2008 International Symposium on Nonlinear Theory and its Applications
NOLTA'08, Budapest, Hungary, September 7-10, 2008

Training and Scoring of Probabilistic Classifiers

Jochen Bröcker†

†Max–Planck–Institut für Physik komplexer Systeme
Nöthnitzer Strasse 34
01187 Dresden, Germany
Email: broecker@pks.mpg.de

Abstract—This contribution discusses aspects of a learning theory for probabilistic classifiers. Classical statistical learning theory focusses mainly on classifiers which give unequivocal response to an input. This is, the output of the classifier is always one of the class labels (after appropriate conversion). Such classifiers are often inadequate though, for example if the classifier is to be used by a community of users with heterogeneous cost–loss profiles. Recently there has been increasing interest in classifiers which provide a probabilistic rather than a deterministic answer, since probability assignments allow for more informed decision making in the face of uncertain risks. The present contribution discusses how to evaluate or "score" probability assignments, leading to the concept of scoring rules. As will be demonstrated, scoring rules need to have certain properties in order to guarantee this evaluation to be logically consistent. Furthermore, scoring rules allow to formulate the training of a probabilistic classifier as empirical risk minimisation, rendering large parts of the theory of statistical learning applicable to the present problem.

1. Probabilistic Forecasts

Assume the objective is to forecast whether a real–world event will occur or not, for example whether on a given day the temperature at Dresden airport at 12 o'clock will fall below 0°C. We define the variable Y, referred to as the *target*[1], to be 1 if that event actually happens and 0 if it does not. As forecasters, we may or may not have some information available which we can employ to build our forecasts. As a *probabilistic forecast* for Y we denote any function ρ which maps our information or *input* data onto a number between zero and one, requiring no further properties so far. To stay with the example above, we might have access to temperature forecasts for Dresden, produced by a numerical weather prediction system. We could take that temperature forecast as the aforementioned input data, while a possible choice for the function ρ could be some sigmoidal function which maps possible temperature values onto the interval $[0, 1]$. So far, this is of course but a classifier with output signal confined between zero and one, and nothing else justifies this simple scheme being called a probabilistic forecast. But what we of course have in mind is this: Suppose there is some underlying probabilistic relationship between the inputs and the targets, and the aim is to find a ρ mimicking that relationship. Suppose we have available a representative set of feature–target pairs, can we learn the underlying relationship from these *training data*? The task is similar to for example the regression problem, where the aim is to mimic a deterministic relationship. The theory of statistical learning is a framework which provides both theoretical understanding and algorithms to learn from data. The basic approach of statistical learning is to start with a sufficiently flexible class of candidate functions and then pick the candidate which exhibits the least error in explaining the training data (according to some appropriate measure of error). Statistical learning for probabilistic relationships is a less well developed theory, and the following sections aim at changing this to the better. In Section 2, the problem at hand will be stated precisely. Furthermore, it is argued informally that the objective of probabilistic forecasting should be to reconstruct (an approximation to) the probability of the event (or possible values of the target) given the inputs. Section 3 revisits scoring schemes, which will provide the appropriate measures of error of a probabilistic forecast. In Section 4 three complementary model classes will be discussed, which allow for constructing probabilistic forecasts. Section 5 concludes. In brief, using appropriate model classes and performance measures, most of the concepts from classical statistical learning carry over to the situation discussed in this paper.

[1] Elsewhere referred to as the *verification* or *observation*

2. Problem Statement and Notation

The primary aim of this section is to settle some notational conventions. The general setup we have in mind is as follows. As in the introduction, let the target Y be a random variable taking the values 0 and 1 only, with $Y = 1$ indicating that the event under concern happened and $Y = 0$ otherwise. The features or inputs X are random variables too, taking values in $\mathbb{R}^d$. The underlying probability measure will be denoted by $\mathbb{P}$. The probabilistic relationship between X and Y is described through the following objects. Let

