Unlocking Unlabeled Data: Ensemble Learning with the Hui-Walter Paradigm for Performance Estimation in Online and Static Settings

Anonymous authors
Paper under double-blind review

Abstract

In the realm of machine learning and statistical modeling, practitioners often work under the assumption of accessible, static, labeled data for evaluation and training. However, this assumption often deviates from reality where data may be private, encrypted, difficult-to-measure, or unlabeled. In this paper, we bridge this gap by adapting the Hui-Walter paradigm, a method traditionally applied in epidemiology and medicine, to the field of machine learning. This approach enables us to estimate key performance metrics such as false positive rate, false negative rate, and prior in scenarios where no ground truth is available. We further extend this paradigm for handling online data, opening up new possibilities for dynamic data environments. Our methodology, applied to two diverse datasets, the Wisconsin Breast Cancer and the Adult dataset, involves partitioning each into latent classes to simulate multiple data populations and independently training models to replicate multiple tests. By cross-tabulating binary outcomes across ensemble categorizers and multiple populations, we are able to estimate unknown parameters through Gibbs sampling, eliminating the need for ground-truth or labeled data. This paper showcases the potential of our methodology to transform machine learning practices by allowing for accurate model assessment under dynamic and uncertain data conditions.

1 Introduction

It is a well-known fact that the amount of data is growing profusely. Most techniques from statistics, machine learning, and deep learning saw their development when data were harder to come by, and thus they have the unstated assumption that applications make predictions based on data that are the same or similar to the training data. Traditional statistics textbooks go to great lengths to stress the difference between interpolation and extrapolation and warn against the dangers of the latter. Much of this data comes in online or via streaming and is unlabeled. The body of research that attempts to move machine learning online still assumes that the data are mostly labeled (Gomes et al., 2019).

In addition, evaluating performance is a required and often daunting task while deploying a machine learning (ML) system to production. Performance is often challenging to measure due to insufficient quantities and a variety of labeled ground-truth data. Additionally, data may be private, encrypted, or infeasible to measure, and the data landscape may constantly be changing for online or streaming algorithms. The actual population distribution is often unknown. Thus, a relatively simple comparison of the machine learning system's output distribution against the population's distribution is impossible.

Fortunately, the problems described are not unique to ML. Epidemiology is another discipline that also relies on precise efficacy measurements with insufficient ground truth and population data. The most prominent example in epidemiology is disease testing. Epidemiologists distinguish between Efficacy and Effectiveness when talking about moving from a laboratory environment, where controlled experiments are possible; this is called Efficacy, to deploying the test to populations of test recipients; this is called effectiveness (Singal et al., 2014). By analogy, a Covid test developed in a laboratory measures effectiveness through false negative and

false positive rates with a double-blind controlled trial. The effectiveness is then measured by deploying that test to an unknown population. In an applied data science setting, this mimics an applied data scientist training a model in a notebook with labeled data and then deploying that model to production. The field of epidemiology faces similar limitations as fields such as finance, health care, and social sciences, where ground truth is impossible or expensive to obtain. Any field where unsupervised learning is ubiquitous can benefit from our approach.

We propose a method that borrows concepts from how epidemiologists measure efficacy or effectiveness to ML systems in a production environment where no gold standard or ground truth is known. This idea is likely to appeal to anyone who has worked in an applied data science environment. We use the Hui-Walter paradigm, which uses Bayesian methods to estimate the prior probability (prevalence) and the false positive and false negative rates of ensemble models trained to detect the outcome running in production or online (Hui & Walter, 1980). This estimate uses binary outcomes of diagnostic tests, i.e., a positive or negative test for Covid, for multiple tests and multiple populations arranged in a table. We then performed Gibbs sampling to estimate the unknown parameters, priors, false positive rates, and false negative rates, without relying on ground truth data. We can perform these estimations because more than one population gives us enough equations and unknowns to solve for the parameters of interest.

We expand on this idea by introducing a closed-form solution where two populations occur naturally, i.e., multiple data centers or regions, and use it to extend this framework to the online setting. When only one population is available, we use latent structure models for both online and offline methods. Many applied settings employ the technique of bagging or other ensemble methods. These ensemble methods naturally extend to our methodology.

In our experiment, we partitioned the Wisconsin Breast Cancer and Adult data sets into latent classes for demonstration purposes to simulate a scenario with multiple data populations and independently train models to simulate multiple tests.

2 Our Contribution

Our contribution is twofold. First, we show how concepts from epidemiology can be modified to help machine learning practitioners when data are static. Second, we extend this framework to work on-line, e.g., for streaming data.

3 Proposed Method

The Hui-Walter paradigm is the technique epidemiologists use to measure the efficacy of effectiveness in the absence of ground truth. The paradigm is based on comparisons of more than one test and their results applied to more than one logical subset of the population Hui & Walter (1980). When evaluating ML systems, the tests are analogous to binary categorizers, the results are the predictions, i.e., 1 or 0 binary classifications correspond to a positive or negative test result, and the population subsets are groups of observations assumed to be in distinct sub-populations.

To obtain distinct sub-populations when only one is available, which is often the case for ML system evaluation, we propose using Latent Structure Analysis, comprised of Latent Class Analysis (LCA) or Latent Profile Analysis (LPA), to partition the data into two populations. Of course, in most applied machine learning settings, multiple distinct subpopulations are easy to come by; for example, a practitioner has various data centers or databases from which to choose.

The framework we propose is two-fold. For a machine learning practitioner working in an industry where often multiple populations are available for selection, we may consider the Hui-Walter paradigm applied to two or more models trained to detect outcomes based on those data. In cases where only one population is present, we use LPA or LCA to separate our data into multiple latent populations.

Then, we give a closed-form solution that applies the Hui-Walter to data streams, measuring the prior probabilities and false positive and false negative rates on unlabeled data streams. Our proposed solution

dynamically assigns latent classes to the streaming data and calculates the unknown performance metrics on the unlabeled data streams.

3.1 Latent Class and Latent Profile Analysis

Latent structure analysis is a statistical method used to identify underlying patterns or structures in data. Latent class analysis is commonly used in psychology, sociology, and public health to identify subgroups or classes within a population based on their responses to a series of categorical variables. Similarly, researchers use latent profile analysis for continuous variables, and both of these models generalize under the moniker latent structure analysis.

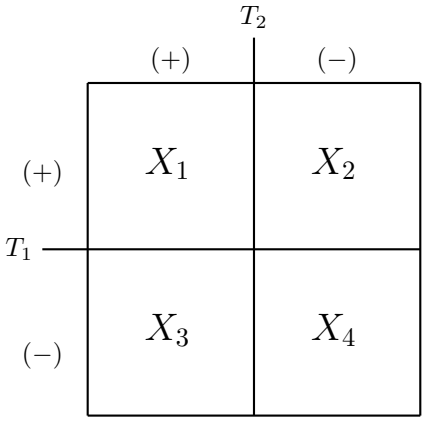
Latent class analysis and latent profile analysis are powerful tools for identifying patterns and structures in data, and they have a wide range of applications in fields such as psychology, sociology, public health, and applied machine learning. They are valuable tools for data scientists who want to understand the underlying structure of a population and identify subpopulations based on categorical, ordinal, or continuous variables Goodman (1974).

