Relationship between Loss Functions and Confirmation Measures

Krzysztof Dembczyński[1] and Salvatore Greco[2] and Wojciech Kotłowski[1] and Roman Słowiński[1,3]

[1] Institute of Computing Science, Poznań University of Technology,
60-965 Poznań, Poland
{kdembczynski, wkotlowski, rslowinski}@cs.put.poznan.pl
[2] Faculty of Economics, University of Catania, 95129 Catania, Italy
salgreco@unict.it
[3] Institute for Systems Research, Polish Academy of Sciences, 01-447 Warsaw, Poland

Abstract. In the paper, we present the relationship between loss functions and confirmation measures. We show that population minimizers for weighted loss functions correspond to confirmation measures. This result can be used in construction of machine learning methods, particularly, ensemble methods.

1 Introduction

Let us define the prediction problem in a similar way as in [4]. The aim is to predict the unknown value of an attribute y (sometimes called *output*, *response variable* or *decision attribute*) of an object using the known joint values of other attributes (sometimes called *predictors*, *condition attributes* or *independent variables*) $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. We consider binary classification problem, in which we assume that $y \in \{-1, 1\}$. All objects for which $y = -1$ constitute decision class Cl_{-1}, and all objects for which $y = 1$ constitute decision class Cl_1. The goal of a learning task is to find a function $F(\mathbf{x})$ (in general, $F(\mathbf{x}) \in \Re$) using a set of training examples $\{y_i, \mathbf{x}_i\}_1^N$ that predicts accurately y (in other words, classifies accurately objects to decision classes). The optimal classification procedure is given by:

$$F^*(\mathbf{x}) = \arg\min_{F(\mathbf{x})} E_{y\mathbf{x}} L(y, F(\mathbf{x})), \tag{1}$$

where the expected value $E_{y\mathbf{x}}$ is over joint distribution of all variables $(y, \mathbf{x})$ for the data to be predicted. $L(y, F(\mathbf{x}))$ is a loss or cost for predicting $F(\mathbf{x})$ when the actual value is y. $E_{y\mathbf{x}} L(y, F(\mathbf{x}))$ is often called *prediction risk*. Nevertheless, the learning procedure can use only a set of training examples $\{y_i, \mathbf{x}_i\}_1^N$. Using this set, it tries to construct $F(\mathbf{x})$ to be the best possible approximation of $F^*(\mathbf{x})$. The typical loss function in binary classification tasks is, so called, 0-1 loss:

$$L_{0-1}(y, F(\mathbf{x})) = \begin{cases} 0 & \text{if } yF(\mathbf{x}) > 0, \\ 1 & \text{if } yF(\mathbf{x}) \leq 0. \end{cases} \tag{2}$$

It is possible to use other loss functions than (2). Each of these functions has some interesting properties. One of them is a population minimizer of prediction risk. By conditioning (1) on $\mathbf{x}$ (i.e., factoring the joint distribution $P(y, \mathbf{x}) = P(\mathbf{x})P(y|\mathbf{x})$), we obtain:

$$F^*(\mathbf{x}) = \arg\min_{F(\mathbf{x})} E_{\mathbf{x}} E_{y|\mathbf{x}} L(y, F(\mathbf{x})). \tag{3}$$

It is easy to see that it suffices to minimize (3) pointwise:

$$F^*(\mathbf{x}) = \arg\min_{F(\mathbf{x})} E_{y|\mathbf{x}} L(y, F(\mathbf{x})). \tag{4}$$

The solution of the above is called *population minimizer*. In other words, this is an answer to a question: what does a minimization of expected loss estimate on a population level? Let us remind that the population minimizer for 0-1 loss function is:

$$F^*(\mathbf{x}) = \begin{cases} 1 & \text{if } P(y = 1|\mathbf{x}) \geq \frac{1}{2}, \\ -1 & \text{if } P(y = -1|\mathbf{x}) > \frac{1}{2}. \end{cases} \tag{5}$$

From the above, it is easy to see that minimizing 0-1 loss function one estimates a region in predictor space in which class Cl_1 is observed with the higher probability than class Cl_{-1}. Minimization of some other loss functions can be seen as an estimation of conditional probabilities $P(y = 1|\mathbf{x})$ (see Section 2).

