Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence

Decision-Theoretic Approximations for Machine Learning

M. Ehsan Abbasnejad

Abstract

Decision theory focuses on the problem of making decisions under uncertainty. This uncertainty arises from the unknown aspects of the state of the world the decision maker is in or the unknown utility function of performing actions. The uncertainty can be modeled as a probability distribution capturing our belief about the world the decision maker is in. Upon making new observations, the decision maker becomes more confident about this model. In addition, if there is a prior belief on this uncertainty that may have obtained from similar experiments, the Bayesian methods may be employed. The loss incurred by the decision maker can also be utilized for the optimal action selection. Most machine learning algorithms developed though focus on one of these aspects for learning and prediction; either learning the probabilistic model or minimizing the loss. In probabilistic models, approximate inference, the process of obtaining the desired model from the observations when its is not tractable, does not consider the task loss. On the other end of the spectrum, the common practice in learning is to minimize the task loss without considering the uncertainty of prediction model. Therefore, we investigate the intersection of decision theory and machine learning considering both uncertainty in prediction model and the task loss.

1 Introduction

Decision theory [Berger, 1985] studies optimal decision under uncertainty. Statistical knowledge obtained from the environment can be used to model this uncertainty. There are various sources of uncertainty that should be considered, such as:

1. **State**: The state of the world where the decision maker is in may be noisy or not fully known. This uncertainty reduces our confidence on the optimal decision that has to be made at any given point. In machine learning the state is usually represented by the parameters of the statistical model.
2. **Utility**: A utility function measures how much performing an action in a given state is of "value" to the decision maker. If this function is unknown or can't be quantified, there is an inherent uncertainty about which decision value is better in a given state.

To model this uncertainty, or the belief on what is known in the world for the decision maker, we resort to the calculus of uncertainty and represent it as a *probability distribution*. This distribution can be derived from the sample data in a typical statistical approach. Other than the sample data, the decision maker can utilize the available information from similar experiments or use problem-specific knowledge of the problem such as the preference of some states to the others. This information is known as the *prior* information. The *posterior* can then be obtained by updating the probabilistic model using the observations and the prior through a process known as *inference*. If inference is performed so that a belief over the state values is found, then the *Bayesian view* through Bayesian methods has been employed. Otherwise, the most likely state is used instead corresponding to the *frequentists view*. In many real-world problems however, exact inference is not possible due to computational cost or intractability; therefore, approximation of the inference is required. It should additionally be noted that one can choose not to incorporate the prior and only consider building a model of the environment from what is observed.

Another type of information that can be used by the decision maker is the knowledge about the consequences of performing an action. This knowledge can be quantified in a *loss* function that measures the output of a decision in a given state of the world. In a simple representation, loss can be thought of as the negative of the utility function. That is, in a given state, the decision that has the best utility corresponds to the minimum loss. The loss function is commonly assumed to be symmetric, that is, the right or wrong predictions have similar cost for the decision maker. In many cases however, we might be interested in asymmetric losses; for instance, consider the train wheel crack detection task where false detection will cost a wheel to be replaced while failing to recognize the crack may lead to a catastrophic accident.

Bayesian decision theory [Bernardo, Jose M ; Smith, 2000] seeks to utilize Bayesian methods to model the uncertainty and the loss function to make an optimal decision. However, the inference in Bayesian methods has developed in-

dependent of the problem-at-hand's loss function (task loss), in other words, it is not *loss-calibrated*. Specially, in machine learning where the final task is prediction, Bayesian decision theory has not been well-utilized. Even in cases where the prior is opted not to be considered, obtaining the optimal model and its parameters in machine learning is challenging. In the following sections we will investigate various aspects of Bayesian decision theory in machine learning and its impacts and applications.

2 Decision Theory for Machine Learning

The concept of risk minimization and the probabilistic representation that was discussed have been of great importance in machine learning community. In this section we detail the overlapping aspects of these two areas.

For machine learning, the decision maker uses a "procedure" that operates on data to produce a decision, then the state and action space can be the same in machine learning. This procedure is a (hypothesis) function of the input data $h \in \mathcal{H}$ where $\mathcal{H}$ is the class of functions we select our hypothesis from. This hypothesis class can be linear, nonlinear functions or the approximated density of the input. Even though decision theory and machine learning are very closely related, machine learning community traditionally focuses on one aspect of the problem either loss incurred from mistakes or distribution of the data for simplicity or the probabilistic representation.

3 Non-parametric Bayesian Methods

Non-parametric Bayesian analysis has gained contemporary attention due to their flexibility in modeling complex phenomena where there are some parameters that we are not able to determine in advance. Examples of these methods are Dirichlet Processes, Gaussian Processes, Latent Dirichlet Allocation and Indian Buffet Process. These methods are distributions with infinitely many parameters that can equivalently be thought of as distributions in function spaces.

Typically approximate inference or sampling is performed in non-parametric Bayesian methods to obtain the posterior. In some cases like the regression in Gaussian processes the posterior is obtained in the closed form. However, the Gaussian process classification is intractable and hence approximation is required.

We can represent the uncertainty of the utility function by a distribution over an unknown utility value as discussed in what is known as *expected expected utility* [Boutilier, 2003]. This is crucial in many real-world problems where the true utility function is not known. This principle can interestingly be jointly used with Gaussian processes where the latent function can be thought of as the unknown utility. Having a distribution over the utility function, we can define additional objectives such as the sparsification objective for Gaussian processes based on these utility values.

In case the loss function is known, the risk can directly be minimized as it is done in where the quantile regression, i.e. estimating the quantile of conditional distribution, is tackled. The loss function measures the difference between the true label and the predicted quantile which leads to a convex function through the convex losses.

4 Conclusion

Machine learning has long utilized decision theory as a sound foundation for decision making under uncertainty. A probabilistic view of the world provides natural representation for many real world problems. The probabilistic representation can lead to point estimation of the parameters as in the frequentists view or learning a belief about them in Bayesian perspective. This probabilistic representation in addition to the loss incurred upon making mistakes provide the solid foundation for decision making. Even though decision-theoretic principles have inspired in machine learning, there has been a little work that utilizes both the loss function and the distribution over the parameter values in learning and prediction tasks.

This amounts to various aspects in machine learning that have disregard either the probabilistic representation or the loss incurred. For example Bayesian inference as the most important aspect of Bayesian learning has been done regardless of the task loss. On the other hand popular frameworks such as ERM simply assumes a uniform distribution over data and focuses on minimizing the average loss.

To summarize in this thesis we seek to investigate Bayesian decision theory in machine learning and in particular look into the following two aspects:

1. **How decision-theoretic view of probabilistic modeling will impact machine learning where the objective is task-specific prediction?**
2. **How inference in probabilistic learning can be performed with respect to the task loss?**

References

[Berger, 1985] James Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, 2. ed edition, 1985.

[Bernardo, Jose M ; Smith, 2000] Adrian Bernardo, Jose M ; Smith. *Bayesian Theory*. Wiley Series in Probability and Statistics, 2000.

[Boutilier, 2003] Craig Boutilier. On the foundations of expected expected utility. *IJCAI'03 Proceedings of the 18th international joint conference on Artificial intelligence*, 2003.