Deriving the Expected Utility of a Predictive Model When the Utilities Are Uncertain

Gregory F. Cooper, M.D., Ph.D. and Shyam Visweswaran, M.D., M.S.
Center for Biomedical Informatics and the Intelligent Systems Program
University of Pittsburgh, Pittsburgh, Pennsylvania

Abstract

Predictive models are often constructed from clinical databases with the goal of eventually helping make better clinical decisions. Evaluating models using decision theory is therefore natural. When constructing a model using statistical and machine learning methods, however, we are often uncertain about precisely how a model will be used. Thus, decision-independent measures of classification performance, such as the area under an ROC curve, are popular. As a complementary method of evaluation, we investigate techniques for deriving the expected utility of a model under uncertainty about the model's utilities. We demonstrate an example of the application of this approach to the evaluation of two models that diagnose coronary artery disease.

1. Introduction

This paper is concerned with how to evaluate the performance of clinical prediction models, such as models used for risk assessment, diagnosis, and prognosis. In particular, we focus on evaluating the performance of models that predict a probability distribution over a discrete outcome variable given a set of clinical features about a patient.

There are many measures for evaluating classification performance (Hand, 1997). Historically, accuracy has been a commonly used measure of classification performance in machine learning. More recently, researchers have increasingly used Receiver Operating Characteristic (ROC) curves (Weinstein & Fineberg, 1980; Provost, Fawcett, & Kohavi, 1998) and their variants and extensions to evaluate classification performance. ROC curves provide an estimate of the various possible sensitivities and specificities of a model in predicting a binary outcome over the range of threshold probabilities.

While ROC curves evaluate the discriminative performance of a model, they do not assess model calibration. A model is well calibrated if the probability predicted for an outcome corresponds closely to the empirical frequency of that outcome. If, for example, a model predicts that $P(rain \mid barometric\ pressure = 29) = 60\%$, then when the barometric pressure is 29, rain empirically occurs about 60% of the time. There are numerous measures for assessing calibration, with the Hosmer-Lemeshow goodness-of-fit measure being a popular one (Hosmer & Lemeshow, 1980). Essentially, these measures indicate how far the predicted probabilities of an outcome variable deviate from the associated empirical frequencies.

In decision making that is based on decision analysis, the expected utility of a model generally depends on both its discriminative performance and its calibration. We might separately assess a model's classification performance and its calibration, and then combine those assessments in some way to obtain an overall measure of performance. Alternatively, we could directly estimate an overall measure of performance of the model, which is the approach taken in this paper. In particular, this paper concentrates on evaluating the expected value of models with an eye toward how they might be used to make decisions, which is often the ultimate purpose for constructing predictive models.

2. Background

Figure 1 shows a simple decision tree (Weinstein & Fineberg, 1980) for a binary decision D with options d_1 and d_2, and a binary outcome R (for result) with values r_1 and r_2. The parameter p equals $P(R = r_1 \mid \boldsymbol{x}, D = d_1)$, where $\boldsymbol{x}$ is a set of variables with assigned states (i.e., features) that are used to predict R. The parameter q equals $P(R = r_1 \mid \boldsymbol{x}, D = d_2)$. The parameters u_{11}, u_{12}, u_{21}, and u_{22} are utilities associated with each of the four branches of the tree. For example, u_{12} denotes the utility of making decision d_1 and having outcome r_2 occur. The utilities are real numbers that range in value from 0 (worst outcome) to 1 (best outcome). The expected utility (EU) of taking decision option d_1 is $EU(d_1) = p \cdot u_{11} + (1 - p) \cdot u_{12}$. Similarly, for decision option d_2, $EU(d_2) = q \cdot u_{21} + (1 - q) \cdot u_{22}$. For a given patient case, if $EU(d_1) > EU(d_2)$, then d_1 is the optimal decision; if $EU(d_2) > EU(d_1)$, then d_2 is the optimal decision; otherwise both decisions are equally optimal. If, in a given case, decision option d_i is taken and the outcome r_j occurs, the utility of having made d_i in light of r_j is u_{ij}, which can be interpreted as the decision-theoretic measure (score) for that decision-outcome combination. Suppose we use prediction model M to derive $P(R = r_1 \mid \boldsymbol{x}, D)$. If a set T of test cases are evaluated, the sum of the utility scores of the cases divided by $|T|$ is an estimate of the expected utility of M applied to such cases.

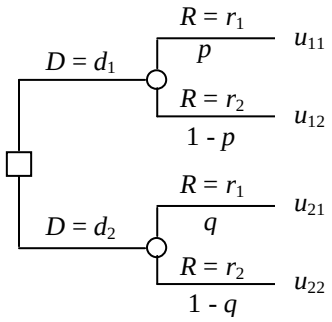

Figure 1. A simple decision tree.