From the other side, Bayesian confirmation measures (see, for example, [5,9]) have paid a special attention in knowledge discovery [7]. Confirmation measure $c(H, E)$ says in what degree a piece of evidence E confirms (or disconfirms) a hypothesis H. It is required to satisfy:

$$c(H, E) = \begin{cases} > 0 & \text{if } P(H|E) > P(H), \\ = 0 & \text{if } P(H|E) = P(H), \\ < 0 & \text{if } P(H|E) < P(H), \end{cases} \tag{6}$$

where $P(H)$ is the probability of hypothesis H and $P(H|E)$ is the conditional probability of hypothesis H given evidence E. In Section 3, two confirmation measures of a particular interest are discussed.

In this paper, we present relationship between loss functions and confirmation measures. The motivation of this study is a question: what is the form of the loss function for estimating a region in predictor space in which class Cl_1 is observed with the positive confirmation, or alternatively, for estimating confirmation measure for a given $\mathbf{x}$ and y? In the following, we show that population minimizers for *weighted* loss functions correspond to confirmation measures. Weighted loss functions are often used in the case of imbalanced class distribution, i.e., when probabilities $P(y = 1)$ and $P(y = -1)$ are substantially different. This result is described in Section 4. The paper is concluded in the last section.

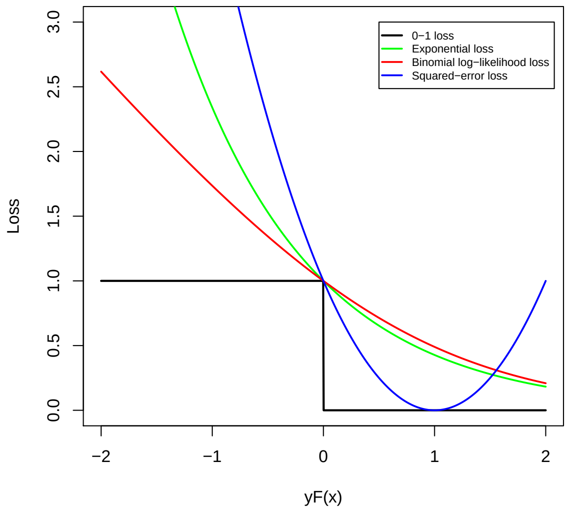

Fig. 1. The most popular loss functions (figure prepared in R [10]; similar figure may be found in [8], also prepared in R)

2 Loss Functions

There are different loss functions used in prediction problems (for a wide discussion see [8]). In this paper, we consider, besides 0-1 loss, the following three loss functions for binary classification:

– exponential loss:

$$L_{exp}(y, F(\mathbf{x})) = \exp(-yF(\mathbf{x})), \tag{7}$$

– binomial negative log-likelihood loss:

$$L_{log}(y, F(\mathbf{x})) = \log(1 + exp(-2yF(\mathbf{x}))), \tag{8}$$

– squared-error loss:

$$L_{sqr}(y, F(\mathbf{x})) = (y - F(\mathbf{x}))^2 = (1 - yF(\mathbf{x}))^2. \tag{9}$$

These loss functions are presented in Figure 2. Exponential loss is used in AdaBoost [6]. Binomial negative log-likelihood loss is common in statistical approaches. It is also used in Gradient Boosting Machines [3]. The reformulation of the squared-error loss (9) is possible, because $y \in \{-1, 1\}$. Squared-error loss is not a monotone decreasing function of increasing $yF(\mathbf{x})$. For values $yF(\mathbf{x}) > 1$ it increases quadratically. For this reason, one has to use this loss function in classification task very carefully.

