BAYESIAN MODELS OF THE AREA UNDER ROC CURVE

1. Model

1.1. **Binary Model.** One distinctive property of the AUC of binary classification is its relationship with the Wilcoxon statistic, first introduced by Hanley and McNeil [*]. Referring to their work, the area under ROC curve is numerically equal to Wilcoxon statistic, which is the probability that a classifier ranks a randomly chosen positive sample before a randomly chosen negative sample. This paper develops binary AUC probabilistic model based on the property.

Let us consider a binary classification task where an instance is labeled as either positive or negative. Suppose a classifier is validated on a test set M, from which all positive instances compose a positive class M_1 and all negative instances compose a negative class M_2.

Consider a random experiment E where we pair a positive sample and a negative sample at random and ascertain whether the classifier estimates a higher score to the positive sample than the negative one. This experiment is exactly a Bernoulli trial since the outcome of E is either success or failure. According to Hanley and McNeil, the probability of success, denoted by p, equals the area under ROC curve. That is,

$$AUC = P(E = success) = p \tag{1}$$

Thus, we compute the posterior probability of p to work out the details about the AUC. Suppose that we repeat the above Bernoulli trial n independent times. We use a random variable X to represent the number of successes in the n trials. Obviously, X follows a Binomial distribution with parameter n and p: $X \sim B(n, p)$, i.e.,

$$\Pr[X = x; n, p] = \binom{n}{x} p^x (1-p)^{n-x}, x = 0, 1, 2, ..., n \tag{2}$$

Since the Beta distribution is conjugate to the Binomial distribution, one natural way is to assign a Beta prior to p: $p \sim Beta(\alpha, \beta)$, i.e.,

$$\Pr[p; \alpha, \beta] = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta))} p^{\alpha-1} p^{\beta-1} \tag{3}$$

We simply set both α and β to 1 before experiment, which yields a uniform distribution. This is consistent with the fact that before the trials we do not favor a higher or lower value for p.

To derive the posterior distribution of p, we need to combine the observed data with our prior knowledge. In this context, the observed data is n and x, which can be directly obtained from the experiment. More specifically, we can pair samples from $M1$ and $M2$ one by one and count the total pairs in correct order, that is

$$n = |m_1| \cdot |m_2|$$

$$x = |\{(a, b) \mid a \in M_1, b \in M_2 \ and \ Score_a > Score_b\}| \tag{4}$$

where $Score_i$ is the score that the classifier assigns to instance i to estimate its likelihood of belonging to positive class. By ranking all samples in advance, we can derive a simpler expression for x:

$$x = \sum_{i \in M_1} Rank_i - \frac{m_1(m_1 + 1)}{2} \tag{5}$$

where $Rank_i$ is the ranking of sample i when all instances are sorted by their scores in increasing order.

1.2. **Multi-class Model.** While the AUC measure is originally designed for binary classification problems, there are some ways to extent it to circumstance with more than two classes. One approach, taken by Hand and Till, is to average all the AUC performance measure between each pair of classes. Averaging in this way avoids considering the prevalence of classes in data, thus makes the measure insensitive to the class distribution. In their measure, the overall multi-class AUC is given by the following equation:

$$AUC = \frac{2}{|C| \left(|C| - 1\right)} \sum_{i,j \in C} A(i, j) \tag{6}$$

where C is the set of predefined classes, and $A(i, j)$ is the pairwise AUC measure based solely on class i and class j. To calculate $A(i, j)$, Hand and Till extend the probabilistic interpretation of the AUC of binary classification to pairwise comparisons in multiple classification. They define $A(i \mid j)$ as the probability that a pair of randomly chosen samples from class i and j will be correctly ranked by classifier according to the estimated score of belonging to class i. While $A(j \mid i)$ is defined as the probability that a pair of randomly chosen samples from class i and j will be correctly ranked by classifier according to the estimated score of belonging to class j. Note that generally $A(i \mid j) \neq A(j \mid i)$ because they are calculated on different bases. Thus, $A(i, j)$ is obtained by averaging $A(i \mid j)$ and $A(j \mid i)$, namely, $A(i, j) = [A(i \mid j) + A(j \mid i)]/2$.

