Journal of
Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7923071

STUDY OF STRUCTURAL AND PROBABILISTIC MODELLING AND MACHINE LEARNING

Sapna Thakur and Dr. Priyanka Bhalerao
Department of Mathematics
Dr. A. P. J. Abdul Kalam University, Indore (M. P.) – 452010
Corresponding Author Email : sapana.thakur15@yahoo.in

Abstract:
Probabilistic modelling provides a frame-work for understanding what learning is, and has therefore emerged as one of the principal theoretical and practical approaches for designing machines that learn from data acquired through experience. The probabilistic framework, which describes how to represent and manipulate uncertainty about models and predictions, plays a central role in scientific data analysis, machine learning, robotics, cognitive science, and artificial intelligence. This article provides an introduction to this probabilistic framework, and reviews some state-of-the-art advances in the field, namely, probabilistic programming, Bayesian optimisation, data compression, and automatic model discovery.

1. Introduction

The key idea behind the probabilistic framework to machine learning is that learning can be thought of as inferring plausible models to explain observed data. A machine can use such models to make predictions about future data, and decisions that are rational given these predictions. Uncertainty plays a fundamental role in all of this. Observed data can be consistent with many models, and therefore which model is appropriate given the data is uncertain. Similarly, predictions, about future data and the future consequences of actions, are uncertain. Probability theory provides a framework for modelling uncertainty. This article starts with an introduction to the probabilistic approach to machine learning and Bayesian inference, and then reviews some of the state-of-the-art in the field. The central thesis is that many aspects of learning and intelligence depend crucially on the careful probabilistic representation of uncertainty. Probabilistic approaches have only recently become a main-stream paradigm in artificial intelligence, roboticsand machine learning [1]. Even now, there is controversy in these fields about how important it is to fully represent uncertainty. For example, recent advances using deep neural networks to solve challenging pattern recognition problems such as speech recognition, image classificationand prediction of words in text [2], do not overtly represent the uncertainty in the structure or parameters of those neural networks. However, my focus will not be on these types of pattern recognition problems, characterised by the availability of large amounts of data, but rather on problems where uncertainty is really a key ingredient, for example where a decision may depend on the amount of uncertainty.

2. Probabilistic Modelling and the Representation of Uncertainty

At a most basic level, machine learning seeks to develop methods for computers to improve theirperformance at certain tasks based on observed data. Typical examples of such tasks might includedetecting pedestrians in images taken from an autonomous vehicle, classifying gene-expressionpatterns from leukaemia patients into subtypes by clinical outcome, or translating English sentencesinto French. However, as we will see, the scope of machine learning tasks is even broader thanthese pattern classification or mapping tasks, and can include optimisation and decision making,compressing data, and automatically extracting interpretable models from data.

Data are the key ingredient of all machine learning systems. But data, even so called “Big Data",is useless on its own until one extracts knowledge or inferences from it. Almost all machinelearning tasks can be formulated as making inferences about missing or latent data from the observed data. We will variously use the terms inference, prediction or forecasting to refer to thisgeneral task. Elaborating the example mentioned above, consider classifying leukaemia patientsinto one of the four main subtypes of this disease, based on each patient's measured gene-expressionpatterns. Here the observed data are pairs of gene-expression patterns and labelled subtypes, andthe unobserved or missing data to be inferred are the subtypes for new patients.

There are many forms of uncertainties in modelling. At the lowest level, model uncertainty isintroduced from measurement noise, e.g., pixel noise or blur in images. At higher levels, amodelmay have many parameters, such as the linear regression, and there is uncertaintyabout which values of these parameters will be good at predicting new data. Finally, at the highestlevels, there is often uncertainty about even the general structure of the model: is linear regressionappropriate or a neural network, if the latter, how many layers, etc.

The probabilistic approach to modelling uses probability theory to express all forms of uncertainty. Probability theory is the mathematical language for representing and manipulating uncertainty, in much the same way as calculus is the language for representing and manipulating rates ofchange. Fortunately, the probabilistic approach to modelling is conceptually very simple: probabilitydistributions are used to represent all the uncertain unobserved quantities in a model (includingstructural, parametric, and noise-related) and how they relate to the data. Then the basic rulesof probability theory are used to infer the unobserved quantities given the observed data. Learning from data occurs through the transformation of the prior probability distributions (definedbefore observing the data), into posterior distributions (after observing data). The application ofprobability theory to learning from data is called Bayesian learning.

