EPJ Web of Conferences **214**, 06004 (2019) https://doi.org/10.1051/epjconf/201921406004
CHEP 2018

Binary classifier metrics for optimizing HEP event selection

Andrea Valassi[1,*]

[1]CERN, Information Technology Department, CH-1211 Geneva 23, Switzerland

Abstract. I discuss the choice of evaluation metrics for binary classifiers in High Energy Physics (HEP) event selection and I point out that the Area Under the ROC Curve (AUC) is of limited relevance in this context, after discussing its use in other domains. I propose new metrics based on Fisher information, which can be used for both the evaluation and training of HEP event selection algorithms in statistically limited measurements of a parameter.

1 Introduction: standard practices for binary classifier evaluation

The training and evaluation of binary classifiers in a given domain are often performed using solutions initially proposed in another domain to solve very different problems. The main idea of this paper is that, instead, problem-specific tools and metrics are often needed. This is the case, in particular, of event selection in High Energy Physics, which is often evaluated using the Area Under the ROC Curve (AUC), even if this is of limited relevance in this context. This paper discusses these issues and has the following outline. The goals and practices for classifier evaluation in Medical Diagnostics and Information Retrieval are reviewed in this Sec. 1 and are then compared in Sec. 2 to the specific challenges of HEP event selection. New metrics based on Fisher information are proposed for parameter estimation problems, in Sec. 3 for evaluating classifiers and in Sec. 4 for their training. Conclusions are given in Sec. 5. I stress that my discussion of binary classifiers in HEP is limited to event selection, even if these tools are also used in HEP for other purposes, such as particle identification.

Discrete binary classifiers are algorithms that map instances of a data sample to one of two classes, signal (or positive) and background (or negative). They are commonly evaluated in terms of a 2x2 "confusion matrix": diagonal elements, known as True Positives (TP) and True Negatives (TN), represent correct decisions, while off-diagonal elements, known as False Positives (FP) and False Negatives (FN), represent incorrect decisions. In HEP event selection, TP and FP are selected signal ($S_{\rm sel}$) and background events ($B_{\rm sel}$), and FN and TN are rejected signal ($S_{\rm rej}$) and background events ($B_{\rm rej}$). In Medical Diagnostics (MD), TP and FP are ill and healthy patients diagnosed as ill, FN and TN are ill and healthy patients diagnosed as healthy. In Information Retrieval (IR), TP and FP are relevant and irrelevant documents retrieved in a query, FN and TN are relevant and irrelevant documents that are not retrieved.

Scoring classifiers differ from discrete classifiers in that they assign to each instance a "score" $\mathcal{D}$, in such a way that instances with a higher score have a higher likelihood to be signal than those with a lower score. A discrete classifier is often obtained by choosing an "operating point" of a scoring classifier, i.e. by choosing a threshold $\mathcal{D}_{\rm thr}$ such that all instances with $\mathcal{D} \geq \mathcal{D}_{\rm thr}$ are classified as signal and all others as background. Scoring classifiers are generally evaluated using either ROC or PRC curves, both of which consist in plotting against each other two metrics defined as ratios of specific elements of the confusion matrix. In both approaches, several techniques also exist for choosing the optimal operating point of the classifier and for summarising its performance using a single-valued "scalar" metric.

*e-mail: andrea.valassi@cern.ch

ROC curves in Medical Diagnostics and Machine Learning

ROCs are the standard tool in Medical Diagnostics. Using HEP terminology (due to space constraints), they are plots of signal efficiency $\epsilon_s = S_{\rm sel}/(S_{\rm sel}+S_{\rm rej})$ as a function of background efficiency $\epsilon_b = B_{\rm sel}/(B_{\rm sel}+B_{\rm rej})$. The ROC was originally introduced for radar applications of signal detection theory in the 1940's. Its use was then extended to psychophysics, i.e. visual and acoustic perception by human observers, and from there to medical imaging and medical diagnostics. It later became a standard tool also in Machine Learning (ML) research [1].

The ROC curve connects the lower-left corner where $(\epsilon_b, \epsilon_s) = (0, 0)$ to the top-right corner $(\epsilon_b, \epsilon_s) = (1, 1)$. One of its obvious features is that it is independent of disease prevalence, i.e. of the ratio $\pi_s = S_{\rm tot}/(S_{\rm tot}+B_{\rm tot})$ between the total number of signal events $S_{\rm tot} = S_{\rm sel}+S_{\rm rej}$ and that of signal plus background events $N_{\rm tot} = S_{\rm tot}+B_{\rm tot}$ (where $B_{\rm tot} = B_{\rm sel}+B_{\rm rej}$) before selection.

The ROC also possesses another remarkable feature, namely the surface below it is convex, i.e. its slope $\ell = d\epsilon_s/d\epsilon_b$ is monotonically decreasing. This is because ℓ is the likelihood ratio between the signal and background hypotheses for the event, and by construction the score $\mathcal{D}$ is a monotonically increasing function of ℓ. Experimental ROCs derived from validation data sets are actually not convex, as they are made up of a sequence of horizontal and vertical steps, but they can be transformed into convex ROCs by considering their "convex hull". The fact that ROCs are convex is very important if the optimal operating point is chosen using the classic cost-benefit analysis described in MD and ML research, where constant benefits $V_{\rm TP}$ and $V_{\rm TN}$ are attributed to each correct decision (TP and TN) and constants costs $K_{\rm FP}$ and $K_{\rm FN}$ to each mistake (FP and FN): the maximum value of the average benefit over all $N_{\rm tot}$ decision is, in fact, the point where the slope $d\epsilon_s/d\epsilon_b$ equals $(1\text{-}\pi_s)/\pi_s(V_{\rm TN}+K_{\rm FP})/(V_{\rm TP}+K_{\rm FN})$, which allows an easy geometrical interpretation of the points on the ROC plane in terms of "iso-benefit" lines, once the prevalence π_s and the cost-benefit matrix are known [1].

