Fisher information metrics for binary classifier evaluation and training

Event selection for HEP precision measurements

Andrea Valassi
(CERN IT-DI-LCG)

CHEP 2018, Sofia – Machine Learning and Physics Analysis session

Why and when I got interested in this topic

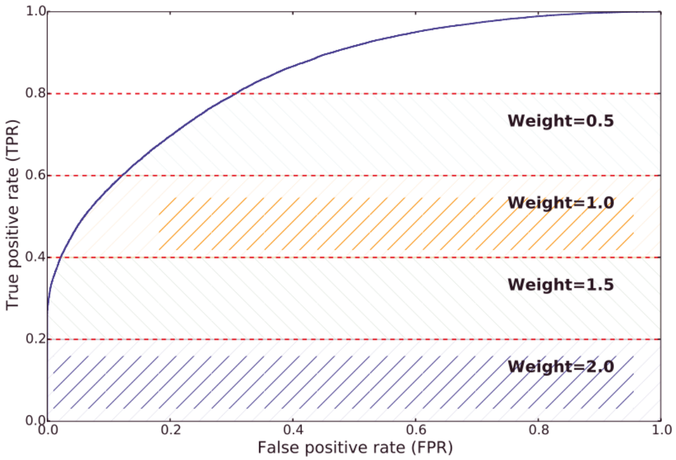


Figure 3: Weights assigned to the different segments of the ROC curve for the purpose of submission evaluation. The x axis is the False Positive Rate (FPR), while the y axis is True Positive Rate (TPR).

T. Blake at al., *Flavours of Physics: the machine learning challenge for the search of $\tau \to \mu\mu\mu$ decays at LHCb* (2015, unpublished). https://kaggle2.blob.core.windows.net/competitions/kaggle/4488/media/lhcb_description_official.pdf (accessed 15 January 2018)

The 2015 LHCb Kaggle ML Challenge:
- Develop an event selection in a search for $\tau \to \mu\mu\mu$

ML binary classifier problem
- *Evaluation: the highest weighted AUC is the winner*

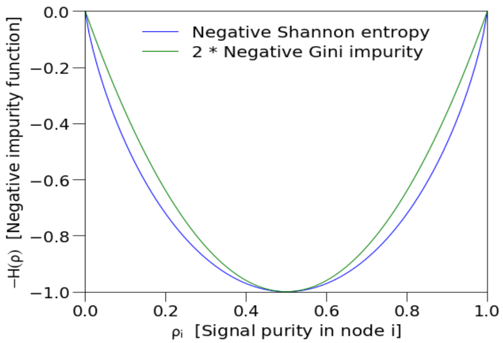
- First time I saw an *Area Under the Roc Curve (AUC)*
- My reaction:
  – What is the AUC? Which other scientific domains use it and why?
  – *Is the AUC relevant in HEP? Can we develop HEP-specific metrics?*

