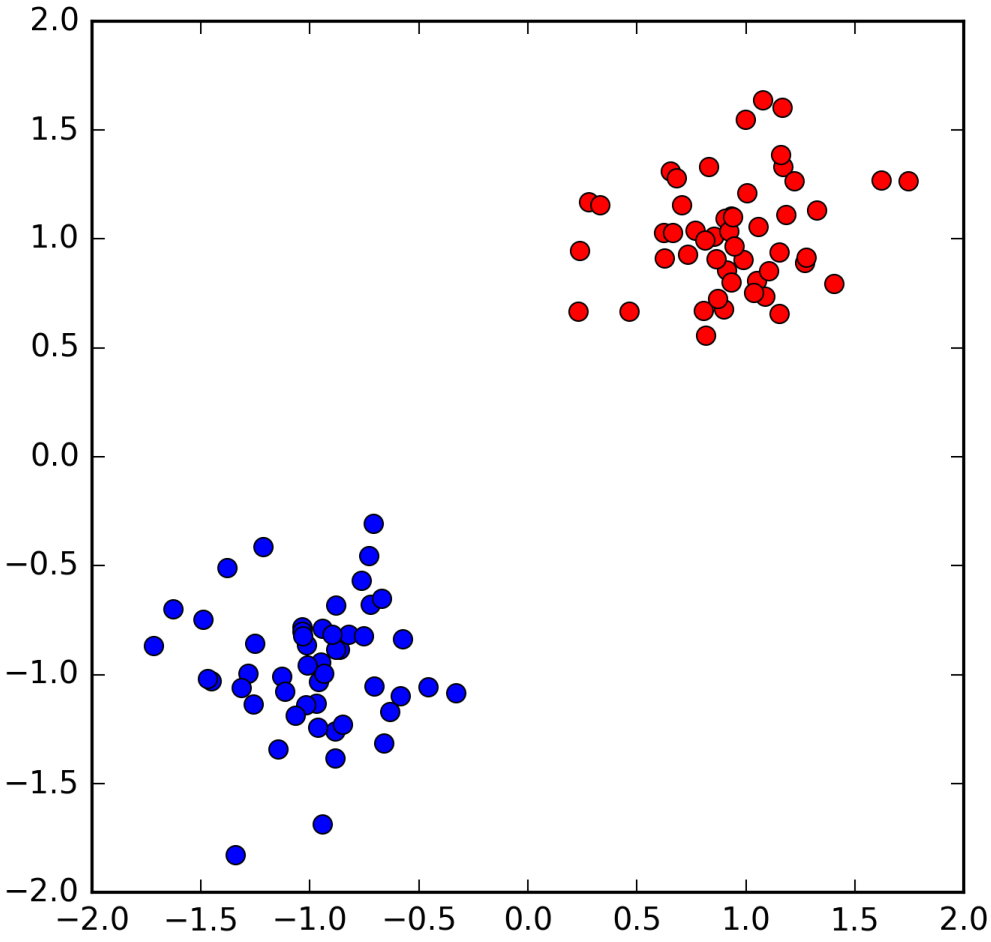

Figure 1.2: Arbitrary data set in two dimensions, divided into two classes (red and blue)

model or an hypothesis from that hypothesis class. The important point here is to distinguish between the universe of possibilities we are considering, which is the *model* or *hypothesis class*, and the specific instantiation of the model, or *hypothesis*, we obtain by setting the parameters and which constitutes an answer to the learning problem.

The *hypothesis class* determines the *inductive bias* of our learning system. We cannot learn anything if we do not assume anything, because this would make it impossible to extrapolate from the data we have, and so we must make some prior assumptions about the problem we are solving. For example, assuming that we can separate the red and blue classes using some horizontal line or some circle of radius 1. This is such a crucial point that we will often talk about the problem of *model selection*, which consists of finding the best *hypothesis class* for a particular problem.