In latent profile analysis, we assume that each latent class follows a normal distribution in its features. Note that the normality assumption applies only to these subpopulations, not their joint distribution, which we model as a Gaussian mixture model. Each latent class $(1, \ldots, K)$ has a density component in the mixture model.

3.2 Hui-Walter

The Hui-Walter paradigm is a framework from epidemiology for estimating the false positive rate, false negative rate, and prior probabilities across multiple populations, with more than one test. In the epidemiological context, the Hui-Walter paradigm was designed for disease tests, i.e. Covid tests, where the results are 0 or 1 for does not-have-the-disease and has-the-disease, respectively. Since we are applying this paradigm to the machine learning context, we will discuss the method in terms of classifications from binary categorizers. The Hui-Walter paradigm suggests that we can form a $n \times m \times k$ contingency table for n binary categorizers, m populations, and k classes. The goal of this method is to estimate a parameter vector $(\hat{\alpha}, \hat{\beta}, \hat{\theta})$ for each model and population, where α is the false positive rate or one minus the specificity, β is the false negative rate or one minus the sensitivity, and θ is the prior probability of the positive class in the population. When $m = n = k = 2$, we can show this result using maximum likelihood estimation (Hui & Walter, 1980).

Figure 1: Two tests on one population.



For one population, consider Table 1. One population is multinomial distributed cell data. The cells X_1, X_4 are where both models (T_1 and T_2) agree, and X_2, X_3 are where the models disagree.

Maximum likelihood estimation (MLE) is a method for estimating the parameters of a statistical model given a set of observations. One common application of MLE is to estimate the parameters of the multinomial distribution (Poleto et al., 2014).

The multinomial distribution is a generalization of the binomial distribution that allows for more than two possible outcomes. Given a set of N independent trials, the multinomial distribution describes the probability of observing a specific combination of counts. The probability mass function of the multinomial distribution is given by:

$$P(x_1, x_2, \ldots, x_k | n, p_1, p_2, \ldots, p_k) = \frac{n!}{x_1! x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$$

where x_i is the number of times the i-th outcome was observed, n is the total number of trials, and p_i is the probability of the i-th outcome occurring.

To perform MLE on the multinomial distribution, we wish to find the values of $p_1, p_2, \ldots, p_k$ that maximize the likelihood of the observed data. We can do this by taking the product of the probabilities of each outcome and maximizing this product for the parameters. This procedure is equivalent to maximizing the log-likelihood function:

$$\mathcal{L}(\mathbf{p}) = \sum_{i=1}^{k} x_i \log p_i$$

The log-likelihood function is easier to work with because the product of many small numbers can be unstable, while the sum of the logs is more stable for calculation.

We can find the maximum likelihood estimates by setting the partial derivatives of the log-likelihood function to zero and solving for p_i. This reasoning leads to the following unbiased estimators:

$$\hat{p}_i = \frac{x_i}{n}$$

One property of the MLE estimates for the multinomial distribution is that they are asymptotically efficient, which means that as the number of trials n increases, the variance of the estimates decreases. However, for small n, the MLE estimates can be highly variable and may not accurately reflect the actual values of the parameters.

In general, MLE is a valuable method for estimating the parameters of the multinomial distribution and has attractive theoretical properties. However, it is crucial to consider the method's limitations, especially when working with small sample sizes.

In this context, the advantage of maximum likelihood for a single population is that p_1 is the probability that both tests produce a positive result. We solve this over-determined system by minimizing the likelihood function. In our simplified setup, we have a parameter vector $p = (p_1, p_2, p_3, p_4)$, which we wish to estimate. Because these probabilities in question are the outcomes of binary classifiers or diagnostic tests, we may express them in terms of the false positive and false negative rates of the models in question and the population's prior probability. Let α_1, α_2 be the false positive rates for T_1 and T_2, let β_1, β_1 be the false negative rates for T_1 and T_2, and let θ be the prior probability for the positive class (base rate) of the population (or the disease in the epidemiological case).

$$\begin{aligned}
p_1 &= \theta(1-\beta_1)(1-\beta_2) + (1-\theta)\alpha_1\alpha_2 \\
p_2 &= \theta(1-\beta_1)\beta_2 + (1-\theta)\alpha_1(1-\alpha_2) \\
p_3 &= \theta\beta_1(1-\beta_2) + (1-\theta)(1-\alpha_1)\alpha_2 \\
p_4 &= \theta\beta_1\beta_2 + (1-\theta)(1-\alpha_1)(1-\alpha_2)
\end{aligned}$$

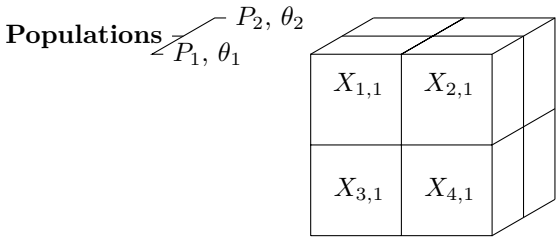


Figure 2: Experimental setup with a $2 \times 2 \times 2$ contingency table. Variables $X_{1,1}, X_{2,1}, X_{3,1}$ and $X_{4,1}$ are cell frequency counts from a product-multinomial distribution.

We need to consider two populations because the base rates of a single population are unknown. We assume that α_i and β_i are the same in both populations for $i = 1, 2$, therefore we have six equations with six unknowns. In other words, p_i where $i = 1, 2, \ldots 8$, can be written algebraically as combinations of 6 variables. Therefore, we use two populations so that the above estimates are solvable (Johnson et al., 2001). At least two populations guarantee that we have as many equations as unknowns. For two populations, we have two independent multinomial distributions. Therefore, the likelihood function is the product of the likelihood functions (Hui & Walter, 1980). Now consider two populations, P_1 and P_2 with prior probabilities θ_1 and θ_2. In that case, our two-by-two-cell data becomes:

Now that we have two populations, we can solve the prior probabilities to eliminate the unknowns in our estimates. Because we have two independent multinomial distributions, which can be considered a product-multinomial distribution, the likelihood function is the product of the individual likelihood functions adjusted with the sample size. Let n_1, n_2 be the total sample size of each population, then the likelihood function is

$$l = l_1 l_2 = \prod_{j}^{2} \prod_{i}^{2} p_{i,j}^{x_{ij}}$$

Where $p_{i,j}$ is the probability of observing the j th outcome in the i th population.