The main reason for the popularity of ROCs in MD (and ML), however, is not their convex shape, but rather the fact that ROCs describe diagnostic accuracy independently of disease prevalence π_s and of the choice of an operating point. This is important because physicians may choose a threshold $\mathcal{D}_{\rm thr}$ using different cost-benefit criteria, and because prevalence may be unknown at the time of diagnosis. The Area Under the ROC Curve (AUC), in particular, emerged over time in both MD and ML as a better measure of accuracy than another popular metric, the overall fraction of correct decisions $\mathrm{ACC} = (\mathrm{TP}+\mathrm{TN})/N_{\rm tot} = (S_{\rm sel}+B_{\rm rej})/N_{\rm tot}$, for precisely these reasons. While ACC strongly depends on π_s and on the choice of an operating point, the AUC is independent of π_s and describes all possible operating points on the ROC. Another advantage of the AUC in MD is that it has a well defined and relevant interpretation, namely it can be interpreted as "the probability that a randomly chosen diseased subject is correctly ranked with greater suspicion than a randomly chosen non-diseased subject" [2].

In spite of their popularity, however, ROCs and AUCs have known limitations. The main problem with the AUC, both in medical diagnostics and in concrete applications of ML techniques in banking and commerce, is that it is misleading when comparing two ROCs that cross: the AUC is built as an integral that gives equal weights to all parts of the ROC curves, but in the end a binary classifier is generally used by choosing a specific operating point, and in this case what counts is which ROC provides the better performance in the region where the operating point is chosen [3]. Another issue that was recently pointed out in ML research is the fact that ROC analysis may not be appropriate and may lead to overly optimistic evaluations in problems involving highly imbalanced data sets, i.e. where prevalence is very low or very high, and PRC curves may give a more informative picture in this case [4].

PRC curves in Information Retrieval

PRC curves are the standard tool in Information Retrieval. They are plots of signal purity, or "precision", $\rho = S_{\rm sel}/(S_{\rm sel}+B_{\rm sel})$ as a function of signal efficiency ϵ_s, or "recall", the two metrics that have essentially been used since the beginning of research in quantitative IR evaluation. ROC analysis was also suggested and thoroughly investigated as an alternative approach to measure IR effectiveness, but never found general acceptance in the field [5].

A popular scalar metric in IR for unranked evaluation, i.e. one which can be computed from the confusion matrix alone at any given point on the PRC, is $F_1 = 2\epsilon_s\rho/(\epsilon_s+\rho)$. Traditional scalar metrics for ranked evaluation, which integrate the knowledge of the whole PRC, include Mean Average Precision (which is related to the area under the PRC or AUCPR), "precision at κ" and precision at a fixed recall ϵ_s. These two last metrics are popular because, while in MD choosing an operating point is a compromise between the benefit of correct diagnoses of both ill and healthy patients (TPs and TNs) and the cost of mistakes (FPs and FNs), in IR users typically decide upfront that they can only afford the effort to retrieve $\kappa = N_{\text{sel}}$ documents, or that they need to retrieve at least a fraction ϵ_s of all existing relevant documents. Discounted Cumulated Gain, $\text{DCG}[\kappa] = \sum_{i=1}^{\kappa} \frac{\text{G}[i]}{\min(1,\log_2 i)}$, is a more recent but already popular metric describing users' gain after retrieving κ documents, taking into account for each document its non-binary true relevance grade $\text{G}[i]$ and the order in which it was retrieved [6].

2 HEP event selection evaluation: a comparison to MD and IR

While "standard" metrics coming from other domains, like the AUC, are frequently used also for evaluating binary classifiers in HEP event selections, the main idea of this paper is that different metrics must be chosen for different problems that have different specific goals.

Event selection is used in HEP in various contexts. During data taking, trigger decisions maximize the volume of useful data that can be stored with limited computing resources [7]: this implies maximising signal purity $\rho = S_{\text{sel}}/N_{\text{sel}}$ while the total numbers of input and output events per unit time, N_{tot} and N_{sel}, are fixed (this is analogous to "precision at κ" in IR). During data analysis, event selections are optimized to minimize errors in parameter measurements and maximize exclusion and discovery potentials of hypothesis tests for new physics models. While parameter measurements and searches require different metrics, I give here some general comments that apply to both, comparing HEP specificities to those of IR and MD in various respects. In Sec. 3, I will then focus on statistically limited parameter measurements.