1.4 Overview of Machine Learning problems

As mentioned above, one class of machine learning problems is *Unsupervised learning*, involving unlabelled examples. In this case, the objective is generally to obtain some information about the structure of the data. A schematic representation of the unsupervised learning process is shown in Figure 1.5

For example, in *clustering*, we may want to organize the data so as to group together similar data points. An example of this is clustering of images obtained from the World Wide Web, to help guide

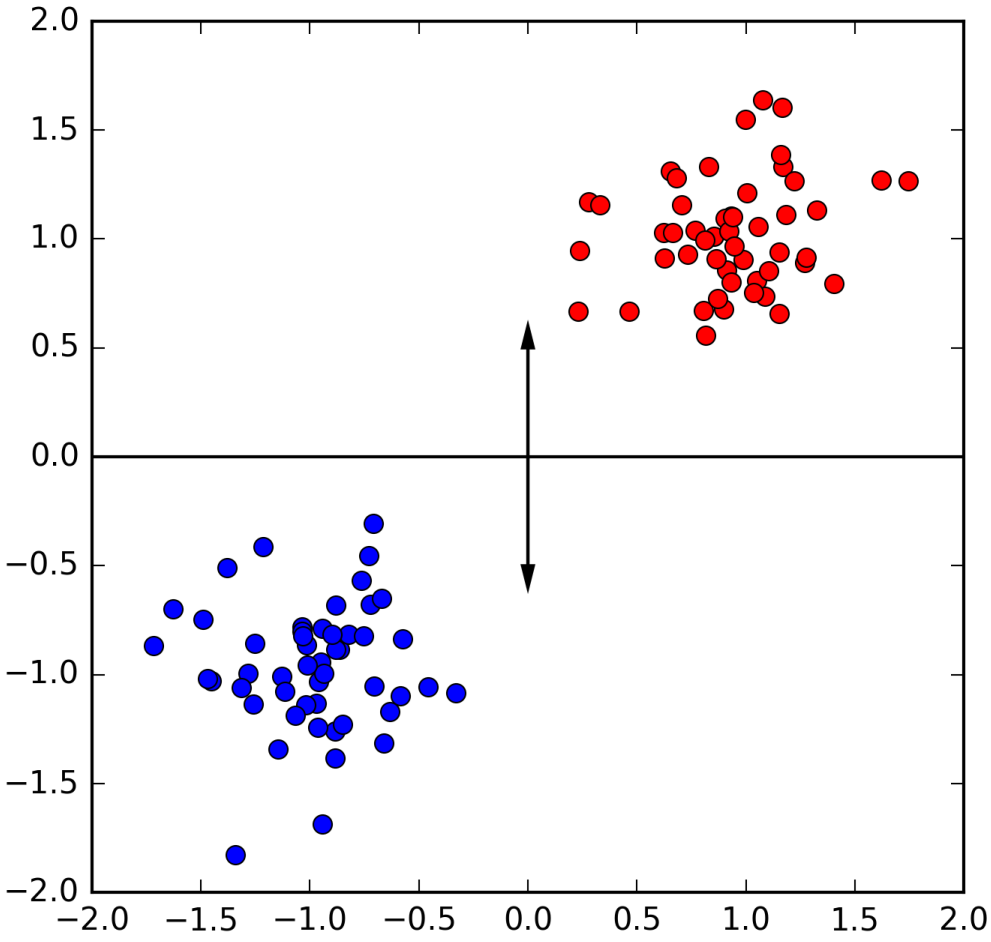

Figure 1.3: One hytpothesis class: horizontal lines to divide the two classes

organize search results. The authors of [5] extracted features both from the images themselves as well as from the text of the pages where the images were found and hypelinks between them. Figure 1.6 gives an example of the resulting clusters for an image search with the keyword “pluto”.

Unsupervised learning gives us new values we can associate with the original data, and so *unsupervised learning* can be used as part of a larger learning task. This is what happens in deep learning, for example.

In *Supervised Learning*, we have a fully labelled data set that provides us not only with the input features for our learning machine but also with the correct answers, allowing us to supervise the learning process directly. Schematically, supervised learning looks like the diagram in Figure 1.7:

From the data we feed into the learner those features that will be used in the future to predict something about new data. But we also use the target values to compute the *empirical error* of the learner during the learning process. In this way, we can improve its performance in correctly predicting the target values.

An example of this is given in [20]. The task consists of identifying faces in photographs. It is a *Classification* task because each segment of the image may be classified as either a face or not a face. The data used for training the classifier consists of a set of labelled images of faces and a set of labelled images that were not faces. Figure 1.8 shows, on the right, an example of the set of face images used in training (non-face images are not shown here) and an example of the application of the final classifier.