$$\begin{aligned} f_0(x) &:= \mathbb{P}(X \in x + \mathrm{d}x | Y = 0), \\ f_1(x) &:= \mathbb{P}(X \in x + \mathrm{d}x | Y = 1), \end{aligned} \tag{1}$$

that is, f_0 and f_1, respectively, are the densities of X given $Y = 0$ and $Y = 1$, respectively. By

$$\pi(x) := \mathbb{P}(Y = 1|X = x) \tag{2}$$

we denote the conditional probability of the event "$Y = 1$" given X, and

$$\bar{\pi} := \mathbb{P}(Y = 1) \tag{3}$$

denotes the *base rate* or *grand probability* of the event "$Y = 1$". Finally,

$$f(x) := \mathbb{P}(X \in x + \mathrm{d}x) \tag{4}$$

denotes the unconditional density of the features X. The Bayes rule entails various relations between these objects, for example $f(x) = f_1(x)\bar{\pi} + f_0(x)(1 - \bar{\pi})$.

Intuitively, one would hope that $\rho(x)$ gives the probability of $Y = 1$ given $X = x$, or

$$\rho(x) = \mathbb{P}(Y = 1|X = x) = \pi(x). \tag{5}$$

We will see in Sect. 3 that many reasonable measures of forecast success support this intuition, that is, they give maximum possible scores if $\rho(X)$ indeed agrees with the probability of $Y = 1$ given X.

A seemingly different way to motivate $\pi(x)$ as a good forecast probability is through reliability. Reliability means that on condition that the forecast (approximately) equals z, the event should occur with a relative frequency (approximately) equal to z, too. As an optimality criterion, reliability is not sufficient to single out a particular forecasting scheme, since *any* conditional probability of the form $\mathbb{P}(Y = 1|I)$ is reliable, independent of what I is. In particular, the base rate $\bar{\pi}$ is reliable as well. Hence, in addition to reliability, the forecast should feature a high correlation with the actual event. This property is known as sharpness. It can be demonstrated that $\pi(x)$ is indeed the forecast which features maximum sharpness among all reliable forecasts which can be written as functions of X. As we will briefly discuss in Sect. 3, it is in fact for the same reason that $\rho(X) = \mathbb{P}(Y = 1|X)$ achieves optimal scores.

3. Scoring Schemes

In this section, the question of how to quantify the performance of forecasts is addressed. Performance measures are important not only in order to rank existing forecast schemes but also in the design of such schemes, for example the tuning of free parameters. Measuring the success of predictions in terms of how "close" they eventually come to the truth is a paradigm which presumably requires no further motivation. The (root) mean squared error is one among many possible variants of this paradigm. If we envisage to formulate our forecasts in terms of probabilities though, the paradigm cannot be applied readily without modification, as the notion of "distance" between forecast and targets ceases to be meaningful if the forecast is a probability assignment but the target is a class label. But probability forecasts essentially quantify how likely a given potential event will come true, thus already providing a sort of self rating. Hence it seems reasonable to value the success of a probability forecast in terms of how confident the forecast was of the event which eventually occurred, in relation to other events which did not. This idea is implemented in the concept of *scores*, explained in Subsection 3.1.

Another popular approach to measuring the quality of probabilistic forecast is the Receiver Operating Characteristic (ROC), briefly discussed in Subsection 3.2. Different from scores, the ROC, albeit taking the probabilistic character of the forecast into account, is insensitive to the reliability of the forecast.

3.1. Brier Score and Ignorance

A *scoring rule* [6, 9, 4, 11] is a function $S(p, z)$ where $p \in [0, 1]$ and z is either zero or one. If $\rho(X)$ is a forecast and Y is the corresponding target, then $S(\rho(X), Y)$ quantifies how well $\rho(X)$ succeeded in forecasting Y. A scoring rule effectively defines two functions $S(p, 1)$, quantifying the score in case the forecast is p and the event happens, and $S(p, 0)$, quantifying the score in case the forecast is again p but the event does not happen. Two important examples are the Ignorance score [6], given by the scoring rule

$$S(p, y) := -\log(p) \cdot y - \log(1 - p) \cdot (1 - y), \tag{6}$$

and the Brier score [1], given by the scoring rule

$$S(p, y) := (y - p)^2 = (1 - p)^2 \cdot y + p^2 \cdot (1 - y). \tag{7}$$

These definitions imply the convention that a smaller score indicates a better forecast.[2]