The population minimizers for these loss functions are as follows:

$$F^*(\mathbf{x}) = \arg\min_{F(\mathbf{x})} E_{y|\mathbf{x}} L_{exp}(y, F(\mathbf{x})) = \frac{1}{2}\log\frac{P(y=1|\mathbf{x})}{P(y=-1|\mathbf{x})},$$
$$F^*(\mathbf{x}) = \arg\min_{F(\mathbf{x})} E_{y|\mathbf{x}} L_{log}(y, F(\mathbf{x})) = \frac{1}{2}\log\frac{P(y=1|\mathbf{x})}{P(y=-1|\mathbf{x})},$$
$$F^*(\mathbf{x}) = \arg\min_{F(\mathbf{x})} E_{y|\mathbf{x}} L_{sqr}(y, F(\mathbf{x})) = P(y=1|\mathbf{x}) - P(y=-1|\mathbf{x}).$$

From these formulas, it is easy to get values of $P(y=1|\mathbf{x})$.

3 Confirmation Measures

There are two confirmation measures of a particular interest:

$$l(H,E) = \log\frac{P(E|H)}{P(E|\neg H)}, \tag{10}$$

$$f(H,E) = \frac{P(E|H) - P(E|\neg H)}{P(E|H) + P(E|\neg H)}, \tag{11}$$

where H is hypothesis, and E is evidence. Measures l and f satisfy two desired properties that are:

- hypothesis symmetry: $c(H,E) = -c(\neg H, E)$ (for details, see for example [5]),
- and monotonicity property M defined in terms of rough set confirmation measures (for details, see [7]).

Let us remark that in the binary classification problem, one tries for a given $\mathbf{x}$ to predict value $y \in \{-1, 1\}$. In this case, evidence is $\mathbf{x}$, and hypotheses are then $y=-1$ and $y=1$. Confirmation measures (10) and (11) take the following form:

$$l(y=1|\mathbf{x}) = \log\frac{P(\mathbf{x}|y=1)}{P(\mathbf{x}|y=-1)}, \tag{12}$$

$$f(y=1|\mathbf{x}) = \frac{P(\mathbf{x}|y=1) - P(\mathbf{x}|y=-1)}{P(\mathbf{x}|y=1) + P(\mathbf{x}|y=-1)}. \tag{13}$$

4 Population Minimizers for Weighted Loss Functions

In this section, we present our main results that show relationship between loss functions and confirmation measures. We prove that population minimizers for weighted loss functions correspond to confirmation measures. Weighted loss functions are often used in the case of imbalanced class distribution, i.e., when probabilities $P(y=1)$ and $P(y=-1)$ are substantially different. Weighted loss function can be defined as follows:

$$L^w(y, F(\mathbf{x})) = w \cdot L(y, F(\mathbf{x})),$$

where $L(y, F(\mathbf{x}))$ is one of the loss functions presented above. Assuming that $P(y)$ is known, one can take $w = 1/P(y)$, and then:

$$L^w(y, F(\mathbf{x})) = \frac{1}{P(y)} \cdot L(y, F(\mathbf{x})). \tag{14}$$

In the proofs presented below, we use the following well-known facts: Bayes theorem: $P(y = 1|\mathbf{x}) = P(y = 1 \cap \mathbf{x})/P(\mathbf{x}) = P(\mathbf{x}|y = 1)P(y = 1)/P(\mathbf{x})$; and $P(y = 1) = 1 - P(y = -1)$ and $P(y = 1|\mathbf{x}) = 1 - P(y = -1|\mathbf{x})$.

Let us consider the following weighted 0-1 loss function:

$$L^w_{0-1}(y, F(\mathbf{x})) = \frac{1}{P(y)} \cdot \begin{cases} 0 & \text{if } yF(\mathbf{x}) > 0, \\ 1 & \text{if } yF(\mathbf{x}) \leq 0. \end{cases} \tag{15}$$

Theorem 1. *Population minimizer of* $E_{y|\mathbf{x}} L^w_{0-1}(y, F(\mathbf{x}))$ *is:*

$$\begin{aligned} F^*(\mathbf{x}) &= \begin{cases} 1 & \text{if } P(y = 1|\mathbf{x}) \geq P(y = 1), \\ -1 & \text{if } P(y = -1|\mathbf{x}) > P(y = -1) \end{cases} \\ &= \begin{cases} 1 & \text{if } c(y = 1, \mathbf{x}) \geq 0, \\ -1 & \text{if } c(y = -1, \mathbf{x}) > 0. \end{cases} \end{aligned} \tag{16}$$

where c is any confirmation measure.