Given the definition and equation, we see that $A(i, j)$ essentially measures how well classes i and j are separated by the classifier without taking any other classes into account. That is to say, in this approach, each pairwise comparison is assumed independent from others. However, the assumption is not consistent with the facts — suppose a classifier can well distinguish between the class “China” and the class “Japan”, then we should believe that it is more likely to separate class “India” and class “Japan” well. Another undesirable fact in this algorithm is that the paris involving classes with little data will have a much higher degree of uncertainty than others.

To exploit the internal connection between the pairwise AUC and solve the problem with small size classes, we modify the previously mentioned probabilistic approach to a hierarchical model. Hyper-parameters are introduced, and serve as connectors which share information across all pairs of classes. In this way, the pairs with little data actually “borrow” knowledge from other pairs and therefore reduce the uncertainty of the estimate.

Consider a multiple classification problem with c different classes. Assume that the classes are labeled as $1, 2, ..., c$ and $m_1, m_2, ..., m_c$ represent the number of samples in each class correspondingly. For each pair of class i and j, define $E_{i,j}$ as a random experiment where we choose a drawn pair from class i and j, and ascertain whether the classifier can correctly rank the pair according to the estimated score of belonging to class i. Note that $E_{i,j}$ is different from $E_{j,i}$. The probability of success on the trial $E_{i,j}$ is equal to $A(i \mid j)$, that is

$$AUC(i \mid j) = P(E_{i,j} = success) = p_{i,j} \tag{7}$$

The overall AUC is the average of all such $p_{i,j}$, of which the posterior distribution will be computed:

$$AUC = \frac{1}{c(c-1)} \sum_{i \neq j} A(i \mid j) = \frac{1}{c(c-1)} \sum_{i \neq j} p_{i,j} \tag{8}$$

Note that we make some transformation to Equation 6 to make it consistent with our model, but Equation 6 and Equation 8 are essentially the same.

Suppose the Bernoulli trial $E_{i,j}$ is repeated $n_{i,j}$ independent times. We use a random variable $X_{i,j}$ to represent the number of successes in the $n_{i,j}$ trials. Similarly to the previous binary version, $X_{i,j}$ follows a Binomial distribution with parameter $n_{i,j}$ and $p_{i,j}$: $X_{i,j} \sim B(n_{i,j}, p_{i,j})$, i.e.,

$$\Pr[x_{i,j}; n_{i,j}, p_{i,j}] = \tbinom{n_{i,j}}{x_{i.j}} {p_{i,j}}^{x_{i,j}} (1 - p_{i,j})^{n_{i,j} - x_{i,j}}, x_{i,j} = 0, 1, 2, ..., n \tag{9}$$

For each $p_{i,j}$, we assign a Beta prior to it: $p \sim Beta(\alpha, \beta)$, i.e.,

$$\Pr[p_{i,j}; \alpha, \beta] = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} {p_{i,j}}^{\alpha - 1} {p_{i,j}}^{\beta - 1} \tag{10}$$

where α and β are hyper-parameters shared across all $p_{i,j} (i = 1, .., c, j = 1, .., c)$. Since α and β can only take on positive values, the exponential distribution would be a good choice of its prior distribution: $\alpha \sim Exp(\lambda_1), \beta \sim Exp(\lambda_2)$, i.e.,

$$\Pr[\alpha; \lambda_1] = \begin{cases} \lambda_1 \mathrm{e}^{-\lambda_1 \alpha} & (\alpha \geq 0) \\ 0 & (\alpha < 0) \end{cases}$$

$$\Pr[\beta; \lambda_2] = \begin{cases} \lambda_2 \mathrm{e}^{-\lambda_2 \beta} & (\beta \geq 0) \\ 0 & (\beta < 0) \end{cases} \tag{11}$$

The observed data in this model is $n_{i,j}$ and $x_{i,j}$, and the value of them can be obtained from the experiments as previously described.