- *Qualitative imbalance.* In MD, healthy and ill patients are both relevant: a diagnostic test is evaluated on its ability to detect a disease in ill patients (TPs) and to rule it out in healthy patients (TNs) [8]. IR is based instead on the idea that a whole class of documents are truly "irrelevant". Irrelevant documents in IR, like background events in HEP, are just a nuisance: what counts is the number of those incorrectly selected (FPs), while the number of those rejected (TNs) is irrelevant. In my opinion, this is the main reason why classifiers are mainly evaluated using ϵ_s and ϵ_b in MD and using ϵ_s and ρ in IR and HEP. Note also that the ACC metric, popular in MD, is invariant under an inversion of positive and negative labels, while ϵ_s, ρ and F_1, popular in IR, are invariant under a change of the TN count [9].
- *Quantitative imbalance.* In MD, relatively balanced class distributions are the norm, even if high skews do exist, e.g. for rare diseases. Data sets with 1 malignant mammography in 150 have been considered as "highly-skewed" (and PRCs have been suggested as more appropriate than ROCs to analyze them) [4]. In IR, instead, it is normal for data sets to be extremely skewed, with less than 1 relevant document in 1000. HEP problems are even more imbalanced, e.g. only 1 in $\sim 10^9$ LHC events (before any preselection) is a Higgs.
- *Cost/benefit models.* Traditional ROC analysis in MD and ML assigns a value to each decision and maximises its average $(\text{TP}\cdot V_{\text{TP}} - \text{FN}\cdot K_{\text{FN}} + \text{TN}\cdot V_{\text{TN}} - \text{FP}\cdot K_{\text{FP}})/(\text{TP+FN+TN+FP})$. This model is not used in IR and HEP classifiers, whose performance does not depend on TN.
- *Non-binary categorization.* A strictly binary categorization as signal or background is used in most IR, MD and ML applications. Only a few metrics use non dichotomous evaluation, such as the Obuchowski measure in MD, example-dependent cost-sensitive classification in ML and graded relevance assessment (e.g. DCG) in IR [6, 10]. In HEP fits of a parameter θ, each signal event i may have a different sensitivity γ_i to it; this is discussed in Sec. 3.
- *Ranking and partitioning.* A very specific feature of HEP, compared to MD and IR, is the routine use of distribution fits. As discussed in Sec. 3, these can be evaluated using metrics

built as sums over different event partitions. Metrics based on the classifier's ranking and computed as sums over event partitions also exist in MD (AUC) and IR (AUCPR and DCG), but these are not appropriate to evaluate the performance of HEP distribution fits.

- *Scale invariance.* Another specific feature of HEP is the need to use different analysis tools and evaluation metrics depending on the scale $S_{\rm tot}$ of the problem. Searches for new physics operate in a regime where $S_{\rm tot}$ is very small and are often optimized by maximizing a metric like $\mathcal{P} = \epsilon_s/(\frac{3}{2} + \sqrt{B_{\rm sel}})$ [11]. With enough events $S_{\rm tot}$, measurements of model parameters of signal processes become possible; the statistical errors scale as $1/\sqrt{S_{\rm tot}}$ and can be minimized in terms of ϵ_s and ρ alone, as discussed in Sec. 3. As $S_{\rm tot}$ grows further, systematic errors become important and can no longer be ignored in classifier evaluation.

In summary, HEP event selection is more similar to IR than to MD, and it is not surprising that ρ and ϵ_s are commonly used for its evaluation. The irrelevance of the TN count for both searches and parameter measurements is by itself a signal that any metric that depends on TN, like the AUC [9], is of limited relevance. As in MD, in particular, the main limitation of the AUC is that it is misleading when comparing two classifiers whose ROC curves cross. HEP event selection, however, has also two specific features that clearly set it apart from other domains, namely different event-by-event sensitivities and the routine use of distribution fits: this implies the need to develop HEP-specific evaluation metrics, as discussed in Sec. 3.

3 Statistical errors in parameter estimation and Fisher information

In this paper, I focus on HEP measurements of a single parameter θ, where systematic errors are negligible with respect to the statistical error $\Delta\theta$. I propose that the evaluation and training of event selection classifiers in this case should be based on the minimization of $\Delta\theta$, or equivalently on the maximization of the Fisher information $\mathcal{I}_\theta = 1/\Delta\theta^2$ about θ [12]. I assume that the signal event distribution depends on θ, while the background distribution does not.

A simple example is the measurement of a total cross section σ_s in a counting experiment, where $\sigma_s^{\rm (meas)} = (M - B_{\rm sel})/\mathcal{L}\epsilon_s$ is derived from the observed count of selected events $M = N_{\rm sel}^{\rm (meas)}$ and the measured luminosity $\mathcal{L}$, while the expected number of selected background events $B_{\rm sel}$ and the signal efficiency ϵ_s are derived from MC simulations. Assuming ϵ_s is independent of σ_s, the expected statistical error $\Delta\sigma_s$ is $\Delta N_{\rm sel}/\mathcal{L}\epsilon_s$, where $\Delta M = \Delta N_{\rm sel} = \sqrt{N_{\rm sel}}$ is the square root of the expected count of selected events $N_{\rm sel} = S_{\rm sel} + B_{\rm sel} = S_{\rm sel}/\rho = \mathcal{L}\epsilon_s\sigma_s/\rho$. This leads to

$$\mathcal{I}_{\sigma_s} = \frac{1}{(\Delta\sigma_s)^2} = \frac{1}{{\sigma_s}^2}\,\epsilon_s\rho S_{\rm tot} = \frac{1}{{\sigma_s}^2}\left(\frac{S_{\rm sel}^2}{S_{\rm sel}+B_{\rm sel}}\right). \tag{1}$$

To minimize $\Delta\sigma_s$, one should then maximise the product $\epsilon_s\rho$. This is a well known strategy to optimize event selections in total cross section measurements by counting experiments [13]. Note that the knowledge of σ_s from previous measurements is needed to compute ρ.