As is known, ROC curves are an imperfect indicator of expected utility, even qualitatively. Indeed, each point on an ROC curve corresponds to a probability decision threshold, which in turn corresponds to a unique equivalence class of utilities. The ROC curve makes no commitment to any particular set of utilities, which can either be a strength or a weakness, depending on the purpose for evaluating a model. Empirically, Moons et al. (1997) provide examples based on medical data in which two models have virtually identical AUROCs but significantly different expected utilities, according to the utility model considered. In contrast, two other models have substantially different AUROCs, but similar expected utilities.

Researchers have applied expected utility as a measure for evaluating predictive models. One example is the evaluation of the Pathfinder system in the early 1990s by Heckerman and Nathwani (1992). More recently, cost-based (or utility-based) performance measures have been studied increasingly in machine learning (Dietterich, et al., 2000). Expected utility is a natural evaluation measure if there is a well-defined decision problem at hand and the relevant utilities can be assessed. In developing a predictive model to apply to future patient cases, however, we may not have precisely defined utilities. One approach to addressing this problem is to represent the utilities as uncertain quantities, that is, as random variables.

Adams and Hand (1999) describe a method for comparing classifiers when misclassification costs are uncertain. They lucidly describe the advantages of treating utilities as random variables in developing an evaluation measure for predictive models. Ultimately, however, the measure they develop is not an expected utility, although it is related to expected utility.

The primary purpose of this paper is to investigate expected utility as an evaluation measure under uncertainty about the utilities. To be clear at the outset, we are not advocating the replacement of other measures of model performance, such as ROC curves and goodness-of-fit measures. Rather, our goal is to provide a complementary measure of performance that emphasizes future decision making uses of a model, while not requiring a firm commitment (on the part of a model-building researcher) to an exact decision making context or set of utilities. The performance measure we describe provides an estimate of how well a model will perform under uncertainty about how it will be used to make decisions in the future.

3. Methods

In this section we first describe a general approach for deriving the expected utility for a model, under uncertainty about the utilities. Next, we introduce an example decision problem that involves diagnosing coronary artery disease. Finally, we derive the expected utility of using a neural network model and a simple Bayes model to diagnose coronary artery disease on a set of patient cases. To do so, we use a numerical method that is straightforward to implement and practical to apply to decision problems that have only a few decision options and outcomes.

3.1. General Formulation

We can generalize the binary decision model in Figure 1 to a model with m decision options and n outcome values.[1] Equation 1 shows the expected utility (EU) of decision option d_i in light of a given set of features $\boldsymbol{x}$.

$$EU(d_i \mid \boldsymbol{x}) = \sum_j P(r_j \mid \boldsymbol{x}, do(d_i)) \cdot u_{ij} \qquad (1)$$

The notation $do(d_i)$ means that the probability in Equation 1 represents the causal effect on outcome r_j of taking (*do*ing) decision option d_i in the context of features $\boldsymbol{x}$.

In a standard decision analysis, there is at least one best possible outcome that is assigned a utility of 1 and at least one worst possible outcome that is assigned a utility of 0. All other outcomes are then assigned values between 0 and 1. The optimal decision is the one that maximizes Equation 1, namely, $\arg\max_{d_i}(EU(d_i \mid x))$, and the expected utility of that decision option is $\max_{d_i}(EU(d_i \mid x))$. The ultimate utility of that decision, in light of the *actual* outcome, is discussed in Equation 4 below.

Standard decision analysis assumes a well-defined decision problem and a decision maker who assesses all the utilities precisely. When constructing a general-purpose predictive model, however, we often are uncertain about both the problem (which

[1] We could further generalize to consider a sequence of decisions, rather than a single decision and to represent continuous-valued decisions and outcomes, but we will not do so here.

reflects uncertainty about how the model will be used in the future to solve problems) and the utilities (which reflect uncertainty about the preferences of some future, unknown decision maker). In this paper we focus on dealing with the uncertainty about utilities, and we assume that a useful decision problem can be envisioned at the time a model is constructed; we show an example in Section 4.[2] In summary, we concentrate in this paper on evaluating the expected utility of models that estimate the probability term in Equation 1.