A score is a "point–wise" (evaluated at every single time instance) measure of performance. It quantifies the success of individual forecast instances by comparing the random variables $\rho(X)$ and Y point-wise. The general quality of a forecasting system (as given here by the random variable $\rho(X)$) is commonly measured by the average score $\mathbb{E}[S(\rho(X), Y)]$, which can be estimated by the empirical mean

$$\mathbb{E}[S(\rho(X), Y)] \cong \frac{1}{N} \sum_{i=1}^{N} S(\rho(x_i), y_i) \tag{8}$$

over a sufficiently large data set (x_i, y_i).

The rationale behind the two mentioned scoring rules, the Ignorance and the Brier score, is rather obvious. If the event occurs, the score should become better (i. e. decrease) with increasing ρ, while if it does not occur, the score should become worse (i.e. *in*crease) with increasing ρ. But why then not taking just $1 - \rho$ if the event occurs,

[2] This convention might run contrary to how the word "score" is used in ordinary parlance. Its virtue though lies in the fact that the divergence function (to be defined later) becomes positive definite.

and ρ if it does not? To see the problem with this "linear" scoring rule, define the *scoring function*

$$s(q, p) := S(q, 1) \cdot p + S(q, 0) \cdot (1 - p) \tag{9}$$

where q, p are two arbitrary probabilities, that is, numbers in the unit interval. Note that the scoring function is the score averaged over cases where the forecast is q but in fact p is the true probability of the event "$Y = 1$". In view of the interpretation of the scoring function, it seems reasonable to require that the average score of the forecast q should be best (i.e. minimal) if and only if q in fact coincides with the true probability of the event "$Y = 1$". This means that the *divergence function* (or loss function)

$$d(q, p) := s(q, p) - s(p, p) \tag{10}$$

has to be positive definite, that is, never negative and zero only if $p = q$. A scoring rule with the corresponding divergence function having this property is called *strictly proper* [4, 3]. The divergence function of the Brier score for example is $d(q, p) := (q-p)^2$, demonstrating that this score is strictly proper. While the Ignorance is proper as well, the linear score though is easily shown to be *improper*.

The mathematical expectation of a strictly proper scores allows for a very interesting decomposition (see [2] for a proof). For any strictly proper scoring rule, define the *entropy* $e(p) := s(p, p)$. Furthermore let $\pi_\rho(r) := \mathbb{P}(Y = 1|\rho(X) = r)$ be the conditional probability of $Y = 1$ given that $\rho(X) = r$. This probability is a function of ρ but is fully calibrated. Then

$$\mathbb{E}S(\rho, Y) = e(\bar{\pi}) - \mathbb{E}d(\bar{\pi}, \pi) + \mathbb{E}d(\rho, \pi_\rho). \tag{11}$$

These terms can be interpreted as follows: The entropy $e(\bar{\pi})$ is the ability of the base rate $\bar{\pi}$ to forecast draws from itself, and hence quantifies the fundamental uncertainty inherent in Y. The term $\mathbb{E}d(\bar{\pi}, \pi)$ is positive definite and quantifies the average divergence of π from its mean. It can hence be considered a generalised variance of π. If the Brier score is used, this term is in fact the ordinary variance of π. The term $\mathbb{E}d(\rho, \pi_\rho)$ is again positive definite and quantifies the imperfect calibration of ρ. As noted earlier, a larger variance of π yields better forecast skill.