If σ_s depends on another parameter θ, the counting experiment may be used to determine θ with a statistical error $\Delta\theta$ given by $\Delta N_{\rm sel} = \frac{\partial N_{\rm sel}}{\partial\theta}\Delta\theta = \sqrt{N_{\rm sel}}$. If ϵ_s is independent of θ, then $\frac{\partial N_{\rm sel}}{\partial\theta} = \epsilon_s\frac{\partial S_{\rm tot}}{\partial\theta}$. This leads to $\frac{1}{(\Delta\theta)^2} = (\frac{1}{S_{\rm tot}}\frac{\partial S_{\rm tot}}{\partial\theta})^2\epsilon_s\rho S_{\rm tot}$, where $(1/S_{\rm tot})(\partial S_{\rm tot}/\partial\theta) = (1/\sigma_s)(\partial\sigma_s/\partial\theta)$ indicates the sensitivity of the total cross section σ_s to θ. This method was used for instance to determine the W mass at LEP2 by measuring the WW cross section at threshold [13].

The measurement of a total cross section σ_s or of a generic parameter θ, however, is most often performed in HEP using a distribution fit, rather than a counting experiment. I consider here the common case of a maximum likelihood fit on a binned histogram where the data are Poisson distributed. Binned fits rely on splitting all events into K disjoint partitions, or "bins", according to the value(s) of one (or more) variable(s) describing the properties of each event. A binned fit can be considered as the combination of K different measurements performed on these disjoint data sets, where θ is determined in each bin from the observed count of selected events $m_k = n_{{\rm sel},k}^{\rm (meas)}$, with an error $(\Delta\theta)_k$ derived from $\Delta n_{{\rm sel},k} = \frac{\partial n_{{\rm sel},k}}{\partial\theta}(\Delta\theta)_k = \sqrt{n_{{\rm sel},k}}$, the square root of the expected number of selected events $n_{{\rm sel},k}$. For statistically limited fits, these K

measurements are independent and a fortiori uncorrelated: the information about θ from the fit is thus equal to the sum of the individual contributions of the K bins,

$$\mathcal{I}_\theta = \frac{1}{(\Delta\theta)^2} = \sum_{k=1}^{K} (\mathcal{I}_\theta)_k = \sum_{k=1}^{K} \frac{1}{(\Delta\theta)_k^2} = \sum_{k=1}^{K} \left(\frac{1}{n_{\mathrm{sel},k}} \frac{\partial n_{\mathrm{sel},k}}{\partial \theta} \right)^2 n_{\mathrm{sel},k} = \sum_{k=1}^{K} \tilde{\epsilon}_k \tilde{\rho}_k \left(\frac{1}{s_{\mathrm{tot},k}} \frac{\partial s_{\mathrm{tot},k}}{\partial \theta} \right)^2 s_{\mathrm{tot},k} \,, \quad (2)$$

where $s_{\mathrm{tot},k}$ is the expected signal event count before selection, $\tilde{\epsilon}_k = s_{\mathrm{sel},k}/s_{\mathrm{tot},k}$ the local signal efficiency (assumed independent of θ) and $\tilde{\rho}_k = s_{\mathrm{sel},k}/(s_{\mathrm{sel},k} + b_{\mathrm{sel},k})$ the local signal purity in bin k, and $\tilde{\phi}_k = (1/s_{\mathrm{tot},k})(\partial s_{\mathrm{tot},k}/\partial\theta)$ represents the local sensitivity of the signal distribution in bin k to the value of θ. The expression for $\mathcal{I}_\theta$ in terms of $n_{\mathrm{sel},k}$ and $\partial n_{\mathrm{sel},k}/\partial\theta$ in Eq. 2 can also be more formally derived as the variance of the Fisher score $\partial \log P(\mathbf{m},\theta)/\partial\theta$ from the joint probability $P(\mathbf{m},\theta) = \prod_{k=1}^{K} e^{-n_{\mathrm{sel},k}} n_{\mathrm{sel},k}^{m_k}/m_k$ for the observation of $\mathbf{m} = \{m_1, \ldots m_K\}$.

The optimal classifier is thus one whose local efficiencies $\tilde{\epsilon}_k$ and purities $\tilde{\rho}_k$ maximize $\mathcal{I}_\theta$ in Eq. 2. Note that the knowledge of θ from previous measurements is needed to compute $\tilde{\rho}_k$. A dimensionless scalar metric between 0 and 1, however, is a more practical tool to evaluate classifiers. In particular, I suggest to use the ratio of $\mathcal{I}_\theta$ to the information $\mathcal{I}_\theta^{(ideal)}$ which could be attained with an "ideal" classifier, achieving $\tilde{\epsilon}_k = 1$ and $\tilde{\rho}_k = 1$ in every bin k,

$$\mathrm{FIP} = \frac{\mathcal{I}_\theta}{\mathcal{I}_\theta^{(ideal)}} = \frac{\sum_{k=1}^{K} \tilde{\epsilon}_k \tilde{\rho}_k \left(\frac{1}{s_{\mathrm{tot},k}} \frac{\partial s_{\mathrm{tot},k}}{\partial \theta} \right)^2 s_{\mathrm{tot},k}}{\sum_{k=1}^{K} \left(\frac{1}{s_{\mathrm{tot},k}} \frac{\partial s_{\mathrm{tot},k}}{\partial \theta} \right)^2 s_{\mathrm{tot},k}} = \frac{\sum_{k=1}^{K} \tilde{\epsilon}_k \tilde{\rho}_k \tilde{\phi}_k^2 s_{\mathrm{tot},k}}{\sum_{k=1}^{K} \tilde{\phi}_k^2 s_{\mathrm{tot},k}} \,. \quad (3)$$