Proof. We have that

$$F^*(\mathbf{x}) = \arg\min_{F(\mathbf{x})} E_{y|\mathbf{x}} L^w_{0-1}(y, F(\mathbf{x})).$$

Prediction risk is then:

$$E_{y|\mathbf{x}} L^w_{0-1}(y, F(\mathbf{x})) = P(y = 1|\mathbf{x}) L^w_{0-1}(1, F(\mathbf{x})) + P(y = -1|\mathbf{x}) L^w_{0-1}(-1, F(\mathbf{x})),$$

$$E_{y|\mathbf{x}} L^w_{0-1}(y, F(\mathbf{x})) = \frac{P(y = 1|\mathbf{x})}{P(y = 1)} L_{0-1}(1, F(\mathbf{x})) + \frac{P(y = -1|\mathbf{x})}{P(y = -1)} L_{0-1}(-1, F(\mathbf{x})).$$

This is minimized, if either $P(y = 1|\mathbf{x})/P(y = 1) \geq P(y = -1|\mathbf{x})/P(y = -1)$ for any $F(\mathbf{x}) > 0$, or $P(y = 1|\mathbf{x})/P(y = 1) < P(y = -1|\mathbf{x})/P(y = -1)$ for any $F(\mathbf{x}) < 0$ (in other words, only the sign of $F(\mathbf{x})$ is important). From $P(y = 1|\mathbf{x})/P(y = 1) \geq P(y = -1|\mathbf{x})/P(y = -1)$, we have that:

$$\frac{P(y = 1|\mathbf{x})}{P(y = 1)} \geq \frac{1 - P(y = 1|\mathbf{x})}{1 - P(y = 1)},$$

which finally gives $P(y = 1|\mathbf{x}) \geq P(y = 1)$ or $c(y = 1, \mathbf{x}) \geq 0$. Analogously, from $P(y = 1|\mathbf{x})/P(y = 1) < P(y = -1|\mathbf{x})/P(y = -1)$, we obtain that $P(y = -1|\mathbf{x}) > P(y = -1)$ or $c(y = -1, \mathbf{x}) > 0$. From the above we get the thesis. □

From the above theorem, it is easy to see that minimization of $L^w_{0-1}(y, F(\mathbf{x}))$ results in estimation of a region in predictor space in which class Cl_1 is observed with a positive confirmation. In the following theorems, we show that minimization of a weighted version of an exponential, a binomial negative log-likelihood, and a squared-loss error loss function gives an estimate of a particular confirmation measure, l or f.

Let us consider the following weighted exponential loss function:

$$L^w_{exp}(y, F(\mathbf{x})) = \frac{1}{P(y)} \exp(-y \cdot F(\mathbf{x})). \tag{17}$$

Theorem 2. *Population minimizer of $E_{y|\mathbf{x}} L^w_{exp}(y, F(\mathbf{x}))$ is:*

$$F^*(\mathbf{x}) = \frac{1}{2} \log \frac{P(\mathbf{x}|y=1)}{P(\mathbf{x}|y=-1)} = \frac{1}{2} l(y=1, \mathbf{x}). \tag{18}$$

Proof. We have that

$$F^*(\mathbf{x}) = \arg\min_{F(\mathbf{x})} E_{y|\mathbf{x}} L^w_{exp}(y, F(\mathbf{x})).$$

Prediction risk is then:

$$E_{y|\mathbf{x}} L^w_{exp}(y, F(\mathbf{x})) = P(y=1|\mathbf{x}) L^w_{exp}(1, F(\mathbf{x})) + P(y=-1|\mathbf{x}) L^w_{exp}(-1, F(\mathbf{x})),$$
$$E_{y|\mathbf{x}} L^w_{exp}(y, F(\mathbf{x})) = \frac{P(y=1|\mathbf{x})}{P(y=1)} \exp(-F(\mathbf{x})) + \frac{P(y=-1|\mathbf{x})}{P(y=-1)} \exp(F(\mathbf{x})).$$