The authors of [20] also report a *semi-supervised learning* approach, where the labelled examples

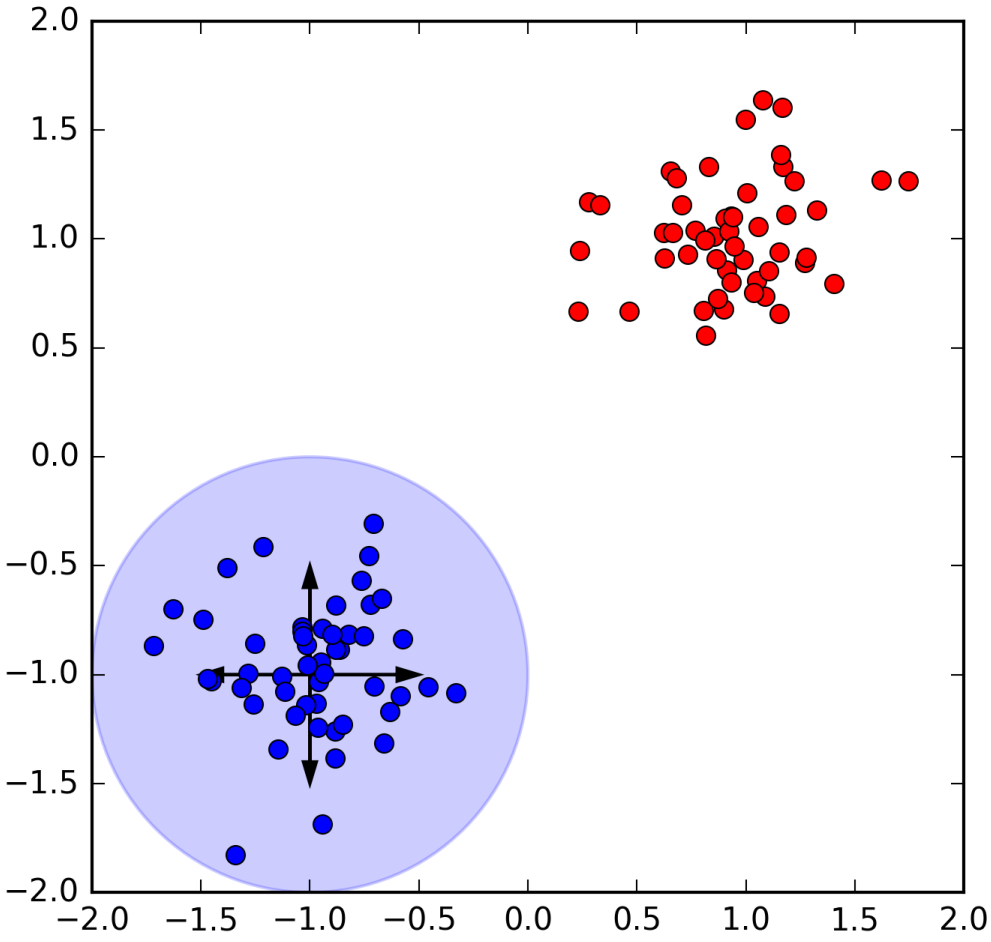

Figure 1.4: Another hytpothesis class: circles of radius 1 to isolate one class from the other

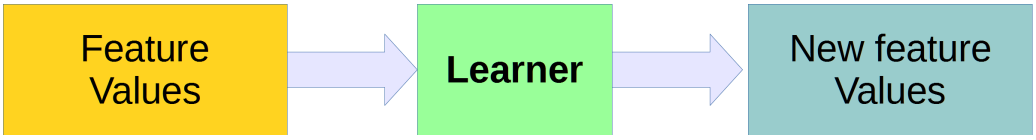

Figure 1.5: Diagram of an unsupervised learning process

used to train the classifier are enriched with unlabelled image data. This requires accounting both for the statistical structure of all data, including the unlabelled data, and the classifier's performance in the labelled data. Another type of machine learning problem is *Reinforcement Learning*. In this type of problem, the task is to optimize some output, like game moves, for example, but the feedback guiding the learner must be given by some heuristic or an evaluation of an eventual outcome and not given by the data. So the learner must improve performance by improving the feedback (e.g. win or lose the game). Figure 1.9 shows a diagram of this learning process.