We want to estimate the parameter vector $\boldsymbol{X} = (\alpha, \beta, \theta)$. A real solution with coordinates that satisfy these constraints is a maximum likelihood estimate. In other cases, when the points of the global maximum are complex or lie outside the unit hypercube, estimates can be obtained by numerical maximization of l with the solution restricted to be within the unit hypercube, again subject to appropriate constraints. Asymptotically, because the observed proportions of cell counts are continuous functions of the parameters, the maximum likelihood solution will converge to (α, β, θ). Let $L = ln(l)$, then we can formulate the information matrix

$$\mathcal{I}(\boldsymbol{X}) = \mathbb{E}[\mathbf{Hess}(\mathbf{X} \otimes \mathbf{X})]$$

In other words, when the MLE resides within the interior point of the feasible set, we can solve for the information matrix by taking the expectation value of the hessian of the outer product of our parameter vector, $\boldsymbol{X}$ (Blasques et al., 2018).

Inverting this matrix gives the variance and covariance matrix. Now that we have shown that there is a maximum likelihood estimate for the parameter vector (α, β, θ) exists for the $2 \times 2 \times 2$ case, it is trivial to see that this generalizes to the $n \times m \times k$ case for arbitrary dimensions of the population outcome tensor. We can generally find a solution to the parameter vector using Bayesian methods such as Monte Carlo estimates or Gibbs sampling to estimate our parameter vector, $\boldsymbol{X} = (\alpha, \beta, \theta)$.

Maximum likelihood estimation is a method used to estimate the parameters of a statistical model given a set of observations. In the case of product-multinomial distributed data, this involves finding the values of the parameters that maximize the likelihood of observing the provided data.

To understand maximum likelihood estimation for product-multinomial distributed data, it is first necessary to understand the concept of a likelihood function. Given a set of observations and a statistical model with unknown parameters, the likelihood function is a function of the parameters that describe the probability of observing the data.

In the case of product-multinomial distributed data, the likelihood function is given by:

$$L(\boldsymbol{X}) = \prod_{i=1}^{n} \prod_{j=1}^{k} p_{ij}^{x_{ij}}$$

where n is the number of observations, k is the number of categories, x_{ij} is the number of times the category j was observed in the population i and $\boldsymbol{X}$ is a vector of the parameters of the model.

The maximum likelihood estimate of the parameters, denoted by $\hat{\boldsymbol{X}}$, is found by maximizing the likelihood function of the parameters:

$$\hat{\boldsymbol{X}} = \arg\max_{\boldsymbol{X}} L(\boldsymbol{X})$$

We can estimate this using numerical optimization techniques.

The maximum likelihood estimate is not necessarily the best estimate of the parameters, and there are other methods of estimation, such as the method of moments and Bayesian estimation, that may be more appropriate in some cases. However, the maximum likelihood estimation is a widely used and well established method with several attractive properties, such as asymptotic efficiency (meaning that it becomes more accurate as the number of observations increases)(Hui & Walter, 1980).

For our simplified case with two populations we get

$$\frac{\partial L}{\partial p_i} = \frac{x_i}{p_i} - \lambda = 0, \frac{\partial L}{\partial p_j} = \frac{x_j}{p_j} - \mu = 0$$

for $i = 1, \ldots, 4$, and for $j = 5, \ldots, 8$. The maximum likelihood then is $\mathbf{p} = (\frac{p_1}{n_1}, \ldots, \frac{p_4}{n_1}, \frac{p_5}{n_2}, \ldots, \frac{p_8}{n_2})$. This closed formula gives the maximum likelihood estimate for the product-multinomial distribution. This MLE extends naturally to populations k.

One of the assumptions of Hui-Walter is that the tests be conditionally independent Hui & Walter (1980). We can test for conditional independence, which can easily be tested by performing a goodness-of-fit hypothesis test on the contingency table with a G-test (Hui & Zhou, 1998).

3.3 Gibbs Sampling

Gibbs sampling, named after the physicist Josiah Willard Gibbs, is an analogy between the sampling algorithm and statistical physics. The algorithm was described by brothers Stuart and Donald Geman in 1984, some eight decades after the death of Gibbs, in their work on algorithmic image restoration (Geman & Geman, 1984). Gibbs sampling is a Markov chain Monte Carlo (MCMC) technique, typically the Metropolis-Hastings algorithm, for estimating the marginal distribution of variables in a multivariate distribution.

In Gibbs sampling, we start with an initial guess for the values of the variables and then iteratively sample each variable from its conditional distribution given the current values of the other variables. This process is repeated many times to obtain a sample from the joint distribution of the variables. Then we use the sample to estimate the marginal distribution of each variable or to compute other statistical quantities of interest.

Gibbs sampling has several advantages over other MCMC techniques. It is relatively simple to implement and easy to obtain good mixing (i.e., the convergence of the samples to the target distribution) in practice. It is also well suited to parallelization, which can speed up the sampling process.

However, Gibbs sampling can be inefficient when the variables are highly correlated or anticorrelated, as it can take many iterations to explore the full joint distribution. It is also sensitive to the choice of initial values and can get stuck in local modes for poorly chosen initial values.

Gibbs sampling finds applications in statistical physics, machine learning, and other fields where it is necessary to sample from multivariate distributions. It is beneficial for estimating the posterior distribution in Bayesian statistical models and is usually used with other MCMC techniques such as Metropolis-Hastings sampling.

Given a multivariate distribution, Gibbs sampling allows us to sample from a conditional distribution rather than to marginalize by integrating over a joint distribution; the joint distribution is often not known explicitly or difficult to sample from directly.

The algorithm is shown in Algorithm 1:

Algorithm 1 Gibbs Sampling

procedure GibbsSample($x_1, \ldots, x_n, T$)
  Initialize $X^{(0)} = (x_1, \ldots, x_n)$
  for $t = 1, 2, \ldots, T$ **do**
    for $i = 1, \ldots, n$ **do**
      Sample $X_i^{(t)}$ from $p(X_i^{(t)} \mid X_1^{(t)}, \ldots, X_{i-1}^{(t)}, X_{i+1}^{(t-1)}, \ldots, X_n^{(t-1)})$
    end for
  end for
  return samples $X^{(0)}, \ldots, X^{(T)}$
end procedure

This algorithm takes a set of variables $X = (x_1, \ldots, x_n)$ and T as input and returns T samples from their joint distribution. It does this by iteratively sampling each variable from its conditional distribution given the current values of the other variables. The algorithm ends after a fixed number of iterations T.

3.4 Hui-Walter Online

Online Machine Learning refers to training machine learning models on data that arrive continuously and in real time rather than using a static data set. In this setting, the model updates its parameters as new data arrive and makes predictions accordingly (Gomes et al., 2019). This approach is helpful in applications such as streaming data analysis, recommendation systems, and dynamic pricing.