Let us compute a derivative of the above expression:

$$\frac{\partial E_{y|\mathbf{x}} L^w_{exp}(y, F(\mathbf{x}))}{\partial F(\mathbf{x})} = -\frac{P(y=1|\mathbf{x})}{P(y=1)} \exp(-F(\mathbf{x})) + \frac{P(y=-1|\mathbf{x})}{P(y=-1)} \exp(F(\mathbf{x})).$$

Setting the derivative to zero, we get:

$$\exp(2F(\mathbf{x})) = \frac{P(y=1|\mathbf{x}) P(y=-1)}{P(y=1|\mathbf{x}) P(y=1)},$$
$$F(\mathbf{x}) = \frac{1}{2} \log \frac{P(y=1|\mathbf{x}) P(y=-1)}{P(y=1|\mathbf{x}) P(y=1)} = \frac{1}{2} l(y=1, \mathbf{x}). \quad \square$$

Let us consider the following weighted binomial negative log-likelihood loss function:

$$L^w_{log}(y, F(\mathbf{x})) = \frac{1}{P(y)} \log(1 + exp(-2y \cdot F(\mathbf{x}))). \tag{19}$$

Theorem 3. *Population minimizer of $E_{y|\mathbf{x}} L^w_{log}(y, F(\mathbf{x}))$ is:*

$$F^*(\mathbf{x}) = \frac{1}{2} \log \frac{P(\mathbf{x}|y=1)}{P(\mathbf{x}|y=-1)} = \frac{1}{2} l(y=1, \mathbf{x}). \tag{20}$$

Proof. We have that

$$F^*(\mathbf{x}) = \arg\min_{F(\mathbf{x})} E_{y|\mathbf{x}} L^w_{log}(y, F(\mathbf{x})).$$

Prediction risk is then:

$$\begin{aligned} &E_{y|\mathbf{x}} L^w_{log}(y, F(\mathbf{x})) = P(y=1|\mathbf{x}) L^w_{log}(1, F(\mathbf{x})) + P(y=-1|\mathbf{x}) L^w_{log}(-1, F(\mathbf{x})) \\ &= \frac{P(y=1|\mathbf{x})}{P(y=1)} \log(1+\exp(-2F(\mathbf{x}))) + \frac{P(y=-1|\mathbf{x})}{P(y=-1)} \log(1+\exp(2F(\mathbf{x}))). \end{aligned}$$

Let us compute a derivative of the above expression:

$$\begin{aligned} \frac{\partial E_{y|\mathbf{x}} L^w_{log}(y, F(\mathbf{x}))}{\partial F(\mathbf{x})} &= -2\frac{P(y=1|\mathbf{x})}{P(y=1)} \frac{\exp(-2F(\mathbf{x}))}{1+\exp(-2F(\mathbf{x}))} + \\ &+ 2\frac{P(y=-1|\mathbf{x})}{P(y=-1)} \frac{\exp(2F(\mathbf{x}))}{1+\exp(2F(\mathbf{x}))}. \end{aligned}$$

Setting the derivative to zero, we get:

$$\begin{aligned} \exp(2F(\mathbf{x})) &= \frac{P(y=1|\mathbf{x})P(y=-1)}{P(y=-1|\mathbf{x})P(y=1)} \\ F(\mathbf{x}) &= \frac{1}{2}\log\frac{P(y=1|\mathbf{x})P(y=-1)}{P(y=-1|\mathbf{x})P(y=1)} = \frac{1}{2} l(y=1, \mathbf{x}). \quad \square \end{aligned}$$

Let us consider the following weighted squared-error loss function:

$$L^w_{sqr}(y, F(\mathbf{x})) = \frac{1}{P(y)}(y - F(\mathbf{x}))^2. \tag{21}$$

Theorem 4. *Population minimizer of $E_{y|\mathbf{x}} L^w_{sqr}(y, F(\mathbf{x}))$ is:*

$$F^*(\mathbf{x}) = \frac{P(\mathbf{x}|y=1) - P(\mathbf{x}|y=-1)}{P(\mathbf{x}|y=1) + P(\mathbf{x}|y=-1)} = f(y=1, \mathbf{x}). \tag{22}$$