I named this metric FIP for Fisher Information Part, because it represents the fraction of theoretically available information that a classifier is able to retain. Equation 3 is interesting for several reasons: first, the classifier is evaluated in each bin in terms of signal purity and efficiency, while the number of True Negatives (rejected background events) and background efficiency are irrelevant; second, the knowledge of local purity and efficiency in each bin is needed to evaluate the classifier, while the global purity and efficiency for the overall selection are not enough; third, different signal events contribute in different ways to the evaluation of the classifier depending on $\tilde{\phi}_k$, the local sensitivity to θ of the bin k they belong to.

The expression in Eq. 3 allows classifier evaluation for a wide range of statistically limited binned fits of a parameter, but it can be simplified in many special cases. To start with, measuring σ_s or θ in a counting experiment is equivalent to a fit with a single bin. As $s_{\mathrm{tot},k}$ and $\tilde{\phi}_k^2$ for $k = 1$ cancel out in the ratio, the FIP metric in this case reduces to $\mathrm{FIP}_1 = \epsilon_s \rho$, which is simply the product of global purity and efficiency, previously discussed for Eq. 1.

FIP metric for a total cross section fit from the score distribution: FIP2

Another very special class of measurements includes binned fits of a parameter θ where the classifier score $\mathcal{D}$ is used as a partitioning variable. I consider the common case where all events are used and $\tilde{\epsilon}_k = 1$ in every bin k, because the classifier is used to partition events, rather than select them or reject them. It is thus no longer necessary to distinguish between total and selected event counts in each bin and I will use simpler symbols with no suffixes for signal ($s_k = s_{\mathrm{sel},k} = s_{\mathrm{tot},k}$), background ($b_k = b_{\mathrm{sel},k} = b_{\mathrm{tot},k}$) and their sum ($n_k = n_{\mathrm{sel},k} = n_{\mathrm{tot},k}$).

An important subcase, which is also an extremely common practice in HEP [14], is the fit of a total signal cross-section σ_s from the one-dimensional (1-D) distribution of the score $\mathcal{D}$. This has two important consequences. First, as long as a change in σ_s can be considered as a global rescaling of the $\mathcal{D}$ distribution for signal events, the sensitivity of this distribution to $\theta = \sigma_s$ is a constant across bins, $\tilde{\phi}_k = \frac{1}{s_k} \frac{\partial s_k}{\partial \sigma_s} = 1/\sigma_s$, and the FIP metric in Eq. 3 reduces to

$$\mathrm{FIP}_2 = \frac{\sum_{k=1}^{K} \tilde{\rho}_k s_k}{\sum_{k=1}^{K} s_k} = \frac{\sum_{k=1}^{K} s_k^2/n_k}{\sum_{k=1}^{K} s_k} = \frac{\sum_{k=1}^{K} n_k \tilde{\rho}_k^2}{\sum_{k=1}^{K} s_k} \,. \quad (4)$$

Second, fitting the 1-D distribution of $\mathcal{D}$ implies that every bin k in the fit effectively corresponds to a separate arc of the ROC curve, subtending two ranges $d\epsilon_s$ and $d\epsilon_b$ of signal and

background efficiency, such that $s_k = S_{\rm tot} d\epsilon_s$ and $b_k = B_{\rm tot} d\epsilon_b$. This implies that the local purity can be computed as $\tilde{\rho}(\epsilon_s) = S_{\rm tot} d\epsilon_s / (S_{\rm tot} d\epsilon_s + B_{\rm tot} d\epsilon_b) = 1/[1 + (d\epsilon_b/d\epsilon_s)(1-\pi_s)/\pi_s]$. FIP$_2$ can then be computed from the ROC alone, as long as prevalence π_s is also known:

$$\mathrm{FIP}_2 = \frac{\sum_{k=1}^{K} \tilde{\rho}_k s_k}{\sum_{k=1}^{K} s_k} = \frac{\int_0^1 \tilde{\rho}(\epsilon_s) S_{\rm tot} d\epsilon_s}{\int_0^1 S_{\rm tot} d\epsilon_s} = \int_0^1 \tilde{\rho}(\epsilon_s) d\epsilon_s = \int_0^1 \frac{d\epsilon_s}{1 + \left(\frac{1-\pi_s}{\pi_s}\right) \frac{d\epsilon_b}{d\epsilon_s}} \,. \tag{5}$$

Some numerical tests [15] that I performed using a Python implementation of Eq. 5 confirmed that FIP$_2$ correctly predicts the average statistical error $\Delta\sigma_s$ on a binned fit of σ_s from the distribution of the score $\mathcal{D}$. In my implementation, I computed the integral in Eq. 5 not on the native ROC obtained from a validation sample, but rather on its convex hull: a stepwise ROC, in fact, leads to an overestimate of FIP$_2$ because signal events contribute with an overestimated purity $\tilde{\rho} = 1$ in the vertical steps where $d\epsilon_s > 0$ and $d\epsilon_b = 0$, while the negative impact of background events is underestimated because they are confined to the horizontal steps where $d\epsilon_s = 0$ and $d\epsilon_b > 0$. It is in any case interesting to note that, given two convex ROCs, one of the two is guaranteed to have a higher FIP$_2$ value than the other if it has a higher ϵ_s for any value of ϵ_b; this can be easily proved using numerical and geometrical arguments.

