Practical Considerations for Performance Measures on Classifiers on Real World Streaming Data

Abstract

Several real world binary classification problems in the domains of automobile, energy and bank are outlined. This paper do not discuss how to solve the problems, but rather how any solution can be evaluated. All problems involve huge data sets which have various limitations related to the question of what is the ground truth. The problems also involve streaming data and earliness of warnings need to be balanced with accuracy. The paper includes practical considerations tailored to the specific properties of each particular problem.

1 Introduction

In the binary classification methods described in this note, we consider the case when the classification rule is basically comparing an output function $q(x)$ to a fixed threshold T, where x is the object that is classified. In a Bayesian network $q(x)$ is usually a probability between 0 and 1.

In general we have what we call a training set, which consist of sample of elements in class zero X_0 and a sample of class one X_1. The classification rule is

$$h_T(x) = \begin{cases} 0 & \text{for} \quad q(x) \leq T \\ 1 & \text{else.} \end{cases} \tag{1}$$

We can count all the false positives by

$$\text{FP(T)} = \sum_{i=1}^{n_0} h_T(x_{0,i}) \tag{2}$$

and the false negatives are given by

$$\text{FN(T)} = \sum_{i=1}^{n_1} 1 - h_T(x_{1,i}). \tag{3}$$

It is clearly seen that the number of false positives and negatives is dependent on which threshold is used. In order to further discuss which threshold to use it is necessary to include information of the prior probabilities, which is the how often does an instance

of class zero happen compared to how often does an instance of class one happen. If the instances are drawn by random then $n_0/(n_0 + n_1)$ and $n_1/(n_0 + n_1)$ can be reasonable priors. Additionally, it also makes sense to include a discussion of the costs of misclassification. For instance it may be more costly to misclassify a cancer patient as not having cancer compared to the opposite. Including a discussion of costs and priors makes the testing regime more complicated, but it is a way to put a number on what is the cost of using this classifier compared to a perfect classifier that never misses. In this way it is possible to compare two classification methods in terms of how expensive are they compared to the perfect classifier. We will say more about this in subsection 1.2.

However, there are also ways to discuss a classifier without taking into account either cost functions or priors, which comes next.

1.1 Concordance probability

False positive rates are $FPR(T) = FP(T)/n_0$ and false negative rates are $FNR(T) = FN(T)/n_1$. When these are plotted against each other, when T is going through all relevant values, we get what is called the receiver operation curve which allow one to visually inspect the effect of varying T. AUROC is a measure of the integral under the ROC curve, which is independent of T and also the priors. It may be difficult to comprehend what AUROC is at this point, but we will return to it afterward.

First, let us define two random variables $Q_0 = q(X_0)$ and $Q_1 = q(X_1)$. Q_0 is basically a random variable that is related to when you pick a random sample of class zero and run the system and then you get an output value (usually a probability of something happening). The same thing can be said for Q_1. The probability $P(Q_1 > Q_0)$ is of interest, because this says what is the probability that if you take one sample from each of the populations, what is the chance that the sample from population one is the highest. This question is independent of T, meaning that discussion about priors and costs are not needed. This probability is called the concordance probability.

Without giving you the details, it is possible to prove that

$$\text{AUROC} = P(Q_1 > Q_0). \tag{4}$$

It is also worth mentioning that nothing need to be assumed about the probability distributions of Q_0 and Q_1 (this is related to the Mann-Whitney U test).

The concordance probability or the AUROC measure is therefore a measure that can be discussed without relating discussing the priors or costs of misclassification. The costs are discussed in the next subsection.

1.2 Empirical risk

In mathematical optimization, statistics, decision theory and machine learning, a loss function or cost function is a function that maps an event or values of one or more variables onto a real number that is intuitively representing some *cost* associated with the event. Loss functions can be used on optimization problems, where an algorithm

or method is optimized by minimizing the loss function. Moreover, loss functions are frequently used to diagnose and compare various algorithms or methods.

In this paper define the *loss function* as a real and lower-bounded function $L(x_i, h(x_i), y_i)$. It takes into account both the x value in addition to $h(x)$ which is the classification rule and the true value y.The empirical risk is found by averaging the loss function on the training set given by

$$R_{emp}(h, \boldsymbol{x}) = n^{-1} \sum_{i=1}^{n} L(x_i, h(x_i), y_i). \tag{5}$$

Notice that L is an array of $n \times 2 \times 2$ elements. Many supervised learning algorithms are optimized by finding the h in a hypothesis space $\mathcal{H}$ that minimizes the empirical risk. This note will not focus on empirical risk minimization, but rather focus on using empirical risk to compare methods.

2 Practical problems

In this section we will outline a number of practical using concordance probability and the empirical risk function.

2.1 Automotive use cases

There are two application scenarios here. The first one is early recognition of lane change manoeuvre. The second is prediction of the need for lane change based on relative dynamics between two vehicles following the same lane.

For the first use case scenario it is required that the concordance probability (AUROC) should be above 0.96 for prediction 1 second before lane crossing and 0.90 for the 2 second prediction.