One of the key benefits of online machine learning is its ability to handle large amounts of data efficiently, which is especially important in scenarios where data volume is increasing rapidly. Additionally, online machine learning algorithms can adapt to changing data distributions, which is critical in applications where the underlying data distribution constantly evolves. This process enables the models to provide accurate predictions even in dynamic environments, improving the overall system performance.

Various algorithms work online, including stochastic gradient descent, k-means, and online support vector machines (Gomes et al., 2019). These algorithms differ in their update rules, but all aim to minimize the cumulative loss incurred over time as new data arrive. A critical aspect of online machine learning is the choice of the loss function, which determines the trade-off between the accuracy of the model and its ability to adapt to new data. However, these methods are still limited to using labeled data and assuming that ground truth is available. Online machine learning provides a powerful and flexible approach to real-time training machine learning models, but fails in some applied settings where the data is encrypted.

We saw above that the maximum likelihood estimate for the product-multinomial distribution is $\mathbf{p} = (\frac{p_1}{n_1}, \ldots, \frac{p_4}{n_1}, \frac{p_5}{n_2}, \ldots, \frac{p_8}{n_2})$. We use this estimate to come up with a naive estimate of the base rate for a population taking $\frac{X_i}{n}$ for $i = 1, 2$. We use this estimate to solve for the other parameters related to false-positive and false-negative rates and keep a tally of the streaming data as a time series. We then use the prior probability of our base class from our training environment to test for significant prior drift.

A closed-form solution to the Hui-Walter estimates without a confidence interval appears in (Hui & Walter, 1980; Enùe et al.). We reformulate our contingency table as follows:

	T_2 (pos)	T_2 (neg)	
T_1 (pos)	a_i	b_i	g_i
T_1 (neg)	c_i	d_i	h_i
	e_i	f_i	n_i

where i ranges over each population. For simplicity, we will focus on $i = 2$ but mention that this can be extended to more populations and classes. By computing the row sums and column sums of this matrix, we look at the following discriminant:

$$F = \pm\sqrt{\left(\frac{g_1e_2 - g_2e_1}{n_1n_2} + \frac{a_1}{n_1} - \frac{a_2}{n_2}\right)^2 - 4\left(\frac{g_1}{n_1} - \frac{g_2}{n_2}\right)\frac{a_1e_2 - a_2e_1}{n_1n_2}}$$

If this discriminant is zero or complex, then this online method cannot be calculated (Hui & Walter, 1980). One possible explanation for this phenomenon is the existence of Simpson's paradox and the algebraic geometry that arises when dealing with three-way contingency tables (Slavkovi´c et al., 2009). The closed solutions to the Hui-Walter estimates are as follows:

$$\begin{aligned}
\hat{\theta_1} &= \frac{1}{2} - \left[\frac{g_1}{n_1}\left(\frac{e_1}{n_1} - \frac{e_2}{n_2}\right) + \frac{g_1}{n_1}\left(\frac{g_1}{n_1} - \frac{g_2}{n_2}\right) + \frac{a_2}{n_2} - \frac{a_1}{n_1}\right]\frac{1}{2F} \\
\hat{\theta_2} &= \frac{1}{2} - \left[\frac{g_2}{n_2}\left(\frac{e_1}{n_1} - \frac{e_2}{n_2}\right) + \frac{g_2}{n_2}\left(\frac{g_1}{n_1} - \frac{g_2}{n_2}\right) + \frac{a_2}{n_2} - \frac{a_1}{n_1}\right]\frac{1}{2F} \\
\hat{\alpha_1} &= \left(\frac{g_1e_2 - e_1g_2}{n_1n_2} + \frac{a_2}{n_2} - \frac{a_1}{n_1} + F\right)\left[2\left(\frac{e_2}{n_2} - \frac{e_1}{n_1}\right)\right]^{-1} \\
\hat{\alpha_2} &= \left(\frac{g_2e_1 - e_2g_1}{n_1n_2} + \frac{a_2}{n_2} - \frac{a_1}{n_1} + F\right)\left[2\left(\frac{g_2}{n_2} - \frac{g_1}{n_1}\right)\right]^{-1} \\
\hat{\beta_1} &= \left(\frac{f_1h_2 - h_1f_2}{n_1n_2} + \frac{d_2}{n_2} - \frac{d_1}{n_1} + F\right)\left[2\left(\frac{e_2}{n_2} - \frac{e_1}{n_1}\right)\right]^{-1} \\
\hat{\beta_2} &= \left(\frac{f_2h_1 - h_2f_1}{n_1n_2} + \frac{d_2}{n_2} - \frac{d_1}{n_1} + F\right)\left[2\left(\frac{g_2}{n_2} - \frac{g_1}{n_1}\right)\right]^{-1}
\end{aligned}$$

Using the above formulas, we can increment the streams' counts and apply the estimates to streaming data, allowing us to examine the false positive and false negative rates, and prior probabilities of streaming data on a given time step. Boosting is a family of algorithms that convert multiple weak learners in parallel to strong learners. With Hui-Walter online, we can leverage multiple weak learners to assess the false positive and false negative rates, and prior probabilities to improve the predictions by making them "prior aware" of the unlabeled data streams. Multiple learning frameworks exist that allow for better predictions by making the learner prior aware (Davis, 2020). Note that the sign of F is either positive or negative and sometimes has no solution (Hui & Walter, 1980; Johnson et al., 2001). So, this value is assigned based on what yields *plausible* solutions, i.e., within a probability simplex.

Our online algorithm runs as follows in Algorithm 2:

In Algorithm 2, $\mathcal{X}$ is the data set and f_1 and f_2 trained on subsets of the columns, T is the number of instances to receive, $\mathcal{T}$ is a contingency table with as many dimensions as populations, $\mathcal{H}$ is the history of previously seen samples. The implementation of $\mathcal{H}$ could cache the previously seen samples in an LRU cache. The procedure calculates the latent profiles on the history $\mathcal{H}$ to assign the data to a population. After the n instances have been predicted, the estimates for θ, α, β start to be calculated and updated. The selection of

Algorithm 2 Online Hui-Walter on Streaming Data

procedure ONLINE HUI-WALTER($\mathcal{X}, T$)
  $t \leftarrow 1$
  Initialize models f_1, f_2 with $\mathcal{X}$
  Initialize three-way contingency table $\mathcal{T}$ and history $\mathcal{H}$
  while $t \leq T$ **do**
    Receive instance x_t
    Predict labels $y_1 = f_1(x_t), y_2 = f_2(x_t)$
    Add x_t to $\mathcal{H}$
    Get LPA profiles of $\mathcal{H}$
    Update cell counts of $\mathcal{T}$
    if t » n **then**
      Calculate $\mathbf{X} = (\theta, \alpha, \beta))$
    end if
    $t \leftarrow t + 1$
  end while
end procedure

n should be reasonable such that there is enough stability in the values for θ, α, β (lessening variation from one instance to the next) without sacrificing too many of the samples. This analogous to a "burn-in" concept commonly seen in MCMC (Hamra et al., 2013), but keep in mind that we are not discarding samples or assuming stationarity has been achieved.