Rather than from the ROC and prevalence, FIP$_2$ can also be derived from the PRC alone, as $\mathrm{FIP}_2 = \int_0^1 \frac{\rho\, d\epsilon_s}{1-(\epsilon_s/\rho)(d\rho/d\epsilon_s)}$, but this is less practical than Eq. 5, because both calculations only make sense on the ROC convex hull, which is easier to compute on the ROC than on the PRC. Note also that these expressions for FIP$_2$ are integrals over the ROC and PRC, just like AUC and AUCPR. Unlike AUC, however, FIP$_2$ is qualitatively and quantitatively relevant for minimizing $\Delta\sigma_s$ in a fit for σ_s, while AUC is irrelevant and misleading, amongst other reasons because it is independent of the prevalence π_s; it is easy to prepare an example [15] to show that, if the maximum AUC is used as the criterion to choose one of two classifiers, the worse classifier, leading to the larger error $\Delta\sigma_s$, is selected for some specific values of π_s. Note also that $\mathrm{FIP}_2 = \int \tilde{\rho} d\epsilon_s$ is an integral of local purity, while AUCPR$= \int \rho d\epsilon_s$ uses global purity.

4 Optimal partitioning and Fisher information

Maximizing the information in Eq. 2 for a binned fit of θ involves two distinct but related steps: choosing the optimal partitioning of events into K bins, i.e. choosing the distributions to fit and the bin size, and choosing the classifier providing the optimal efficiencies $\tilde{\epsilon}_k$ and purities $\tilde{\rho}_k$ in those bins. As discussed in Sec. 3, I consider the case where the score $\mathcal{D}$ is (one of) the partitioning variable(s), so that all events are used and $\tilde{\epsilon}_k = 1$ in each bin k.

The relevant question to ask for optimizing partitioning is under which circumstances there is a benefit in splitting the events in one bin $k = 0$ into two separate bins $k = 1, 2$, so that $n_0 = n_1 + n_2$ and $s_0 = s_1 + s_2$. From Eq. 2, the information "inflow" [12] in this split is given by

$$\Delta\mathcal{I}_\theta = \frac{1}{n_1}\left(\frac{\partial n_1}{\partial\theta}\right)^2 + \frac{1}{n_2}\left(\frac{\partial n_2}{\partial\theta}\right)^2 - \frac{1}{n_0}\left(\frac{\partial n_0}{\partial\theta}\right)^2 = \frac{(n_1 \frac{\partial n_2}{\partial\theta} - n_2 \frac{\partial n_1}{\partial\theta})^2}{n_1 n_2 n_0} = \frac{n_1 n_2}{n_0}\left(\tilde{\rho}_1 \frac{1}{s_1}\frac{\partial s_1}{\partial\theta} - \tilde{\rho}_2 \frac{1}{s_2}\frac{\partial s_2}{\partial\theta}\right)^2 . \tag{6}$$

Equation 6 clearly shows that the optimal partitioning strategy consists in separating events into bins with different values of $\tilde{\rho}_k \tilde{\phi}_k = \tilde{\rho}_k \frac{1}{s_k}\frac{\partial s_k}{\partial\theta}$. In other words, if $\tilde{\rho}$ and $\tilde{\phi}$ were observable properties of each event, an optimal strategy for measuring θ would consist in the fit of the 1-D distribution of $\tilde{\rho}\tilde{\phi}$, or alternatively the fit of the 2-D distribution of $\tilde{\rho}$ versus $\tilde{\phi}$. I note, in passing, that these are optimal variables for both binned and unbinned fits.

Neither $\tilde{\rho}$ nor $\tilde{\phi}$, however, are observable properties of an event. The question, then, is which partitioning variables should be used to ensure that different bins of their distributions contain events with different values of $\tilde{\rho}\tilde{\phi}$. When formulated in these terms, this is a clear example of a ML problem that can be solved by a regression algorithm: given the multi-dimensional space of event observables $\mathbf{x}$, the problem consists in finding the two functions of these observables whose distributions $\mathcal{R}_{\tilde{\rho}}(\mathbf{x})$ and $\mathcal{R}_{\tilde{\phi}}(\mathbf{x})$ "best" approximate $\tilde{\rho}(\mathbf{x})$ and $\tilde{\phi}(\mathbf{x})$,

or in other words maximally separate phase space regions with different relative abundances of signal and background, and signal events with different sensitivities to the parameter θ.

If a scoring classifier is used for event selection, the score $\mathcal{D}$ is often used instead of a regression predictor $\mathcal{R}_{\tilde{\rho}}$ for $\tilde{\rho}$, even if $\mathcal{D}$ is actually the output of a classification algorithm (the distinction is blurred). As discussed, total cross sections are often measured by fitting the $\mathcal{D}$ distribution; this is an optimal strategy, because $\mathcal{D}$ is "best" trained to separate regions with different values of $\tilde{\rho}$, while a predictor $\mathcal{R}_{\tilde{\phi}}$ for $\tilde{\phi}$ is not needed as $\tilde{\phi}(\mathbf{x}) = 1/\sigma_s$ is a constant.