Some examples of reinforcement learning applications include robotics, for locomotion control and other tasks such as object manipulation, autonomous vehicle control, operations research (pricing, routing, marketing), and games. In this course we will focus on supervised and unsupervised learning and will not cover semi-supervised or reinforcement learning problems.

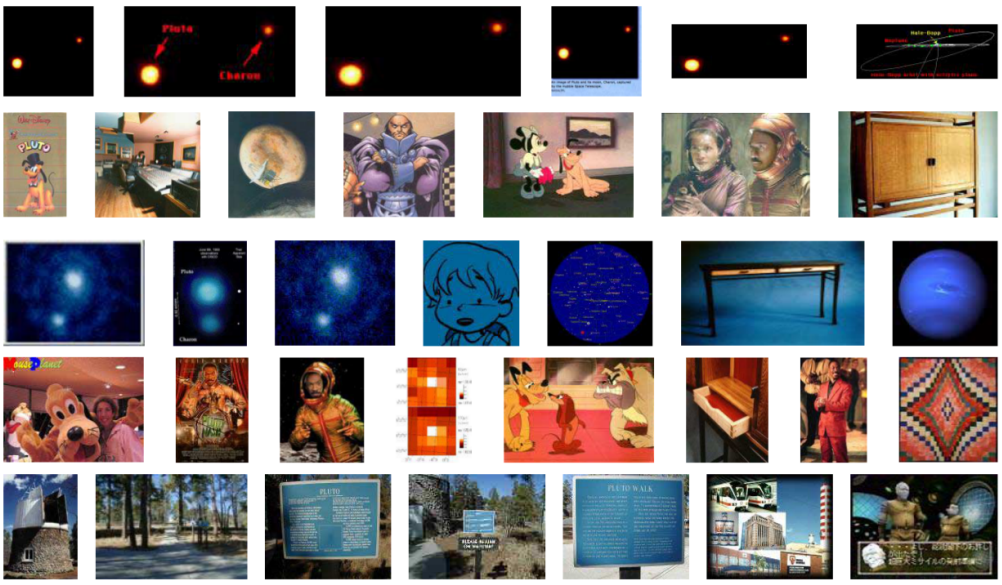

Figure 1.6: Clusters obtained for the image results of a search for “pluto”. Each cluster is a row of images. Figure from [5].

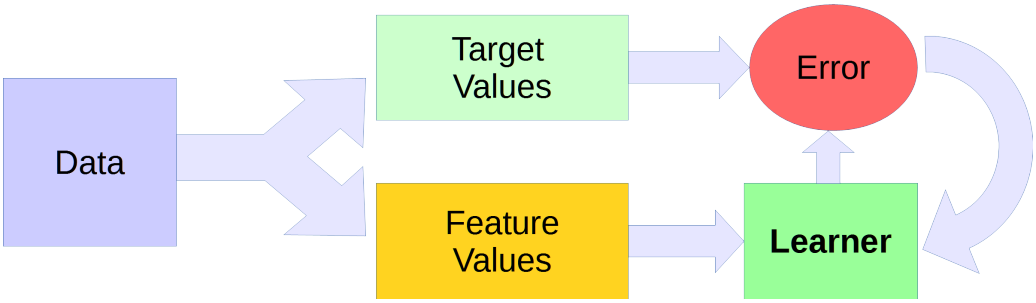

Figure 1.7: Diagram of a supervised learning process

Figure 1.8: Labelled face images used in training (left) and the result of applying the classifier (right). For more details see the original paper [20].

1.5 Goals and course outline

The main goal of this course is to provide a foundation on theoretical and practical aspects of machine learning so the student can get some experience with common machine learning techniques, understand the concepts, be able to follow the literature, acquire the skills to handle scientific computation problems and understand the algorithms from their mathematical specifications.