4 Experimental Results

4.1 Data Sets

4.1.1 Wisconsin Breast Cancer Data Set

The Wisconsin breast cancer data set is a clinical data collection used to analyze breast cancer tumors. It was created in the early 1990s by Dr. William H. Wolberg, a physician and researcher at the University of Wisconsin-Madison. The data set contains 569 observations and includes information on patient characteristics, such as age, menopausal status, and tumor size, as well as diagnosis and treatment of cancer. In the diagnosis of 569 observations, 357 were benign and 212 were malicious.

The Wisconsin Breast Cancer Data Set has been widely used in machine learning and data mining to predict breast cancer diagnosis and prognosis. Researchers have used the data set to develop algorithms and models to classify tumors as benign or malignant and predict the probability of breast cancer recurrence or survival. These predictions can guide the treatment of patients with breast cancer and identify potential risk factors for the disease.

The Wisconsin breast cancer data set has been widely used in research studies and has contributed significantly to our understanding of breast cancer. However, it is essential to note that the data set has some limitations. For example, data is based on a specific population of patients and may not represent all cases of breast cancer. Furthermore, the data are from the 1990s and may not reflect more recent advances in breast cancer diagnosis and treatment.

Despite these limitations, the Wisconsin breast cancer data set remains an essential resource for researchers studying breast cancer and working to improve the diagnosis and treatment of this disease. It has contributed significantly to our understanding of breast cancer and will continue to be a valuable resource for researchers in the field.

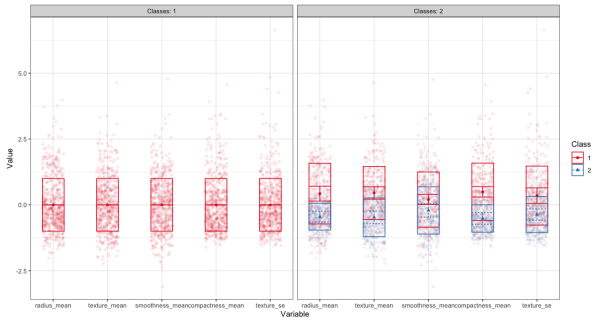


Figure 3: Five features from the Wisconsin Breast Cancer Data Set examined with Latent Profile Analysis

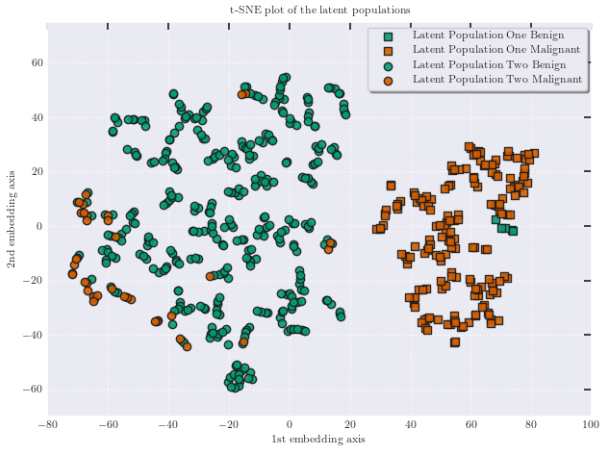


Figure 4: T-SNE dimension reduction of both latent populations.

In addition to its role in medicine, it has been used as an almost canonical data set for data science education and benchmarking of ML algorithms. It is one of the most downloaded data sets on the UCI Machine Learning Repository(Dua & Graff, 2017)[1].

We trained a random forest model and a support vector classifier model using these data. To ensure that the models are independent, we trained each model on mutually exclusive columns of the data.

4.1.2 Adult Data Set

This is another canonical data set from the UCI Machine Learning Repository(Dua & Graff, 2017) [2]. It contains census data from 1994 for the task of predicting whether an adult's income exceeds $50k/year. There are 32,561 observations described with a mix of categorical and continuous characteristics such as age, sex, education level, occupation, capital gains, and capital losses.

For our investigation, we partitioned it into two populations based on sex: Male and Female.

4.2 Latent Profile Analysis

We performed a Latent Profile Analysis (LPA) on the Wisconsin Breast Cancer data set to obtain two populations. Due to collinearities in the data set, a subset of features was selected to estimate the latent profiles. This subset was determined via hierarchical clustering on Spearman rank-order correlations and selecting one feature from each cluster, where clusters were defined as being separated by at least a distance

[1]https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29
[2]https://archive.ics.uci.edu/ml/datasets/Adult

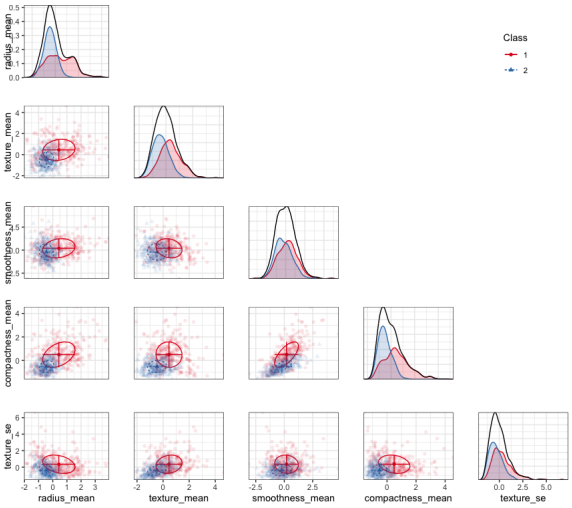


Figure 5: Bivariate plot of 2 Latent Profiles

of 1 per Ward's linkage. LPA, assuming varying variances and varying covariances, was used to fit models assuming 1 through 10 latent classes. The approximate weight of evidence (AWE) criterion was used to identify the optimal number of classes, which was 2. AWE was used instead of BIC for model selection due to its ability to select more parsimonious models (Banfield & Raftery, 1993), especially relevant because we assumed complex parameterization (varying variances and varying covariances).

In Figure 5, we can see two distinct latent classes or subpopulations among the data. Figure 3 shows the columns of the data partitioned into one and then two latent classes. In Figure 4, we performed the t-SNE (Maaten & Hinton, 2008) dimension reduction of the columns of the Wisconsin Breast Cancer data set and labeled it according to the latent classes found with the analysis of the latent profile. We see that latent profile analysis and t-SNE separated the data into two latent populations.