Let $U = \{u_{11}, u_{12}, ..., u_{1n}, u_{21}, u_{22}, u_{2n}, ..., u_{mn}\}$ be the set of utilities in the decision model. Equation 2 modifies Equation 1 to explicitly include U and to specify that the probability function reflects the predictions of a particular model M.

$$EU(d_i \mid \boldsymbol{x}, M, U) = \sum_{j=1}^{n} P(r_j \mid \boldsymbol{x}, M, do(d_i)) \cdot u_{ij} \qquad (2)$$

Equation 3 shows the decision option that has the maximum expected utility when we use model M to generate probabilities of outcome R.

$$f(M, U, \boldsymbol{x}) = \arg\max_{i=1}^{m}(EU(d_i \mid \boldsymbol{x}, M, U)) \qquad (3)$$

Let c be a patient case in the test dataset, consisting of a feature set $\boldsymbol{x}$ and an outcome R. Let $g(c)$ be a function whose value equals j, if and only if $R = r_j$ in case c. For example, in a test dataset of 100 cases, if the value of R in case 3 is r_2, then g(case 3) = 2. Let $\boldsymbol{x}_c$ be the features associated with case c. For example, if the only feature is *age*, and case 3 has *age* = 45, then $\boldsymbol{x}_3$ = (*age* = 45). Equation 4 expresses the utility of the decision given by Equation 3 in light of the actual outcome in case c, as given by $g(c)$.

$$u(M, U, c) = u_{f(M,U,\boldsymbol{x}_c),g(c)} \qquad (4)$$

Let the probability distribution $P(U \mid \boldsymbol{x})$ represent the uncertainty about the values of the utilities, given feature set $\boldsymbol{x}$. Thus, $P(U \mid \boldsymbol{x})$ encodes our belief about the utilities of a future decision maker who we envision using model M in making decision D. Since we are uncertain about the utility values, we integrate Equation 4 over the probability of the joint values of the utilities, as shown in Equation 5.

$$EU(M, c) = \int_U u(M, U, c) \cdot P(U \mid \boldsymbol{x}_c)\, dU \qquad (5)$$

Equation 6 shows the expected utility of applying model M in making decisions for the test cases in T. $EU(M,T)$ can also be interpreted as the expected utility of using M in making decisions in the future on cases drawn from the distribution that generated the cases in T. If we wish to select a predictive model for an envisioned future decision maker to use in cases such as T, we should select the model M^* for which $EU(M^*,T)$ is maximum relative to all the models considered.

$$\begin{aligned} EU(M, T) &= \int_U \left(\frac{1}{|T|} \sum_{c \in T} u(M, U, c)\right) \cdot P(U \mid x_c)\, dU \\ &= \frac{1}{|T|} \sum_{c \in T} \int_U u(M, U, c) \cdot P(U \mid x_c)\, dU \end{aligned} \qquad (6)$$

3.2. An Example

In this section we describe an example that illustrates the application of Equation 6 to a particular decision problem. In doing so, we discuss the key ideas needed to apply that equation in other decision problems.

The example clinical problem involves diagnosing coronary artery disease (CAD). We use outcome r_1 to represent *no CAD at the present time* and r_2 to represent *CAD at the present time*. We use d_1 to represent *a diagnosis of no CAD*, and d_2 to represent *a diagnosis of CAD*, again, at the present time. We assume the best situation (from the patient or patient-surrogate perspective) is to diagnose no CAD and have no CAD, and therefore $u_{11} = 1$. We assume that diagnosing the absence of CAD when a patient actually has CAD is the worst outcome ($u_{12} = 0$), because such a patient may not receive timely treatment for a serious disease. We also assume that $u_{21} > u_{22}$, which expresses that it is better to not have CAD and be misdiagnosed as having it, than to have CAD and be correctly diagnosed as having it. Subject to these constraints, for the purpose of illustration (and without loss of generality) we assume a uniform prior over all values of u_{21} and u_{22}, which expresses our uncertainty about the values these utilities will have in future applications of this CAD decision model.

We assume $p = q$ in Figure 1, because the diagnosis of CAD (or, alternatively, of no CAD) at the present time will not influence whether the patient has CAD at the present time. The diagnosis can, however, influence subsequent patient states, and such influence is captured by the utilities in the model. Making the above assumptions transforms Figure 1 to Figure 2. Following Figure 2, Equation 7 specializes Equation 6 to integrate over just the two uncertain utilities in the example problem.

$$EU(M, T) = \qquad (7)$$

$$\frac{1}{|T|} \sum_{c \in T} \int_{u_{21}, u_{22}} u(M, [u_{21}, u_{22}], c) \cdot P(u_{21}, u_{22})\, du_{21} du_{22},$$

[2] While modeling the uncertainty of the structural decision problem is beyond the scope of this paper, it is an interesting problem for future research.