3.2. The Receiver Operating Characteristic

The ROC [5] is a concept originating in signal detection, but it is applicable to any binary classification problem. The ROC curve for a certain classifier ρ comprises a plot of the *hit rate*

$$H(\delta) := \mathbb{P}(\rho \geq \delta|Y = 1) \tag{12}$$

versus the *false–alarm rate*

$$F(\delta) := \mathbb{P}(\rho \geq \delta|Y = 0), \tag{13}$$

with δ acting as a parameter along the curve.It follows readily from the definitions that both H and F are monotonously decreasing functions of δ with limits 0 for increasing δ and 1 for decreasing δ, whence the ROC curve is a monotonously *in*creasing arc connecting the points $(0, 0)$ and $(1, 1)$. Monotonous transformations of the classifier do not change the ROC, as is easily derived from its definition.

Arguably, a classifier ρ_1 should be taken as better than another classifier ρ_2, if for any fixed false–alarm rate F, the hit rate H_1 of ρ_1 is equal or larger than the hit rate H_2 of ρ_2. If this is the case, we will refer to ρ_1 as being *never inferior* to ρ_2. It can be demonstrated that the classifier $\pi(X)$ is never inferior to any classifier of the form $\rho(X)$ (this follows from the Neyman–Pearson–Lemma [10] and the fact that $\pi(X)$ is a monotonically increasing function of the likelihood ratio (see Eq. 14).

If we have to compare two arbitrary classifiers ρ_1 and ρ_2, then the notion of "never inferior" is not so useful, as the two ROC curves might cross. This is a problem if a criterion is required in order to rank classifiers, as there is no reason why the ROC curves corresponding to any two classifiers should not cross. Hence, summary statistics of ROC curves are needed, for example the area under the ROC curve or *AUC* (which is positively oriented). It can be shown that the AUC gives the probability that on an instance when the event takes place, the classifier is actually larger than on an independent instance when the event does not take place. A never inferior classifier is obviously never inferior with respect to AUC.

4. Three Model Classes

In this section, three model classes will be presented. The first two, Kernel Estimators and Nearest Neighbor Models, are variants of their cousins well known in statistical learning. The third, logistic regression, has been thoroughly investigated in statistics [7].

4.1. Kernel Estimator

At the basis of this model class is the identity

$$\pi(x) = \frac{\bar{\pi} f_1(x)}{f(x)} = \frac{1}{1 + \frac{(1-\bar{\pi}) f_0(x)}{\bar{\pi} f_1(x)}}, \tag{14}$$

which suggests to first estimate the probability densities $f_0(x)$ and $f_1(x)$ and then replace the respective quantities in Equation (14) by their estimates. So–called kernel estimators [12] provide a simple yet powerful method to estimate probability densities from data. Versatile implementations of this technique are available [8]. Since kernel estimators do essentially all computations upon evaluation, it is worthwhile to use efficient code in these steps. On the other hand, kernel estimators allow for simple leave–one–out crossvalidation, a feature which is very conventient for model validation.

4.2. Nearest Neighbor Approach

Nearest neighbor (NN) models [7] are among the most popular model types in statistical learing. Common to all NN models is that for a given query point x, the training set is searched for a few nearest neighbors $x_i, i = 1 \dots n$ of x. The number of requested neighbors n might be fixed, or alternatively all neighbors within a certain radius δ of the query point are considered, in which case n depends on x. Then a (usually rather simple) local model is fitted to the selected feature–target pairs $(x_i, y_i), i = 1 \dots n$ and finally evaluated at the query point x. Most common are local averages, that is, we estimate

$$\rho(x) = \frac{\sum y_i w_i}{\sum w_i}, \tag{15}$$

where the sum runs over all i so that x_i is among the chosen neighbors of x, and the w_i are weights which depend on the distance between x and x_i. Giving fewer weight to points which are further away from the query point renders the estimate more smooth. Otherwise, if the query point is varied, the estimator jumps a little every time a new point enters the neighborhood.