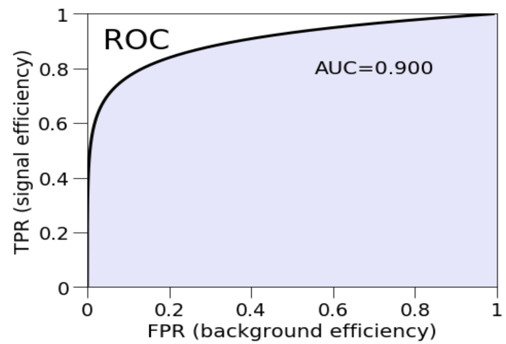
Overview – the scope of this talk

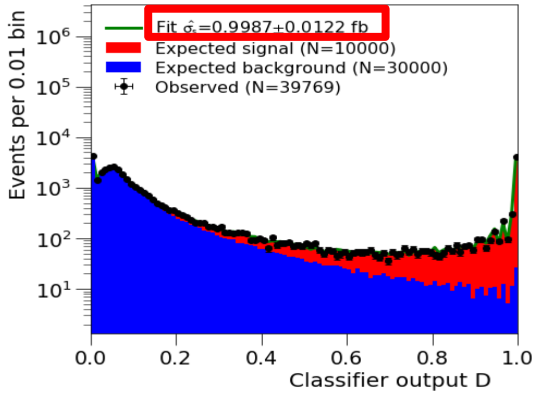
- *Different domains and/or problems → Need different metrics*
  - Always keep your final goal in mind
- Focus on a specific HEP example: *event selection to minimize statistical error $\Delta\theta$ in an analysis for the point estimation of θ*
  - Do not focus on: tracking, systematic errors, trigger, searches…
- *Whenever you take a decision, base it on the minimization of $\Delta\theta$*
  - Metrics for physics precision → final goal: minimize $\Delta\theta$
  - Metrics for binary classifier evaluation → (is the AUC relevant?)
  - Metrics for binary classifier training → (are standard ML metrics relevant?)

Training, Evaluation, Physics: one metric to bind them all?

Example: event selection using a Decision Tree for a parameter fit







TRAINING

- (either) **Gini impurity**
  - *Economics: inequality*
  - *Ecology: diversity*
- (or) **Shannon information**
  - *Information theory: entropy*

EVALUATION

- **ROC Curve** **(Receiver Operating Characteristic)**
  - *Signal detection: radar detection*
  - *Psychophysics: sensory detection*
- **AUC (Area Under the ROC Curve)**
  - *Radiology, Medicine: diagnostic accuracy*

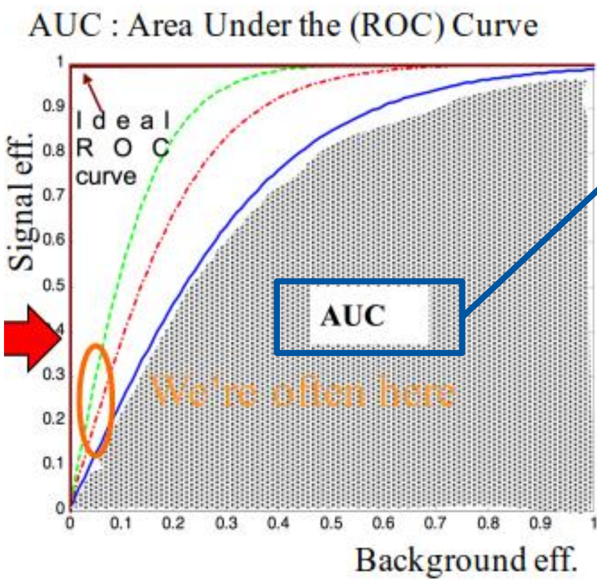
PHYSICS

- **Precision**
  - *Parameter estimation: measurement error $\Delta\theta$*

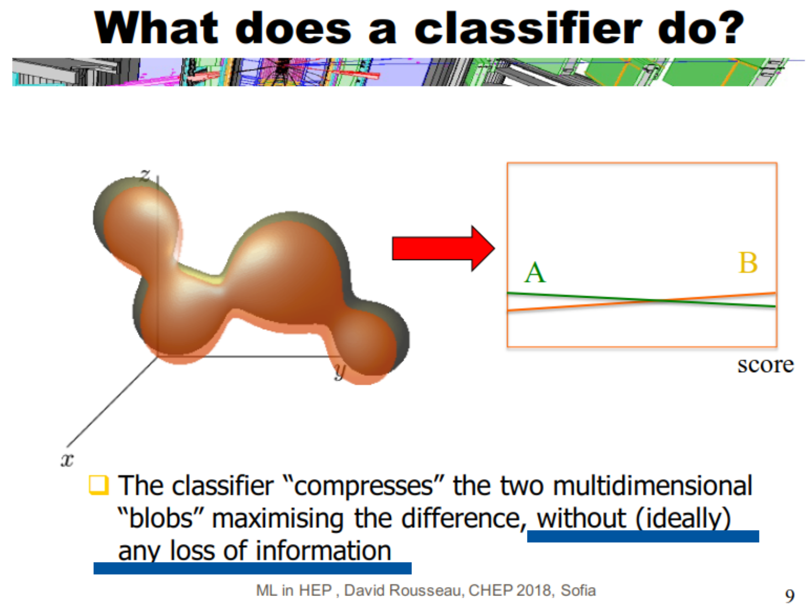
Proposal: use metrics based on Fisher Information in all three steps