For the second use case scenario it is required that the concordance probability (AUROC) should be above 0.96 for prediction 1 second before lane crossing and 0.90 for the 2 second prediction.

These two application scenarios seem straightforward to calculate.

Questions:

1. What will be the sizes of n_0 and n_1.
2. Are each test sample completely independent?
3. Are the shapes of $P(Q_0)$ and $P(Q_1)$ equal? If so, we can also do the hypothesis test.
4. Are you sure that a cost invariant test is sufficient for your use?

2.2 Financial use case: Default prediction of clients

There are two application scenarios here. The first one is prediction of whether a client will default within two years and the second is related to the benefit of a marketing campaign.

Discussion using concordance probability

In the first use case scenario, it is required that the concordance probability (AUROC) should be above 0.90.

This use case involves to predict the probability p_i of defaulting for certain customer i that is applying for a loan in a bank. The y's can take value 0, which is non defaulting and value 1, which is defaulting. In the cases where a loan is given, each y_i is determined by whether the loan has defaulted or not, exactly two years later. This means that the both $\boldsymbol{q_0}$ and $\boldsymbol{q_1}$ are severely biased. Estimating the concordance probability will only be relevant if we use the AMIDST software as an addition to the existing software.

Questions for the first use case scenario:

1. Is it ok to test on only these samples that are known? Should we try to make a different sample (by expert estimates or similar) that is independent on whether a loan was given to them or not.
2. What will be the sizes of n_0 and n_1.
3. Are each test sample completely independent?
4. Are the shapes of $P(Q_0)$ and $P(Q_1)$ equal? If so, we can also do the hypothesis test.

Discussion using empirical risk

In the second use case scenario it is required that the benefit of a AMIDST induced marketing campaign should be more than 5 percent higher than a normal campaign.

Based on the size of the loan it is possible to reason about the cost of defaulting $c_i(0|1)$ and also the cost of declining the loan application if the customer actually would have not defaulted $c_i(1|0)$. The classification problem reduces to whether $c_i(0|1)p_i$ is higher than $c_i(1|0)(1-p_i)$, which is the same as comparing p_i with $c_i(1|0)/(c_i(0|1+c_i(1|0))$.

In order to discuss the cost of using this classification method, compared to a perfect classifier, we propose this implementation of the loss function

$$L(x_i, h(x_i), y_i) = \begin{cases} 0 & \text{for} \quad h(x_i) = 0 \quad \& \quad y_i = 0 \\ c_i(1|0) & \text{for} \quad h(x_i) = 1 \quad \& \quad y_i = 0 \\ c_i(0|1) & \text{for} \quad h(x_i) = 0 \quad \& \quad y_i = 1 \\ 0 & \text{for} \quad h(x_i) = 1 \quad \& \quad y_i = 1. \end{cases} \tag{6}$$

In this context, the loss function is only partially known. It is only known at the x_is where the subject was part of the default marketing campaign and a loan was actually given. We define a function $h_{pre}(x) : \Omega_X \to \Omega_Y$ as the decision rule which involves that the bank decided to offer a loan and also that the subject decided to accept the loan more than two years ago. Consequently, y_i is only known given $h_{pre}(x_i) = 1$. We define $\boldsymbol{x_{acc}} = \{x_i \in \boldsymbol{x} | h_{pre}(x_i) = 1\}$, $\boldsymbol{y_{acc}} = \{y_i \in \boldsymbol{y} | h_{pre}(x_i) = 1\}$ and the sizes of $\boldsymbol{x_{acc}}$ and $\boldsymbol{y_{acc}}$ are equal to n_{acc}.

If we let $y_{acc,i}$ be an element of $\boldsymbol{y_{acc}}$ and $x_{acc,i} \in \boldsymbol{x_{acc}}$ be a corresponding element to $y_{acc.i}$, then an estimate of empirical risk is

$$R_{emp}(h, \boldsymbol{x_{acc}}) = n_{acc}^{-1} \sum_{i=1}^{n_{acc}} L(x_{acc,i}, h(x_{acc,i}), y_{acc,i}). \tag{7}$$

This approximation must be treated with care because the $x_{acc,i}$s are not taken randomly, but they are filtered by when the decision rule $h_{pre}(x_i)$ is equal to one. However, $R_{emp}(h, X_{acc})$ has a practical interpretation. It is the extra cost of using the decision rule h instead of a perfect classifier on data that are already filtered by $h_{pre}(x_i)$. Moreover, this number can be compared to $R_{emp}(h_{pre}, \boldsymbol{x_{acc}})$, which is the cost associated with using h_{pre} on $\boldsymbol{x_{acc}}$. It is therefore possible to outline if there is a financial gain of using a two stage filter, that is using h_{pre} prior to h, compared to only using h_{pre}.

The two stage scenario is of cause an interesting scenario by itself, but it is probably more interesting to see whether it makes sense to use h instead of h_{pre}.

Questions for the second use case scenario:

1. How can we possibly find out which is least costly of h and h_{pre}? This is the key problem in my opinion.
2. What will be the sizes of n_0 and n_1.
3. Are each test sample completely independent?