Probabilistic modelling also has some conceptual advantages over alternatives as a normative theoryfor learning in artificially intelligent (AI) systems. How should an AI system represent and updateits beliefs about the world in light of data? The Cox axioms dene some desiderata for representingbeliefs; a consequence of these axioms is that `degrees of belief', ranging from “impossible" toabsolutely certain", must follow all the rules of probability theory [3]. This justifies the useof subjective Bayesian probabilistic representations in AI. An argument for Bayesian representationsin AI that is motivated by decision theory is given by the Dutch-Book theorems. The argumentrests on the idea that the strength of beliefs of an agent can be assessed by asking the agent whetherit would be willing to accept bets at various odds (ratios of payos). The Dutch-Book theoremsstate that unless an AI system's (or human's, for that matter) degrees of beliefs are consistentwith the rules of probability it will be willing to accept bets that are guaranteed to lose money[4]. Because of the force of these and many other arguments on the importance of a principledhandling of uncertainty for intelligence, Bayesian probabilistic modelling has emerged not only asthe theoretical foundation for rationality in AI systems but also as a model for normative behaviourin humans and animals and much research is devoted tounderstanding how neural circuitry may be implementing Bayesian inference [5].

3. Flexibility through Non-Parametrics

The best way to understand non-parametric models is through comparison to parametric ones. In aparametric model, there are anumber of parameters, and no matter how much trainingdata are observed, all the data can do is set theseparameters that control futurepredictions. In contrast, nonparametric approaches have predictions that grow in complexity withthe amount of training data, either by considering a nested sequence of parametric models withincreasing numbers of parameters or by starting out with a model with infinitely many parameters.

For example, in a classification problem, whereas a linear (i.e., parametric) classifier will alwayspredict using a linear boundary between classes, a nonparametric classifier can learn a nonlinearboundary whose shape becomes more complex with more data. Many nonparametric models can bederived starting from a parametric model and considering that happens as the model grows to thelimit of infinitely many parameters [6]. Clearly, fitting a model with infinitely many parametersto finite training data would result in\over fitting", in the sense that the model's predictions mightfect quirks of the training data rather than regularities that can be generalised to test data.

3.1 Probabilistic Programming

The basic idea in probabilistic programming it to use computer programs to represent probabilisticmodels[2],[7]. One way to do this is for the computer program to dene a generator fordata from the probabilistic model, i.e., a simulator. This simulator makes calls to arandom number generator in such a way that repeated runs from the simulator would sampledifferent possible data sets from the model. This simulation framework is more general than thegraphical model framework described previously since computer programs can allow constructssuch as recursion (functions calling themselves) and controlo/p statements (e.g., if statementsresulting in multiple paths a program can follow) which are difficult or impossible to represent in a finite graph. In fact, for many of the recent probabilistic programming languages that are based onextending Turing-complete languages (a class that includes almost all commonly-used languages),it is possible to represent any computable probability distribution as a probabilistic program [8].

The full potential of probabilistic programming comes from automating the process of inferringunobserved variables in the model conditioned on the observed data. Conceptually,conditioning needs to compute input states of the program that generate data matching the observeddata. Whereas normally we think of programs running from inputs to outputs, conditioninginvolves solving the inverse problem of inferring the inputs (in particular the random number calls)that match a certain program output. Such conditioning is performed by a universal inferenceengine, usually implemented by Monte Carlo sampling over possible executions over the simulatorprogram that are consistent with the observed data. The fact that defining such universal inferencealgorithms for computer programs is even possible is somewhat surprising, but it is related to thegenerality of certain key ideas from sampling such as rejection sampling, sequential Monte Carlo and \approximate Bayesian computation" [9].