(Fisher Information about θ ~is $I_\theta = 1/(\Delta\theta)^2$ – maximize I_θ to minimize $\Delta\theta$)

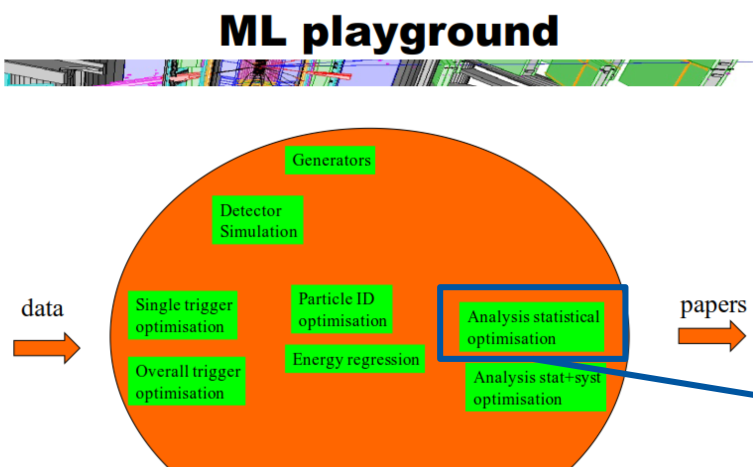
CHEP plenary this morning



I will argue against AUC's for evaluation in HEP



I will discuss the retention of *Fisher* information in classifiers



I will describe one problem in analysis statistical optimization

Binary classifier evaluation – reminder

Discrete classifiers: the confusion matrix

Binary decision:
signal or background

	true class: **P**ositives (HEP: **signal Stot**)	***true class***: **N**egatives (HEP: **background Btot**)
classified as: positives (HEP: **selected**)	**True Positives (TP)** (HEP: selected signal **Ssel**)	**False Positives (FP)** (HEP: selected bkg **Bsel**)
classified as: negatives (HEP: **rejected**)	**False Negatives (FN)** (HEP: rejected signal **Srej**)	**True Negatives (TN)** (HEP: rejected bkg **Brej**)

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$$

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

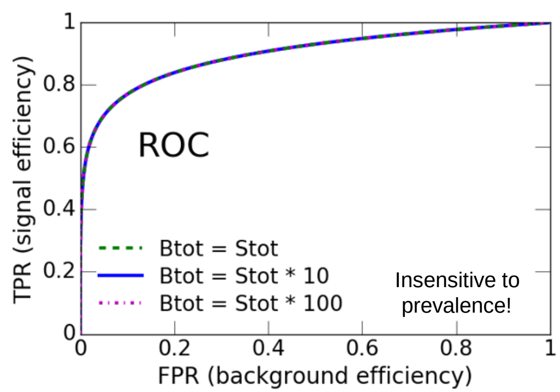
$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} = 1-\mathrm{FPR}$$

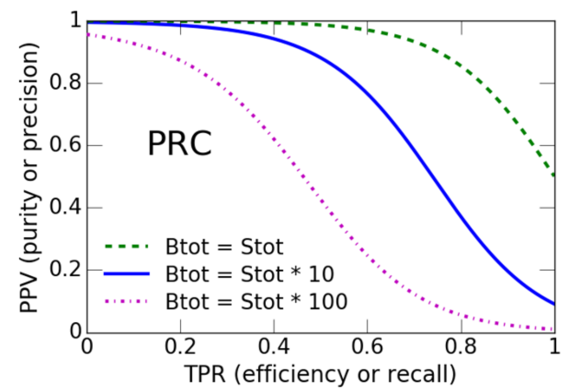
Prevalence $\pi_s = \frac{S_{\mathrm{tot}}}{S_{\mathrm{tot}}+B_{\mathrm{tot}}}$