4.3 Hui-Walter

4.3.1 Wisconsin Breast Cancer Data Set

Table 1: Wisconsin Breast Cancer Data

Experimental Results				
Model	TPR	TNR	FPR ($\hat{\alpha}$)	FNR ($\hat{\beta}$)
Model One	0.690	0.956	0.044	0.310
Model Two	0.547	0.899	0.101	0.453
Population	Prior ($\hat{\theta}$)			
Population One	0.819			
Population Two	0.035			

True Values				
Model	TPR	TNR	FPR	FNR
Model One	0.789	0.969	0.031	0.211
Model Two	0.549	0.868	0.132	0.451
Population	Prior			
Population One	0.635			
Population Two	0.122			

In our experiment, we split the data set into two latent populations using latent profile analysis. Then, we split the data set into a training and test data set, sizes 200 and 369, respectively. We then trained two classifiers, a support vector machine and a random forest classifier, in different columns of the data in the training data set. The experimental results contain the estimated parameters $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\theta}$. These parameters are the output of the Gibbs sampler, where initial guesses were assumed to each follow a $Beta(1, 1)$ distribution. Underneath in Table 4, the true values were calculated from the true labels and model classifications.

Table 2: Model Predictions Over The Two Latent Populations

		Model Outcome	
Population	Model	Malignant	Benign
Population One	Model One	109	80
	Model Two	89	100
Population Two	Model One	10	170
	Model Two	19	161

The experimental results in Table 4 show the results of the support vector and the random forest ensemble models on the Wisconsin Breast Cancer data set within each latent population. We compared the results of the models to the ground truth labels in the data set so that we could assess how well Hui-Walter performs when no labels are present. The experimental results of the Hui-Walter Sampling with Gibbs are shown in Table 4, and the actual results of the ground truth for comparison are shown in Table 4.

Table 5 contains a three-way contingency table with the categorizations of both models on both populations. Population 1 has more true malignant cases with 120 malignant tumors.

In Table 6, we have the confidence intervals for each parameter found during Gibbs sampling. All the confidence intervals for the errors contain the parameter we found during the evaluation phase of the experiment. However, the true prevalence (prior or base rate) is outside the confidence interval.

Table 3: Estimated Parameters

Parameter	CI	Mean	SD
$\hat{\theta}_1$	(0.693, 0.947)	0.819	0.064
$\hat{\theta}_2$	(3.9e-07, 0.084),	0.035	0.026
$\hat{\alpha}_1$	(0.004, 0.081)	0.044	0.020
$\hat{\alpha}_2$	(0.054, 0.150)	0.101	0.025
$\hat{\beta}_1$	(0.198, 0.425)	0.310	0.058
$\hat{\beta}_2$	(0.357, 0.547)	0.453	0.049

4.3.2 Adult Data Set

We used the sex of the adult to partition the data set, resulting in two populations of $10,771$ females and $21,790$ males. We performed an $\frac{80}{20}$ split on the Adult data set to produce a training and test data set, sizes $26,048$ and $6,513$, respectively. We trained two classifiers, a linear regression (LR) and a random forest classifier (RF), on different columns of the data in the training data set.

Table 4: Adult Data Set

Experimental Results				
Model	TPR	TNR	FPR ($\hat{\alpha}$)	FNR ($\hat{\beta}$)
Model One (LR)	0.608	0.992	0.008	0.392
Model Two (RF)	0.793	0.990	0.010	0.207
Population	Prior ($\hat{\theta}$)			
Population One (Female)	0.084			
Population Two (Male)	0.316			

True Values				
Model	TPR	TNR	FPR	FNR
Model One (LR)	0.446	0.942	0.058	0.554
Model Two (RF)	0.523	0.906	0.094	0.477
Population	Prior			
Female	0.110			
Male	0.305			

Table 5: Model Predictions Over The Two Subpopulations

		Model Outcome	
Population	Model	\$50K & Under	Over \$50K
Female	Model One (LR)	2003	123
	Model Two (RF)	1968	158
Male	Model One (LR)	3521	866
	Model Two (RF)	3258	1129

Table 6: Estimated Parameters

Parameter	CI	Mean	SD
$\hat{\theta}_1$	(0.0656, 0.103)	0.0837	0.00961
$\hat{\theta}_2$	(0.293, 0.338),	0.316	0.0114
$\hat{\alpha}_1$	(0.000109, 0.0153)	0.00816	0.00407
$\hat{\alpha}_2$	(5.15e-6, 0.0198)	0.00979	0.00567
$\hat{\beta}_1$	(0.353, 0.428)	0.392	0.0193
$\hat{\beta}_2$	(0.165, 0.247)	0.207	0.0209

Similarly to results from the Wisconsin Breast Cancer data set, the Hui-Walter method obtained close estimates of some of the parameters. The 95% confidence intervals for the rates of adults with over $\$50K$ income contained the ground truth rate for the male population (θ_2) and was 0.007 within the ground truth rate of the female population (θ_1).

Our experiments show that the model's false positive and false negative rates and prior can be estimated on unknown data sets with Gibbs sampling. Although we assumed the default distribution $Beta(1, 1)$ for the errors and priors, the results from both data sets are generally promising. With more iterations or a more specific distributional assumption, the results would improve.

4.4 Hui-Walter Online

Our main goal in this research is to answer how we estimate false positive and false negative rates of binary categorizers to a stream of unlabeled data without available ground truth. Current online machine learning theory still assumes that labels are available for the data in question (Gomes et al., 2019). Figure 6 shows an experiment where we streamed the Wisconsin Breast Cancer Data set with the `river` python library for online machine learning (Montiel et al., 2020) [3]. We trained a support vector machine using the mean radius and mean texture and a Random Forest Classifier using mean concavity and mean symmetry. The mean radius of the features, the mean texture, the mean radius and the mean texture were chosen, so each model had features from one leg of the dendrogram into which the features of this data set fall into (Pantazi et al., 2002). These features are different from the features used in the non-streaming experiment. The support vector machine is 90% accurate with $\alpha_1 = FPR = 0.18$ and $\beta_2 = FNR = 0.06$. The Random Forest is 88% accurate with $\alpha_2 = FPR = 0.23$ and $\beta_2 = FNR = 0.09$.

The categorizers then applied predictions to the online data and, for each time step, the contingency table was updated, the new sample was recorded, and latent profile analysis was applied to the new sample, plus the complete history of the samples. Because latent profiles are calculated when a new sample arrives, the latent class of a single sample will often flip. Due to this, Figure 6 shows the history of the parameters of interest calculated for the last latent profile. Additionally, due to the instability of the discriminant for Hui-Walter, the first 250 time steps are omitted because they contain discontinuities. The literature on Hui-Walter states that Hui-Walter may be better at estimating the test agreement than the actual false positive and false negative rates (Bertrand et al., 2005), and the results of our experiment in Figure 6 show the parameters that come close to the desired metrics of the training environment scaled up or down by a factor of 2 for α_i, β_i for $i = 1, 2$. Prior probabilities θ_1, θ_2 are more stable. Empirically, the priors are $\theta_1 = 0.11$ and $\theta_2 = 0.51$, calculated by taking the number of true values from the labels available in our experimental setup.