3.2 Bayesian optimisation

Consider the very general problem of finding the global maximum of an unknown function whichis expensive to evaluate (say, evaluating the function requires performing lots of computation, orconducting an experiment). Mathematically, for a function f on a domain X, the goal is to find aglobal maximiser x*:

$$x^* = \arg\max_{x \in \mathcal{X}} f(x).$$

Bayesian optimisation poses this as a problem in sequential decision theory: where should oneevaluate next so as to most quickly maximize f, taking into account the gain in information aboutthe unknown function f [10]? For example, having evaluated at three points measuring thecorresponding values of the function at those points, $f(x_1; f(x_1)); (x_2; f(x_2)); (x_3; f(x_3))g$, whichpoint x should the algorithm evaluate next, and where does it believe the maximum to be? This isa classic machine intelligence problem with a wide range of applications in science and engineering,e.g., from drug design to robotics where the function could be the drug's efficacy or the speed ofa robot's gait respectively. Basically, it can be applied to any problem involving the optimisationof expensive functions; the qualifier expensive" comes because Bayesian optimisation might usesubstantial computational resources to decide where to evaluate next, and these resources have tobe traded with the cost of function evaluations.

3.3 Data Compression

Consider the problem of compressing data so as to communicate it or store it in as few bits aspossible, in such a manner that the original data can be recovered exactly from the compressed data.Methods for such lossless data compression are ubiquitous in information technology, from computerhard drives to data transfer over the internet. Data compression and probabilistic modelling aretwo sides of the same coin, and Bayesian machine learning methods are increasingly advancing thestate of the art in compression. The connection between compression and probabilistic modellingwas established in Shannon's seminal work on the source coding theorem [11] which states that thenumber of bits required to compress data is bounded by the entropy of the probabilitydistribution of the data. All commonly used lossless data compression algorithms (e.g., gzip, etc)can be viewed as probabilistic models of sequences of symbols.

4. Automatically Discovering Interpretable Models from Data

One of the grand challenges of machine learning is to fully automate the process of learning andexplaining statistical models from data. This it the goal of the Automatic Statistician, a system thatcan automatically discover plausible models from data, and explain what it has discovered in plainEnglish [12]. This could be useful to almost any field of endeavour that is reliant on extractingknowledge from data. In contrast to much of the machine learning literature which has beenfocused on extracting increasing performance improvements on pattern recognition problems usingtechniques such as kernel methods, random forests, or deep learning, the Automatic Statisticianneeds to build models that are composed of interpretable components, and to have a principledway of representing uncertainty about model structures given data. It also needs to be able to givereasonable answers not just for big data sets but also for small ones. Bayesian approaches providean elegant way of trading of the complexity of the model and the complexity of the data, andprobabilistic models are compositional and interpretable as described previously [13].

Probabilistic Model for Attributes A probabilistic relational model O specifies a probability distribution overall instantiations K of the relational schema. It consists ofthe qualitative dependency structure, P, and the parameters associated with it, Q*R. The dependency structure isdefined by associating with each attribute $ a set of parents Pa $. Each parent has the form $ S5Twhere Sis either empty or a single slot %. (PRMs alsoallow dependencies on longer slot chains, but we have chosen to omit those for simplicity of presentation.) To understand the semantics of this dependence, note that US5 isa multi-set of values V in S5. We use the notion of aggregation from database theory to define the dependenceon a multi-set; thus, Uwill depend probabilistically onsome aggregate property W8XV. In this paper, we use themedian for ordinal attributes, and the mode (most commonvalue) for others. When V is single-valued, both reduce toa dependence on the value of U S5T.

The quantitative part of the PRM specifies the parameterization of the model. Given a set of parents for an attribute,we can define a local probability model by associating withit a conditional probability distribution (CPD). For each attribute we have a CPD that specifies Y[Z Pa $

Definition 1: A probabilistic relational model (PRM) O fora relational schema P is defined as follows. For each classand each descriptive attribute, we havea set of parents Pa $, and a conditional probability distribution (CPD) that represents Y_^`$[Z Pa $.Given a relational skeleton LAM , a PRM O specifies adistribution over a set of instantiations K consistent with LM :

$$P(\mathcal{I} \mid \sigma_r, \Pi) = \prod_{x \in \sigma_r(X)} \prod_{A \in \mathcal{A}(x)} P(x.A \mid \mathrm{Pa}(x.A))$$

where LCMare the objects of each class as specified bythe relational skeleton LAM (in general we will use the notationL to refer to the set objects of each class as definedby any type of domain skeleton).For this definition to specify a coherent probability distributionover instantiations, we must ensure that our probabilisticdependencies are acyclic, so that a random variabledoes not depend, directly or indirectly, on its own value.Moreover, we want to guarantee that this will be the casefor any skeleton. For this purpose, we use a class dependencygraph, which describes all possible dependenciesamong attributes. In this graph, we have an (intra-object) edge $T $ if $T is a parent of . If $%T is a parent of $, and 9 021=3C4*6*+%*-, we have an (inter-object)edge 9 T . If the dependency graph of is acyclic, then it defines a legal model for any relationaP lskeleton L M [14].