Scoring classifiers: ROC and PRC curves

Continuous output:
probability to be signal

Vary the binary decision
by varying the cut
on the scoring classifier





Binary classifier evaluation in other domains

Medical Diagnostics (MD) $\rightarrow$ e.g. diagnostic accuracy for cancer

– *Symmetric: all patients important, both truly ill (TP) and truly healthy (TN)*
– ROC-based analysis (because ROC insensitive to prevalence)
  - *AUC interpretation: probability that diagnosis gives greater suspicion to a randomly chosen sick subject than to a randomly chosen healthy subject*

Information Retrieval (IR) $\rightarrow$ e.g. find pages in Google search

– *Asymmetric: distinction between relevant and non-relevant documents*
– PRC-based evaluation: precision and recall (= purity and efficiency in HEP)
  - Single metric: e.g. Mean Average Precision ~ area under PRC (AUCPR)

Oversimplification: $\boxed{\mathrm{AUC} = \int_0^1 \epsilon_s d\epsilon_b = 1 - \int_0^1 \epsilon_b d\epsilon_s}$ (MD) vs. (IR) $\boxed{\mathrm{AUCPR} = \int_0^1 \rho \, d\epsilon_s}$

Evaluation: (main) specificities of HEP

1. Qualitative asymmetry: signal interesting, background irrelevant
   – Like Information Retrieval: use purity and efficiency (precision and recall)
     • *True Negatives and the AUC are irrelevant in HEP event selection*

2. Distribution fits: several disjoint bins, not just a global selection
   – Analyze *local signal efficiency and purity in each bin*, not just global ones
   – Frequent special case: fits involving distributions of the scoring classifier

3. Signal events not all equal: they may have different sensitivities
   – Example: only events close to a mass peak are sensitive to the mass

Illustrated in the following by three examples (1=FIP1, 1+2=FIP2, 1+2+3=FIP3)

Evaluation: Fisher Information Part (FIP)

- Evaluation of an event selection from its effect on the error $\Delta\hat{\theta}$
  - *Compare to "ideal" case where there is no background*

- FIP: fraction of "ideal" FI that is retained by the real classifier
  - Range in [0,1] $\rightarrow$ 0 if no signal, 1 if select all signal and no background
  - Qualitatively relevant: higher is better $\rightarrow$ maximize FIP to minimize $\Delta\hat{\theta}$
  - Numerically meaningful: related to $\Delta\hat{\theta}$

- For a binned fit of θ from a (1-D or multi-D) histogram:
  - Consider only statistical errors $\rightarrow$ sum information from the different bins

$$\mathrm{FIP} = \frac{\mathcal{I}_\theta^{(\mathrm{real\ \ classifier})}}{\mathcal{I}_\theta^{(\mathrm{ideal\ \ classifier})}} = \frac{\sum_{i=1}^{m} \epsilon_i \rho_i \times \frac{1}{S_i}\left(\frac{\partial S_i}{\partial \theta}\right)^2}{\sum_{i=1}^{m} \frac{1}{S_i}\left(\frac{\partial S_i}{\partial \theta}\right)^2}$$

Remember from the previous slide:
1. Qualitative asymmetry: use ε and ρ (as in IR)
2. Distribution fit: need local ε_i and ρ_i in each bin
3. Signal events not all equal: need sensitivity $\frac{1}{S_i}\frac{\partial S_i}{\partial \theta}$