The first part of this course will focus on supervised learning. Broadly speaking, the task of learning

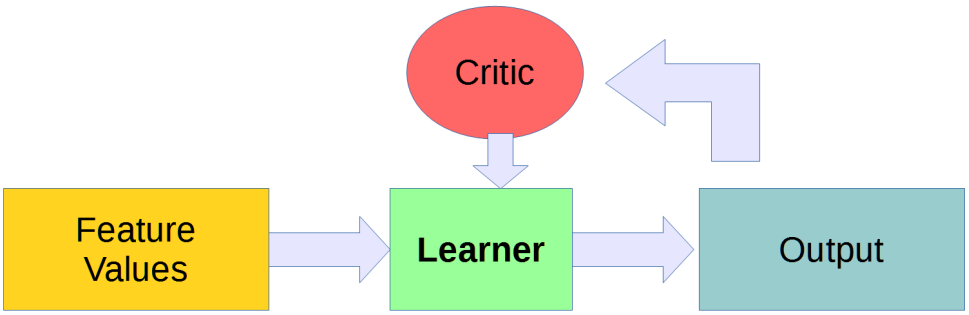

Figure 1.9: Diagram of a reinforcement learning process

how to predict an attribute of a universe of examples from a data set which includes both the observed features used for the prediction and the attribute to be predicted. The second part will cover some basics of learning theory and more detailed aspects of model selection. The third part will be dedicated to ensemble methods and the final part to unsupervised learning algorithms.

1.6 Further Reading

1. Alpaydin [2], Chapter 1
2. Mitchell [18], Chapter 1
3. Marsland [17], Chapter 1, sections 1.1 through 1.4.

Bibliography

[1] Uri Alon, Naama Barkai, Daniel A Notterman, Kurt Gish, Suzanne Ybarra, Daniel Mack, and Arnold J Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12):6745–6750, 1999.

[2] Ethem Alpaydin. *Introduction to Machine Learning*. The MIT Press, 2nd edition, 2010.

[3] David F Andrews. Plots of high-dimensional data. *Biometrics*, pages 125–136, 1972.

[4] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, New York, 1st ed. edition, oct 2006.

[5] Deng Cai, Xiaofei He, Zhiwei Li, Wei-Ying Ma, and Ji-Rong Wen. Hierarchical clustering of www image search results using visual. Association for Computing Machinery, Inc., October 2004.

[6] Guanghua Chi, Yu Liu, and Haishandbscan Wu. Ghost cities analysis based on positioning data in china. *arXiv preprint arXiv:1510.08505*, 2015.

[7] Le Cun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems*, pages 396–404. Morgan Kaufmann, 1990.

[8] Pedro Domingos. A unified bias-variance decomposition. In *Proceedings of 17th International Conference on Machine Learning. Stanford CA Morgan Kaufmann*, pages 231–238, 2000.

[9] Hakan Erdogan, Ruhi Sarikaya, Stanley F Chen, Yuqing Gao, and Michael Picheny. Using semantic analysis to improve speech recognition performance. *Computer Speech & Language*, 19(3):321–343, 2005.

[10] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

[11] Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007.

[12] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[13] Patrick Hoffman, Georges Grinstein, Kenneth Marx, Ivo Grosse, and Eugene Stanley. Dna visual and analytic data mining. In *Visualization'97., Proceedings*, pages 437–441. IEEE, 1997.

[14] Chang-Hwan Lee, Fernando Gutierrez, and Dejing Dou. Calculating feature weights in naive bayes with kullback-leibler measure. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 1146–1151. IEEE, 2011.

[15] Stuart Lloyd. Least squares quantization in pcm. *Information Theory, IEEE Transactions on*, 28(2):129–137, 1982.

[16] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA., 1967.

[17] Stephen Marsland. *Machine Learning: An Algorithmic Perspective*. Chapman & Hall/CRC, 1st edition, 2009.

[18] Thomas M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997.

[19] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19):2507–2517, 2007.

[20] Roberto Valenti, Nicu Sebe, Theo Gevers, and Ira Cohen. Machine learning techniques for face analysis. In Matthieu Cord and Pádraig Cunningham, editors, *Machine Learning Techniques for Multimedia*, Cognitive Technologies, pages 159–187. Springer Berlin Heidelberg, 2008.

[21] Giorgio Valentini and Thomas G Dietterich. Bias-variance analysis of support vector machines for the development of svm-based ensemble methods. *The Journal of Machine Learning Research*, 5:725–775, 2004.

[22] Jake VanderPlas. Frequentism and bayesianism: a python-driven primer. *arXiv preprint arXiv:1411.5018*, 2014.