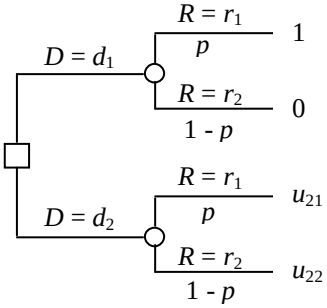

Figure 2. The decision tree used for the CAD decision problem example.

where the integral is over all joint values of u_{11} and u_{12}, and the square brackets enclose a list of only the uncertain utilities and leave implicit u_{11} and u_{12}, which are constant.

For simplicity, we applied a straightforward numerical integration method to estimate Equation 7 as follows:

$$EU(M,T) = \frac{1}{|T|}\sum_{c\in T}\sum_{u_{21}=0}^{1 by k}\sum_{u_{22}=0}^{1 by k} u(M,[u_{21},u_{22}],c)\cdot P(u_{21},u_{22}), \quad (8)$$

where k is a fraction that indicates how the utility variables are incremented from 0 to 1 in the sums. For example, for the CAD problem, we used k = 0.01, and therefore the utility variables in the inner two sums of Equation 8 took on the values 0, 0.01, 0.02, …, 1.0.

For the CAD problem, when $u_{21} > u_{22}$ then $P(u_{21}, u_{22}) = b$, where b is a constant; otherwise, $P(u_{21}, u_{22}) = 0$. Since $P(u_{21}, u_{22})$ must sum to 1.0 over all the values of u_{21} and u_{22} in Equation 8, it follows that b = 1/5050 in that equation.

It is straightforward to record the joint values of u_{21} and u_{22} that yield the maximum and minimum values of $EU(M, T)$, as derived by Equation 8. We report such values in Section 4.

The advantage of applying the computational approach illustrated by Equation 8 is that it is (1) easy to implement, (2) adequately efficient for simple decision problems, and (3) provides the flexibility to use an arbitrary probability distribution on the unknown utilities in the sum. Nonetheless, more complex decision problems will likely require more sophisticated methods for approximating the integral in Equation 6, such as Monte Carlo integration methods.

4. Results

In this section we illustrate the application of the methods described in Section 3 to the CAD decision problem. We first introduce the dataset and machine-learning methods that we used, and then present the results.

4.1. Dataset

We used data that were collected at the Cleveland Clinic Foundation by Robert Detrano, M.D., Ph.D. and that are available from the heart disease directory of the UCI Machine Learning repository (www.ics.uci.edu/~mlearn/MLRepository.html). A primary reason for choosing this dataset is that it is non-proprietary and publicly available, and thus, other researchers can validate and extend the results we report here, using that dataset. The data consist of 303 patient cases, 139 of whom were diagnosed with coronary artery disease (CAD) based on angiographic evidence of at least 50% narrowing of one or more coronary blood vessels. The remaining 164 cases were considered to have no CAD. The values of 13 variables were recorded for each patient, representing demographic, symptom, sign, and laboratory information.

4.2. Machine Learning Methods

We split the data randomly into a training set of 210 (~70%) cases and a test set of 93 cases (~30%). As examples of predictive models that have been applied frequently in machine learning, we chose to learn Neural Network (NN) and Simple Bayes (SB) models. We emphasize that our purpose is not to analyze the performance of these models, per se, but rather to use them to illustrate the application of the methods described in Section 3. To learn the NN and SB models, we used software that is available in the Weka library (Weka v3.3.6, www.cs.waikato.ac.nz/ml/weka) with default learning settings. For each method, we induced a predictive model from the training set that we then applied to the test set to obtain the results reported below.

4.3. Experimental Results

Figure 3 shows ROC curves for each of SB and NN. The results described in the caption suggest that the two methods are only borderline statistically significantly different when using a 95% confidence interval (CI), which we computed using a bootstrap method with 1000 samples.

Table 1 shows the EU results of NN and SB, as well as the expected utility for an optimal (Opt) model. An optimal (Opt) model produces the correct prediction with probability 1. As with the ROC analysis, while SB performs better than NN, the statistical significance of that difference is tenuous, as shown by the 95% confidence intervals (CI) that were derived using bootstrap sampling.

As an indication of the computation times required, it took an average of 22,840 ms (averaged over 10 runs) to derive the EU of the NN model using the numerical method given by Equation 8. We used

a desktop computer with a 500 Mz Pentium III processor and 384 MB of RAM.