[3] https://github.com/online-ml/river

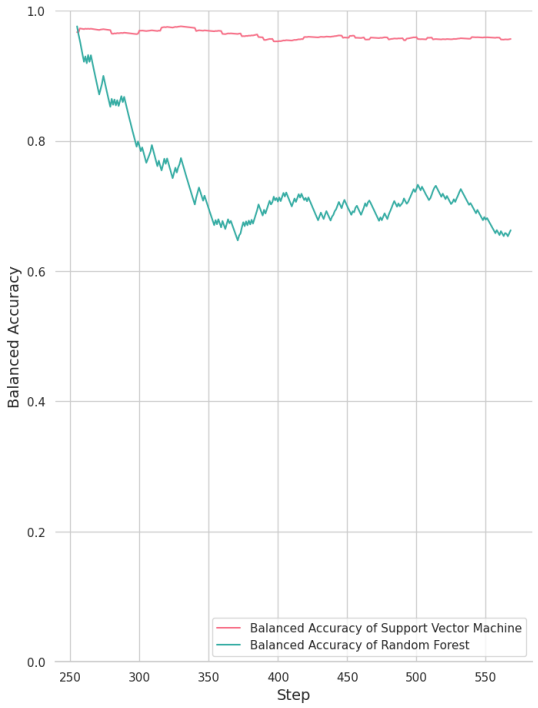


Figure 6: Online-Hui Walter applied to the Wisconsin Breast Cancer data set in a stream with the `river` python library for online machine learning Montiel et al. (2020) to simulate online data and the balanced accuracy for both classifiers was calculated over two latent subpopulations starting after a burn in period of 250 streaming samples.

Table 9 contains the mean absolute error of the values estimated with Hui-Walter versus those estimated in the training environment. We calculate the mean absolute error using the last 200 steps. This technique is similar to the burn-in period in Bayesian statistics [4].

The Rand index (RI) is a measure of the similarity between two cluster assignments, typically the ground truth labels and the clustering results Rand (1971). It is calculated by considering the number of pairs of data points that are either in the same cluster in both assignments (true positives, TP) or in different clusters in both assignments (true negatives, TN). Given the total number of pairs of data points, N, the Rand index can be computed as follows.

$$RI = \frac{TP + TN}{\binom{N}{2}}$$

The RI ranges from 0 to 1, where a value of 1 indicates perfect agreement between the two cluster assignments, and a value of 0 suggests that the assignments are completely dissimilar.

Now, let us discuss the relationship between the Rand index and accuracy. In the context of binary classification, precision is the proportion of true positive (TP) and true negative (TN) instances out of the total number of instances (P). It can be calculated as:

$$\text{Balanced Accuracy} = \frac{TPR + TNR}{2}$$

[4]Data and Experiments available: https://github.com/kslote1/hui-walter

In clustering problems, the Rand index can be considered a measure of the "accuracy" of a clustering algorithm when compared to the ground truth labels. However, it is essential to note that there are some differences between these two measures.

The Rand index is designed for clustering problems and considers both true positive and true negative pairs, whereas the accuracy is designed for classification problems and is based on the number of correctly classified instances. The Rand index is a measure of similarity between two cluster assignments and does not directly evaluate the quality of a single clustering result. On the contrary, accuracy directly evaluates the performance of a classification model. In summary, while the Rand index shares some similarities with accuracy, they are designed for different types of problem (clustering versus classification), and their interpretations and use cases are not entirely the same.

Classifier	Balanced Accuracy	Rand Index	Accuracy
RF	0.96	0.78	0.86
SVM	0.73	0.82	0.90

Table 7: Wisconsin Breast Cancer Data Set: Comparison of the mean Balanced Accuracy estimated by Hui-Walter on the online data, Rand Index, and Ground Truth Accuracy for Random Forest and Support Vector Machine

Classifier	Balanced Accuracy	Rand Index	Accuracy
LR	0.76	0.71	0.82
RF	0.77	0.70	0.81

Table 8: Adult Data Set: Comparison of the mean Balanced Accuracy estimated by Hui-Walter on the online data, Rand Index, and Ground Truth Accuracy for Logistic Regression and Random Forest

In order to compare our method to established methodologies from unsupervised learning, we compare the balanced accuracy found for our online implementation of Hui-Walter with the Rand Index which has been shown to relate the accuracy to the unsupervised learning setting (Garg & Kalai, 2018). For the Wisconsin Breast Cancer data set, we compare the balanced accuracy for the support vector machine and the random forest to the Rand index for both classifiers over both latent sub-populations. For the random forest model, we found a Rand index of 0.78 and a balanced accuracy of 0.96 from Hui-Walter, and the ground truth accuracy is 0.86. For the support vector machine, the balanced accuracy is 0.73, the RI is 0.82, and the ground truth accuracy is 0.90. These can be seen in Table 7 where we can see that online Hui-Walter gives a clear improvement over perhaps the only available baseline metric for this problem. The results are even more compelling for the analogous experiment run on the adult data set; as we can see in Table 8, the Balanced Accuracy is closer to the ground truth accuracy than the Rand index. The classifiers in question (Random Forest, Support Vector Machine, and Logistic Regression) were all chosen for their simplicity and ubiquity in production settings.

Table 9: Mean Absolute Error for Streaming Parameters

Parameter	MAE
θ_1	0.35
θ_2	0.03
α_1	0.10
α_2	0.15
β_1	0.05
β_2	0.09

Although this test yielded reasonable estimates for most parameters, the main issue with this approach for online estimation is the need for more standard errors, which is given by Bayesian estimation for the static case (Johnson et al., 2001). The worst estimate on the streaming data, θ_1, also closely matches the static data case found with the Gibbs sampling method.

Conclusion

Our results show that the Hui-Walter method works very well for static data and gives *plausible* results for online data, and this method requires further improvements. The data from this experiment are in a three-way contingency table, and a log-linear model could improve the estimate as is in (Hui & Zhou, 1998). Additionally, the algebraic geometry of the three-way contingency tables could yield different results, as there is a relationship between tensor factorization and the parameters of the product-multinomial model (Dunson & Xing, 2009). One of the limitations of using explicit formulas for the $2 \times 2 \times 2$ case is that sometimes the solutions do not have *plausible* solutions (Hui & Walter, 1980). In other words, it is possible to get a solution for the false positive rate greater than one or less than zero. Furthermore, the performance of the straightforward solutions on the online data lacked confidence intervals, which are problematic for applied settings, and the solutions are not as precise as the Gibbs sampler. One suggestion for further research on how this can be improved is to leverage online Gibbs sampling (Kim et al., 2016; Dupuy et al., 2017).