Definition 2: A probabilistic relational model O with referenceuncertainty has the same components as in Definition1. In addition, for each reference slot %] #with0;1=35476*+%7-N 9 , we have:

a set of attributes Y,+-.0/5.0/02"3+%*-74 9>;

$$P(\mathcal{I} \mid \sigma_o, \Pi) = \prod_{x \in \sigma_o(X)} \prod_{A \in \mathcal{A}(x)} P(x.A \mid \text{Pa}(x.A))$$
$$\prod_{\rho \in \mathcal{R}(x), \text{Range}[\rho]=Y} \frac{P(x.S_\rho = v[x.\rho] \mid \text{Pa}(x.S_\rho))}{|Y_v|}$$

a new selector attribute V 6 within which takes onvalues in the cross-product space Y,+-".0/1.0/02"3+%*-;a set of parents and a CPD for V 6.To define the semantics of this extension, we must definethe probability of reference slots as well as descriptive attributes:

5. Conclusion

We evaluated the methods on several real-life data sets,comparing standard PRMs, PRMs with reference uncertainty(RU), and PRMs with existence uncertainty (EU).Our experiments used the Bayesian score with a uniformDirichlet parameter prior with equivalent sample size , and a uniform distribution over structures. We first tested whether the additional expressive power allowsus to better capture regularities in the domain. Towardthis end, we evaluated the likelihood of test data given ourlearned models. Unfortunately, we cannot directly comparelikelihoods, since the PRMs involve different sets ofprobabilistic events. Instead, we compare the two variantsof PRMs with structural uncertainty, EU and RU, to “baseline”models which incorporate link probabilities, but makethe “null” assumption that the link structure is uncorrelatedwith the descriptive attributes. For reference uncertainty,the baseline has +%7-8for each slot. For existence uncertainty,it forces U % to have no parents in the model.

References

1. Murphy, K. P. Machine learning: a probabilistic perspective (MIT press, 2012).
2. Hinton, G. et al. Deep neural networks for acoustic modeling in speech recognition: Theshared views of four research groups. Signal Processing Magazine, IEEE 29, 82{97 (2012).
3. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutionalneural networks. In Advances in neural information processing systems, 1097{1105 (2012).
4. Sermanet, P. et al. Overfeat: Integrated recognition, localization and detection using convolutionalnetworks. In International Conference on Learning Representations (ICLR 2014)(arXiv preprint arXiv:1312.6229, 2014).
5. Ghahramani, Z. Bayesian nonparametrics and the probabilistic approach to modelling. Philo-sophical Transactions of the Royal Society A 371, 20110553 (2013).
6. Marcus, G. F. & Davis, E. How robust are probabilistic models of higher-level cognition? Psychological Science 24, 2351{2360 (2013).
7. Goodman, N. D. et al. Relevant and robust a response to Marcus and Davis (2013). Psychological Science 0956797614559544 (2015).
8. Doya, K., Ishii, S., Pouget, A. & Rao, R. P. N. Bayesian Brain: Probabilistic Approaches to Neural Coding (The MIT Press, 2007).
9. Neal, R. M. MCMC using Hamiltonian dynamics. In S. Brooks A. Gelman, G. J. &Meng, X.-L. (eds.) Handbook of Markov Chain Monte Carlo (Chapman & Hall / CRC Press, 2010).

10. Girolami, M. &Calderhead, B. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73, 123{214 (2011).
11. Sutskever, I., Vinyals, O. & Le, Q. V. Sequence to sequence learning with neural networks. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. & Weinberger, K. (eds.) Advances in Neural Information Processing Systems 27, 3104{3112 (Curran Associates, Inc., 2014).
12. Orbanz, P. &Teh, Y. W. Bayesian nonparametric models. In Encyclopedia of MachineLearning (Springer, 2010).
13. Hjort, N., Holmes, C., M uller, P. & Walker, S. (eds.) Bayesian Nonparametrics (Cambridge University Press, 2010).
14. Lu, C. & Tang, X. Surpassing human-level face verification performance on LFW with GaussianFace. In Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI)(2015).