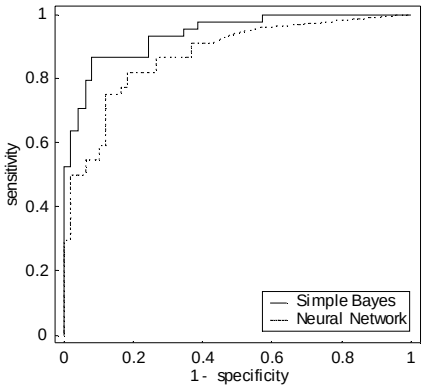

Figure 3. The AUROC for Simple Bayes is 0.9374 [95% CI: 0.8648, 0.9760] and for the Neural Network it is 0.8711 [95% CI: 0.7854, 0.9315].

Table 1. The EU of various models for the example CAD decision problem.

Method	Numerical *EU*	95% CI
Opt	0.6814	[0.6162, 0.7466]
NN	0.6178	[0.5386, 0.6984]
SB	0.6455	[0.5712, 0.7199]
Opt - NN	0.0636	[-0.0429, 0.1676]
Opt - SB	0.0359	[-0.0586, 0.1325]
SB - NN	0.0276	[-0.0799, 0.1348]

Table 2 lists the values of u_{21} and u_{22} that maximize and that minimize the absolute or relative EUs. Note that for the test set that we used, the minimum EU for SB - NN is 0, meaning that there are no utility settings for which NN performs better than SB; this insight cannot be gleaned from the ROC curve in Figure 3.

Table 2. The maximum and minimum EUs of various models for the CAD decision problem.

Method	Max EU	u_{21}, u_{22}	Min EU	u_{21}, u_{22}
Opt	0.9905	0.99, 0.98	0.5269	any, 0
NN	0.9258	0.99, 0.98	0.5084	0.06, 0.05
SB	0.9875	0.99, 0.98	0.5260	0.23, 0.03
Opt - NN	0.1189	0.65, 0.64	0	any, 0
Opt - SB	0.0739	0.40, 0.39	0	any, 0
SB - NN	0.0761	0.93, 0.92	0	any, 0

As a form of visual sensitivity analysis, Figure 4 plots the differences in EUs between SB and NN. A darker color indicates better performance of SB over NN; white indicates no difference. Figure 4 shows (1) that NN does relatively poorly when u_{22} is high and u_{21} is high (dark areas), and (2) NN and SB perform about the same when u_{22} is low (light areas).

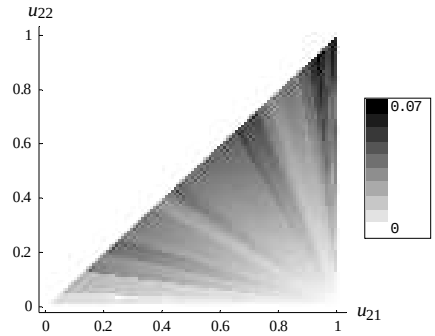

Figure 4. Plot of *EU*(SB) – *EU*(NN) with the constraint $u_{21} > u_{22}$.

5. Summary

In this paper we introduced a basic approach to evaluating predictive models based on expected utility as an evaluation measure. This approach emphasizes two main points: (1) the importance of our considering future decision-making uses of predictive models being constructed, and (2) the explicit incorporation of our uncertainty about such future uses into the evaluation of a model.

We applied the approach to an example decision problem. The EU results were consistent with the results of an ROC analysis, but also revealed interesting insights that would not be apparent in an ROC analysis.

Acknowledgements

This research was supported by grants from the National Library of Medicine (R01-LM008374) and the National Science Foundation (IIS-0325581).

References

Adams NM, Hand DJ (1999). Comparing classifiers when the misallocation costs are uncertain. *Pattern Recognition, 32*, 1139-1147.

Dietterich T, Margineantu D, Provost F, Turney P (Eds.). (2000). *Workshop on Cost-Sensitive Learning.*

Hand DJ (1997). *Construction and Assessment of Classification Rules*. Chichester, UK: Wiley.

Heckerman DE, Nathwani B (1992). An evaluation of the diagnostic accuracy of Pathfinder. *Computers and Biomedical Research, 15*, 56-79.

Hosmer DW, Lemeshow S (1980). A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics, A10*, 1043-1069.

Moons KGM, et al. (1997). Application of treatment thresholds to diagnostic-test evaluation: An alternative to the comparison of areas under receiver operating characteristic curves. *Medical Decision Making, 17*, 447-454.

Provost F, Fawcett T, Kohavi R (1998). The case against accuracy for comparing induction algorithms. In: *Proceedings of the International Conference on Machine Learning.*

Weinstein MC, Fineberg HV (1980). *Clinical Decision Analysis.* Philadelphia, PA: W.B. Saunders.