Proof. We have that

$$F^*(\mathbf{x}) = \arg\min_{F(\mathbf{x})} E_{y|\mathbf{x}} L^w_{sqr}(y, F(\mathbf{x})).$$

Prediction risk is then:

$$\begin{aligned} E_{y|\mathbf{x}} L^w_{sqr}(y, F(\mathbf{x})) &= P(y=1|\mathbf{x}) L^w_{sqr}(1, F(\mathbf{x})) + P(y=-1|\mathbf{x}) L^w_{sqr}(-1, F(\mathbf{x})), \\ E_{y|\mathbf{x}} L^w_{sqr}(y, F(\mathbf{x})) &= \frac{P(y=1|\mathbf{x})}{P(y=1)}(1 - F(\mathbf{x}))^2 + \frac{P(y=-1|\mathbf{x})}{P(y=-1)}(1 + F(\mathbf{x}))^2. \end{aligned}$$

Let us compute a derivative of the above expression:

$$\frac{\partial E_{y|\mathbf{x}} L^w_{log}(y, F(\mathbf{x}))}{\partial F(\mathbf{x})} = -2\frac{P(y=1|\mathbf{x})}{P(y=1)}(1 - F(\mathbf{x})) + 2\frac{P(y=-1|\mathbf{x})}{P(y=-1)}(1 + F(\mathbf{x})).$$

Setting the derivative to zero, we get:

$$F(\mathbf{x}) = \frac{P(y=1|\mathbf{x})/P(y=1) - P(y=-1|\mathbf{x})/P(y=-1)}{P(y=1|\mathbf{x})/P(y=1) + P(y=-1|\mathbf{x})/P(y=-1)},$$

$$F(\mathbf{x}) = \frac{P(\mathbf{x}|y=1) - P(\mathbf{x}|y=-1)}{P(\mathbf{x}|y=1) + P(\mathbf{x}|y=-1)} = f(y=1,\mathbf{x}). \quad \square$$

5 Conclusions

We have proven that population minimizers for weighted loss functions correspond directly to confirmation measures. This result can be applied in construction of machine learning methods, for example, ensemble classifiers producing a linear combination of base classifiers. In particular, considering ensemble of decision rules [1,2], a sum of outputs of rules that cover **x** can be interpreted as an estimate of a confirmation measure for **x** and a predicted class.

Our future research will concern investigation of general conditions that loss function has to satisfy to be used in estimation of confirmation measures.

References

1. Błaszczyński, J., Dembczyński, K., Kotłowski, W., Słowiński, R., Szeląg, M.: Ensemble of Decision Rules. Research Report RA-011/06, Poznań University of Technology (2006)
2. Błaszczyński, J., Dembczyński, K., Kotłowski, W., Słowiński, R., Szeląg, M.: Ensembles of Decision Rules for Solving Binary Classification Problems with Presence of Missing Values. In: Greco et al. (eds.): Rough Sets and Current Trends in Computing, Lecture Notes in Artificial Intelligence, Springer-Verlag **4259** (2006) 318–327
3. Friedman, J. H.: Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics, 5 **29** (2001) 1189–1232
4. Friedman, J. H.: Recent Advances in Predictive (Machine) Learning. Dept. of Statistics, Stanford University, `http://www-stat.stanford.edu/~jhf` (2003)
5. Fitelson, B.: Studies in Bayesian Confirmation Theory. Ph.D. Thesis, University of Wisconsin, Madison (2001)
6. Freund, Y., Schapire, R. E.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. System Sci. **55** 1997 119–139
7. Greco, S., Pawlak, Z., Słowiński, R.: Can Bayesian confirmation measures be useful for rough set decision rules? Engineering Applications of Artificial Intelligence **17** (2004) 345–361
8. Hastie, T., Tibshirani, R., Friedman, J. H.: Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer (2003)
9. Kyburg, H.: Recent work in inductive logic. In: Lucey, K.G., Machan, T.R. (eds.): Recent Work in Philosophy. Rowman and Allanheld, Totowa, NJ, (1983) 89–150
10. R Development Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, `http://www.R-project.org`, Vienna, (2005)