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. It is argued that each of the problems presented in this paper can be discussed using two measures; concordance probability related to the output function and the empirical risk function in some forms. The paper includes practical considerations tailored to the specific properties of each particular problem.

1 Introduction

We consider the random variables X and Y with joint probability distribution $P(X, Y)$ and output spaces Ω_X and $\Omega_Y = \{0, 1\}$, respectively. We consider a continuous output function $q : \Omega_X \to \mathbb{R}^+$ and a family of hypothesis functions $\mathcal{H}$, where each element $h_T : \Omega_X \to \Omega_Y$ has the form

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

Also, let X_0 and X_1 be random variables with probability distributions $P(X|Y = 0)$ and $P(X|Y = 1)$, respectively. We define the random variables $Q_0 = q(X_0)$ and $Q_1 = q(X_1)$.

In this paper, we will discuss the continuous output function in the light of the Mann-Whitney U test and the concordance probability $P(Q_1 > Q_0)$. It is important to mention that in the context of concordance probability, the shapes of $P(Q_0)$ and $P(Q_1)$ may be different.

Also, we want to mention that the concordance probability is exactly equal to the area under the receiver operating characteristic curve (ROC) and the common language effect size of the Mann-Whitney U test.

In terms of discussing the family of hypothesis functions, we have chosen empirical risk as the quantity of interest. This involves defining a loss function and the risk function is simply the expected loss in a frequentist perspective.

1.1 Mann-Whitney U test

In statistics, the Mann-Whitney U test (also called the Mann-Whitney-Wilcoxon (MWW), Wilcoxon rank-sum test, or Wilcoxon-Mann-Whitney test) is a nonparametric test of the null hypothesis that two populations are the same against an alternative hypothesis, especially that a particular population tends to have larger values than the other. It has greater efficiency than the t-test on non-normal distributions and it is nearly as efficient as the t-test on normal distributions.

We define a training set $\mathbf{x} \times \boldsymbol{y}$ with n input-output pairs (x_i, y_i), independently drawn from $P(X, Y)$. From the training set we have two populations $\boldsymbol{q_0} = \{q(x_i), | \, y_i = 0\}$ and $\boldsymbol{q_1} = \{q(x_i), | \, y_i = 1\}$. Their sizes are n_0 and n_1 so that $n_0 + n_1 = n$. Calculating the U statistics is straightforward, where these two values are obtained

$$U_0 = \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} H(\, q_j - q_i \,) \quad \text{and} \quad U_1 = \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} H(\, q_i - q_j \,). \tag{2}$$

Here $H(\cdot)$ is the heaviside step function and notice that $U_0 + U_1 = n_0 n_1$. For large samples, each U is approximately normally distributed. In that case, the standardized value

$$z = \frac{U_0 - m_U}{\sigma_U}, \tag{3}$$

where m_U and σ_U are the mean and standard deviation of U given by

$$m_U = \frac{n_0 n_1}{2} \quad \text{and} \quad \sigma_U = \sqrt{\frac{n_0 n_1 (n_0 + n_1 + 1)}{12}}. \tag{4}$$

Significance of test can be checked in tables of the normal distribution. Although, such an hypothesis test is interesting by itself, we are more interested in the concordance probability $P(Q_1 > Q_0)$ which is defined by

$$P(Q_1 > Q_0) = \frac{U_1}{n_0 n_1}. \tag{5}$$

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 on $\Omega_X \times \Omega_Y \times \Omega_Y$. The value of the loss function at an arbitrary point $(x, h(x), y)$ is interpreted as the loss, or cost, of taking the decision $h(x)$ at x, when the right decision is y. Notice that in this paper, the loss function is dependent on x as well. This is of high practical use, because a certain misclassification might be more expensive that another.

In the frequentist perspective, the expected loss is often referred to as the risk function. It is obtained by taking the expected value over the loss function with respect to the probability distribution $P(X, Y) : \Omega_X \times \Omega_Y \to \mathbb{R}^+$. The *risk function* is given by

$$R(h) = \int_{\Omega_X, \Omega_Y} L(x, h(x), y) dP(x, y). \qquad (6)$$

In the case when the costs are independent of x and also that there is no cost related to correct classification, the risk function reduces to the well known expected cost of misclassification (ECM)

$$ECM = c(1|0)p(1|0)p_0 + c(0|1)p(0|1)p_1. \qquad (7)$$

Here, $c(1|0)$ is the cost for misclassifying an item of class zero as class one and $p(1|0)$ is the misclassification probability given class zero. The quantities $c(0|1)$ and $p(0|1)$ are defined equivalently, while p_0 and p_1 are the priors.

In general, the risk $R(h)$ cannot be computed because the distribution $P(x, y)$ is unknown. However, we can compute an approximation, called empirical risk, by averaging the loss function on the training set given by

$$R_{emp}(h, \boldsymbol{x}) = n^{-1} \sum_{i=1}^{n} L(x_i, h(x_i), y_i). \qquad (8)$$

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 paper 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

$$L(x_i, h(x_i), y_i) = \begin{cases} 0 & \text{for} \quad h(x_i) = 0 \quad \& \quad y_i = 0 \\ c_i(1|0) & \text{for} \quad h(x_i) = 1 \quad \& \quad y_i = 0 \\ c_i(0|1) & \text{for} \quad h(x_i) = 0 \quad \& \quad y_i = 1 \\ 0 & \text{for} \quad h(x_i) = 1 \quad \& \quad y_i = 1. \end{cases} \tag{9}$$

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

$$R_{emp}(h, \boldsymbol{x_{acc}}) = n_{acc}^{-1} \sum_{i=1}^{n_{acc}} L(x_{acc,i}, h(x_{acc,i}), y_{acc,i}). \tag{10}$$

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?