For fits of a generic parameter θ like a particle mass or coupling, one or more kinematic observables (such as invariant masses or transverse momenta) are generally used to separate events with different values of $\tilde{\phi}$, rather than building a regression predictor $\mathcal{R}_{\tilde{\phi}}$ for $\tilde{\phi}$. In the "optimal observable" method [16], specific functions of the observed event kinematics that would provide optimal separation if computed from the true event kinematics are used, but the separation power of these observables is degraded by the finite dector resolution. Another approach that I would suggest to try out, but I have not yet worked on concretely, consists in storing the event-by-event MC weight derivative $\gamma_i = \frac{1}{|\mathcal{M}_i|^2}\frac{\partial |\mathcal{M}_i|^2}{\partial \theta}$ for a sample of unweighted simulated signal events, and training a predictor $\mathcal{R}_{\tilde{\phi}}(\mathbf{x})$ against γ_i so that it would "best" predict $\tilde{\phi}(\mathbf{x})$. It is easy to see, in fact, that the bin-by-bin sensitivity $\tilde{\phi}_k = \frac{1}{s_k}\frac{\partial s_k}{\partial \theta}$ is simply the average $\langle\gamma\rangle_k$ of the event-by-event sensitivities γ_i in bin k. A 2018 paper by Brehmer et al. [17] has suggested a similar approach using likelihood gradients. Note that background contamination effectively decreases the bin-by-sensitivity to $\tilde{\rho}_k\tilde{\phi}_k$ because it dilutes s_k signal events of average sensitivity $\tilde{\phi}_k$ with b_k background events whose sensitivity is 0.

Fisher information for training classification and regression trees

A partitioning strategy using a classifier $\mathcal{D}$ to predict $\tilde{\rho}$ and a predictor $\mathcal{R}_{\tilde{\phi}}$ for $\tilde{\phi}$, however, is only truly optimal to minimize the measurement error on θ if these algorithms are "best" trained according to a metric that is relevant to $\Delta\theta$. I focus on Decision Tree (DT) algorithms [18], which are commonly used in HEP [14] and are the natural companions of binned fits as both describe distributions in terms of disjoint event partitions, the "bins" of a fit and the "nodes" of a tree. DT algorithms proceed by iteratively splitting nodes into two sub-nodes. Given a node with n_0 events, amongst all possible splits into two nodes with n_1 and n_2 events, that with the highest loss $-\Delta\mathcal{H} = n_0 h_0 - n_1 h_1 - n_2 h_2$ in the impurity $\mathcal{H} = \sum_k n_k h_k$ is chosen, and the process stops when the loss -$\Delta\mathcal{H}$ is below a threshold, and/or other conditions are met. The impurity h_k of a node k would generally be computed as a function of the local purity $\tilde{\rho}_k$ using one of two criteria in classification trees, Gini diversity $\mathcal{H}_{\text{Gini}} = \sum_k 2n_k\tilde{\rho}_k(1-\tilde{\rho}_k)$ or Shannon information entropy $\mathcal{H}_{\text{entropy}} = \sum_k n_k[\tilde{\rho}_k \log_2 \tilde{\rho}_k + (1-\tilde{\rho}_k)\log_2(1-\tilde{\rho}_k)]$, and using the mean squared error $\mathcal{H}_{\text{MSE}} = \sum_k n_k[\frac{1}{n_k}\sum_{i\in k}(\gamma_i - \langle\gamma\rangle_k)^2]$ in a regression tree for $\tilde{\phi} = \langle\gamma\rangle$. My proposal is to use Fisher information in both cases, $\mathcal{H}_{\text{Fisher}} = -\mathcal{I}_\theta = -\sum_k n_k\tilde{\rho}_k^2\tilde{\phi}_k^2$. Maximising the impurity loss -$\Delta\mathcal{H}$ would thus be equivalent to minimising the Fisher information gain $\Delta\mathcal{I}_\theta$ in Eq. 6.

In practice, an approach to optimize a measurement of θ could be the following: first, train a regression tree $\mathcal{R}_{\tilde{\phi}}$ for $\tilde{\phi}$ on signal MC events, against γ_i; then, train a classification tree $\mathcal{D}$ (i.e. a regression tree for $\tilde{\rho}$) on both signal and background MC, using the predictor $\mathcal{R}_{\tilde{\phi}}$ in the loss function; finally, perform a 2-D binned fit on $\mathcal{D}$ and on the predictor $\mathcal{R}_{\tilde{\phi}}$.

With respect to the Gini and entropy criteria for classification, $\mathcal{H}_{\text{Fisher}}$ has two main differences: first, it is asymmetric in $\tilde{\rho}$, e.g. pure signal nodes ($\tilde{\rho}_k = 1$) are very valuable, while pure background nodes ($\tilde{\rho}_k = 0$) are worthless; second, it weighs nodes according to their sensitivity $\tilde{\phi}$. Quite surprisingly, however, for problems like σ_s measurements where $\tilde{\phi}$ is a constant, it is easy to show that the Gini and Fisher splitting strategies are equivalent, because

$$\frac{-\Delta\mathcal{H}_{\text{Gini}}}{2} = -s_1\left(1-\frac{s_1}{n_1}\right) - s_2\left(1-\frac{s_2}{n_2}\right) + (s_1+s_2)\left(1-\frac{s_1+s_2}{n_1+n_2}\right) = \frac{(s_1 n_2 - s_2 n_1)^2}{n_1 n_2 (n_1+n_2)} = -\Delta\mathcal{H}_{\text{Fisher}}\,. \quad (7)$$