We have shown how to use the Hui-Walter paradigm to estimate false positive and false negative rates and prior probabilities when no ground truth is available. One of the core assumptions of statistical reasoning and its offshoot machine learning is that all of the models' possible data is collected ahead of time, and a practitioner only needs to sample appropriate training and holdout test sets. This assumption is far from the reality of many machine learning applications worldwide. Machine learning practitioners increasingly adapt to the reality that models make predictions based on previously unseen unlabeled data that change in the shape of the distribution over time. In this paper, we tackle a common problem in applied machine learning. Specifically, we derived a way to measure the efficacy or effectiveness of machine learning models in a situation where the data set is unknown and there are no assumptions about the data distribution. We have also shown how this methodology applies to static and streaming data.

Broader Impact Statement

This research is expected to significantly impact the field of applied machine learning by providing a methodology to assess the efficacy of the model in the common but challenging circumstances of incomplete, unknown, or dynamically changing data sets. The Hui-Walter method, despite its limitations, as highlighted in our study, offers a robust starting point for practitioners dealing with data ambiguity. Additionally, our explorations of potential improvements, such as the application of online Gibbs sampling, open a promising path towards refining this technique. Ultimately, the ability to gauge the effectiveness of machine learning models even in the absence of complete data will empower practitioners across industries, driving more confident and informed decision-making processes.

References

Jeffrey D Banfield and Adrian E Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, 49: 803–821, 1993. ISSN 0006341X, 15410420. doi: 10.2307/2532201. URL `http://www.jstor.org/stable/2532201`.

Philippe Bertrand, Jacques Bénichou, Philippe Grenier, and Claude Chastang. Hui and walter's latent-class reference-free approach may be more useful in assessing agreement than diagnostic performance. *Journal of Clinical Epidemiology*, 58:688–700, 7 2005. ISSN 08954356. doi: 10.1016/j.jclinepi.2004.10.021.

Francisco Blasques, Paolo Gorgi, Siem Jan Koopman, and Olivier Wintenberger. Feasible invertibility conditions and maximum likelihood estimation for observation-driven models. *Electronic Journal of Statistics*, 12:1019–1052, 2018. ISSN 19357524. doi: 10.1214/18-EJS1416.

Jim Davis. Posterior adaptation with new priors. 7 2020. URL `http://arxiv.org/abs/2007.01386`.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

David B. Dunson and Chuanhua Xing. Nonparametric bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, 104:1042–1051, 2009. ISSN 01621459. doi: 10.1198/jasa.2009.tm08439.

Christophe Dupuy, Francis Bach, and Francis Bach@ens Fr. Online but accurate inference for latent variable models with local gibbs sampling, 2017. URL http://jmlr.org/papers/v18/16-374.html.

Claes Enùe, Marios P Georgiadis, and Wesley O Johnson. Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown.

Vikas Garg and Adam T Kalai. Supervising unsupervised learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/72e6d3238361fe70f22fb0ac624a7072-Paper.pdf.

Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, 1984. doi: 10.1109/TPAMI.1984.4767596.

Heitor Murilo Gomes, Jesse Read, Albert Bifet, Jean Paul Barddal, and João Gama. Machine learning for streaming data: State of the art, challenges, and opportunities. *SIGKDD Explor. Newsl.*, 21(2):6–22, nov 2019. ISSN 1931-0145. doi: 10.1145/3373464.3373470. URL https://doi.org/10.1145/3373464.3373470.

Leo A Goodman. Exploratory latent structure analysis using both identifiable and unidentifiable models, 1974. URL https://about.jstor.org/terms.

Ghassan Hamra, Richard MacLehose, and David Richardson. Markov Chain Monte Carlo: an introduction for epidemiologists. *International Journal of Epidemiology*, 42(2):627–634, 04 2013. ISSN 0300-5771. doi: 10.1093/ije/dyt043. URL https://doi.org/10.1093/ije/dyt043.

S L Hui and S D Walter. Estimating the error rates of diagnostic tests, 1980. URL https://about.jstor.org/terms.

Siu L Hui and Xiao H Zhou. Evaluation of diagnostic tests without gold standards. *Statistical Methods in Medical Research*, 7(4):354–370, 1998. doi: 10.1177/096228029800700404. URL https://doi.org/10.1177/096228029800700404. PMID: 9871952.

Wesley O Johnson, Joseph L Gastwirth, and Larry M Pearson. Screening without a "gold standard": The hui-walter paradigm revisited, 2001. URL https://academic.oup.com/aje/article/153/9/921/124729.

Yongdai Kim, Minwoo Chae, Kuhwan Jeong, Byungyup Kang, and Hyoju Chung. An online gibbs sampler algorithm for hierarchical dirichlet processes prior. In Paolo Frasconi, Niels Landwehr, Giuseppe Manco, and Jilles Vreeken (eds.), *Machine Learning and Knowledge Discovery in Databases*, pp. 509–523, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46128-1.

Laurens Van Der Maaten and Geoffrey Hinton. Visualizing data using t-sne, 2008.

Jacob Montiel, Max Halford, Saulo Martiello Mastelini, Geoffrey Bolmier, Raphael Sourty, Robin Vaysse, Adil Zouitine, Heitor Murilo Gomes, Jesse Read, Talel Abdessalem, and Albert Bifet. River: machine learning for streaming data in python. 12 2020. URL http://arxiv.org/abs/2012.04740.

Stefan V. Pantazi, Yuri Kagolovsky, and Jochen R. Moehr. Cluster analysis of wisconsin breast cancer dataset using self-organizing maps. *Studies in health technology and informatics*, 90:431–6, 2002.

Frederico Z. Poleto, Julio M. Singer, and Carlos Daniel Paulino. A product-multinomial framework for categorical data analysis with missing responses. *Brazilian Journal of Probability and Statistics*, 28:109–139, 2014. ISSN 01030752. doi: 10.1214/12-BJPS198.

William M Rand. Objective criteria for the evaluation of clustering methods, 1971.

Amit G. Singal, Peter D.R. Higgins, and Akbar K. Waljee. A primer on effectiveness and efficacy trials. *Clinical and Translational Gastroenterology*, 5, 1 2014. ISSN 2155384X. doi: 10.1038/ctg.2013.13.

Aleksandra B Slavkovi´c, Slavkovi´c Stephen, E Fienberg, Paolo Gibilisco, Eva Riccomagno, Maria Piera Rogantin, and Henry P Wynn. *Algebraic and Geometric Methods in Statistics*. Cambridge University Press, 2009. doi: 10.1017/CBO9780511642401.