4.3. Logistic Regression

Logistic regression [7] assumes a model of the form

$$\rho(x) = \lambda\left(\beta_0 + x\beta^t\right) \tag{16}$$

where β_0 resp. β are parameters to be determined. The *link function* λ can be any monotonous function mapping $\mathbb{R}$ to the unit interval, but a popular choice is

$$\lambda(z) = \frac{\exp(z)}{1 + \exp(z)}. \tag{17}$$

In other words, logistic regression assumes that the *logarithmic odds ratio* $\log(\frac{\rho}{1-\rho})$ is a linear function of the inputs. The parameter vector is often determined using the log–likelihood (which is equivalent to using the Ignorance) as emirical risk, but other scores work just as well. Logistic models inherit various useful properties from linear models, as long as strictly proper scores are used in the empirical risk minimisation. The reason is that locally around the optimum, risk minimisation is equivalent to weighted linear regression. For example, logistic regression allows for an easy calculation of the leave–one–out parameters. These are only approximately true, but the error should be small as long as there are sufficiently many data points so that leaving out one of them amounts to a small perturbation of the risk functional only. To minimize the empirical risk, a Newton–Raphson algorithm can be used. Good initial guesses for the parameter β are obtained by fitting a conventional linear model $z = \beta_0 + x\beta^t$ to the modified targets $\tilde{y}_i = 2y_i - 1$. As any learning problem, logistic models need regularization if there are many and highly correlated inputs, in order to avoid large model variance. We found that a robust way to achieve this is to determine the parameters β^t by standard ridge–regression [7] on the modified targets $\tilde{y}_i$. The ridge penalty can be optimized using, for example, leave–one–out cross–validation. The so determined parameters are used in a logistic model of the form

$$\rho(x) = \lambda(\beta_0 + \beta_s \cdot (x\beta^t))$$

with only two remaining parameters β_0, β_s to be determined. Finally, note that the link function (and the parameters β_0, β_s) have no influence on the ROC. Hence, in terms of ROC, a logistic model is equivalent to its "linear core".

5. Conclusion

This brief contribution discusses aspects of a statistical learning approach to probabilistic forecasting. Both scoring techniques as well as model classes are presented, which render the risk minimisation principle, which is fundamental to statistical learning, applicable in the present situation.

Acknowledgements

Fruitful discussions with the members of Centre for the Analysis of Time Series (CATS), London School of Economics, in particular Liam Clarke, Milena Cuellar and Leonard A. Smith, are kindfully acknowledged. I am indebted to Christian Merkwirth (UJ Kraków) and Jörg Wichard (FMP–Berlin) for revealing to me various secrets of statistical learning.

References

[1] Glenn W. Brier. Verification of forecasts expressed in terms of probabilities. *Monthly Weather Review*, 78(1), 1950.

[2] Jochen Bröcker. Decomposition of proper scores. (submitted to *Journal of Forecasting*, available as Tech.Rep. from http://edoc.mpg.de, document ID: 324199.0).

[3] Jochen Bröcker and Leonard A. Smith. Scoring probabilistic forecasts: The importance of being proper. *Weather and Forecasting*, 22(2):382–388, 2007.

[4] Thomas A. Brown. Probabilistic forecasts and reproducing scoring systems. Technical Report RM–6299–ARPA, RAND Corporation, June 1970.

[5] James P. Egan. *Signal detection theory and ROC analysis*. Academic Press series in cognition and perception. Academic Press, first edition, 1975.

[6] I. J. Good. Rational decisions. *Journal of the Royal Statistical Society*, XIV(1):107–114, 1952.

[7] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, first edition, 2001.

[8] Alexander Ihler. Kernel density estimation toolbox for matlab.

[9] J. L. Kelly, Jr. A new interpretation of information rate. *Bell System Technical Journal*, 35, 1956.

[10] Alexander M. Mood, Franklin A. Graybill, and Duane C. Boes. *Introduction to the Theory of Statistics*. McGraw-Hill Series in Probability and Statistics. McGraw-Hill, 1974.

[11] Leonard J. Savage. Elicitation of personal probabilities and expectation. *Journal of the American Statistical Association*, 66(336), 1971.

[12] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, first edition, 1986.