One advantage of the approach I suggest is that it makes it extremely easy to interpret and visualize the loss functions used for classifier training in terms of the metrics used for

its evaluation, because both aim at maximising the information $\mathcal{I}_\theta$. Taking the example of a classification tree, which is essentially a regression tree $\mathcal{D}$ for $\tilde{\rho}_k$ in my approach, this is best understood by plotting the different ROCs and the FIP_2 distribution for the same algorithm, using many different training and validation samples drawn randomly from the same toy model distribution [15]. While in the "ideal" case the ROC passes through $(\epsilon_b, \epsilon_s) = (0, 1)$ and FIP_2 equals 1, possessing the knowledge of the real multi-dimensional distribution would only allow reaching a "limit" ROC and a limit value of FIP_2. When testing a classifier against a toy model, the distribution of FIP_2 over the training sample would generally be higher than the limit, while on the validation sample it would generally be lower than the limit, with the distance between the two peaks representing a measure of overtraining in the algorithm.

5 Conclusions, outlook and acknowledgements

The main message of this paper is that binary classifiers must be evaluated using different metrics, specific to the goals of the problem to which they are applied, and that no single scalar metric exists which is appropriate to any classification problem in any domain. In particular, I pointed out that AUC is of limited relevance in HEP event selection and proposed new metrics based on Fisher information, which can be used for both the evaluation and training of event selection algorithms in statistically limited measurements of one parameter.

I thank L. Canali, M. Schulz, R. Schoefbeck, A. Read, A. Sciabà, D. Rousseau, D. Düllmann, D. Giordano, G. Lo Presti, G. Dissertori, L. Brenner, M.-O. Bettler, P. Manzano, P. Azzurri, P. Seyfert, R. Chierici, T. Keck, T. Dorigo and V. Ciulli for many useful discussions. I am grateful to my former colleagues in ALEPH, where many of the ideas I reported in this article originated. I thank S. Gleyzer and the anonymous referee for useful feedback on my initial submission. Plots and further details on this work are available in the slides of the CHEP 2018 talk [15] described in this paper. This research will be discussed in more depth in an upcoming longer paper, including more complete references and acknowledgements.

References

[1] W. W. Peterson, T. G. Birdsall, Un. Michigan Tech. Report 13 (1953); W. P. Tanner, J. A. Swets, Psych. Rev. 61 (1954) 401; L. B. Lusted, Science 171 (1971) 1217; C. E. Metz et al., Radiology 109 (1973) 297; F. J. Provost, T. Fawcett, Proc. KDD-97.

[2] D. M. Green, J. Ac. Soc. Am. 36 (1964) 1042; D. J. Goodenough et al., Radiology 105 (1972) 199; J. A. Hanley, B. J. McNeil, Radiology 143 (1982) 29; J. A. Swets, Inv. Radiology 14 (1979) 109, Science 240 (1988) 1285; A. P. Bradley, Pat. Rec. 30 (1997) 1145.

[3] X. H. Zhou et al., *Statistical Methods in Diagnostic Medicine*, Wiley (2002); N. M. Adams, D. J. Hand, Pat. Rec. 32 (1999) 1139; C. Drummond, R. C. Holte, Proc. KDD-00.

[4] J. Davis et al., Proc. IJCAI-05; J. Davis, M. Goadrich, Proc. ICML-06; T. Saito, M. Rehmsmeier, PLoS One 10 (2015) e0118432; H. He, E. A. Garcia, IEEE Trans. Knowl. Data Eng. 21 (2009) 1263.

[5] C. W. Cleverdon, *Cranfield report* (1962); J. A. Swets, Science 141 (1963) 245; C. J. van Rijsbergen, *Information retrieval*, Butterworths (1979); D. Hull, Proc. SIGIR 1993.

[6] K. Järvelin, J. Kekäläinen, Proc. SIGIR 2000.

[7] J. Albrecht and LHCb Coll., J. Phys. Conf. Ser. 623 (2015) 012003.

[8] B. J. McNeil et al., N. Engl. J. Med. 293 (1975) 211.

[9] M. Sokolova, G. Lapalme, IPM 45 (2009) 427; A. Luque et al., Symmetry 11 (2019) 47.

[10] N. A. Obuchowski, Stat. Med. 25 (2006) 481; B. Zadrozny, C. Elkan, Proc. KDD-01.

[11] G. Punzi, Proc. PhyStat2003; LHCb Coll., Phys. Rev. D 97 (2018) 032010.

[12] A. van den Bos, *Parameter Estimation for Scientists and Engineers*, Wiley (2007).

[13] D. Gelé et al., Proc. LEP2 Workshop (1996); OPAL Coll., Eur. Phys. J. C 1 (1998) 395.

[14] D0 Coll., Phys. Rev. D 78 (2008) 012005.

[15] A. Valassi, CHEP 2018 presentation slides, doi:10.5281/zenodo.1303387.

[16] D. Atwood, A. Soni, Phys. Rev. D 45 (1992) 2405; M. Davier et al., Phys. Lett. B 306 (1993) 411; M. Diehl, O. Nachtmann, Z. Phys. C 62 (1994) 397.

[17] J. Brehmer et al., Phys. Rev. Lett. 121 (2018) 111801.

[18] L. Breiman et al., *Classification And Regression Trees*, Chapman and Hall (1984).