[FIP1] Cross-section in counting experiment

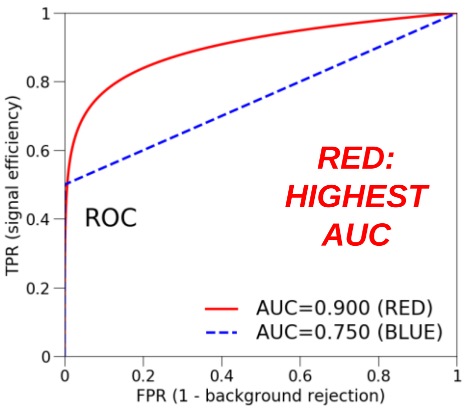
- Counting experiment: measure a single number N_{meas}
  - Well-known since decades: maximize $\varepsilon_s * \rho$ to minimize statistical errors

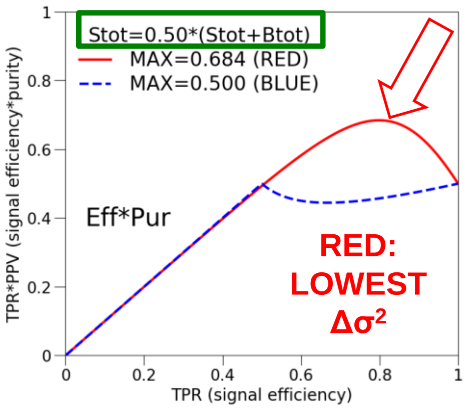
- FIP special case:
  - Counting experiment (1 bin) → *global* signal efficiency and purity
  - *Cross-section fit $\theta=\sigma_s$ → all events have equal sensitivity* $\frac{1}{S_i}\frac{\partial S_i}{\partial \sigma_s} = \frac{1}{\sigma_s}$

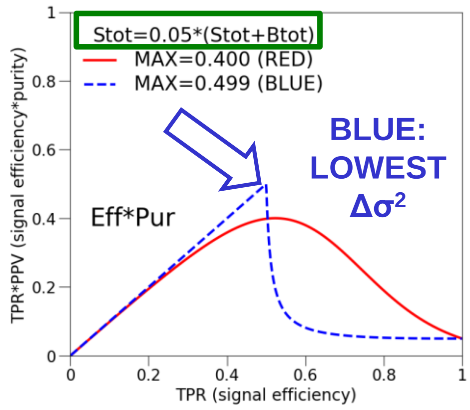
$$\mathrm{FIP} = \frac{\mathcal{I}_\theta^{(\text{real classifier})}}{\mathcal{I}_\theta^{(\text{ideal classifier})}} = \frac{\sum_{i=1}^{m} \epsilon_i \rho_i \times \frac{1}{S_i}\left(\frac{\partial S_i}{\partial \theta}\right)^2}{\sum_{i=1}^{m} \frac{1}{S_i}\left(\frac{\partial S_i}{\partial \theta}\right)^2} \rightarrow \boxed{\mathrm{FIP1} = \varepsilon_s * \rho}$$

Examples of issues in AUCs – *crossing ROCs*

- Cross-section measurement by counting experiment
  - Maximize FIP1=ε_s*ρ → Minimize the statistical error $\Delta\sigma^2$
- Compare two classifiers: red (AUC=0.90) and blue (AUC=0.75)
  - The red and blue ROCs cross (otherwise the choice would be obvious!)
- Choice of classifier achieving minimum $\Delta\sigma^2$ *depends on S_{tot}/B_{tot}*
  - *Signal prevalence 50%*: choose classifier with higher AUC (red)
  - *Signal prevalence 5%*: choose classifier with lower AUC (blue)
  - AUC is irrelevant – and ROC is only useful if you also know prevalence







	FIP1	AUC
Range in [0,1]	YES	YES
Higher is better	YES	NO
Numerically meanigful	YES	NO

Optimal partitioning in distribution fits

- Does information I_θ increase if I split a bin into two ($n \rightarrow n_L+n_R$)?
  - Information gain is $\Delta I_\theta = \left(\rho_L \frac{1}{s_L}\frac{\partial s_L}{\partial\theta} - \rho_R \frac{1}{s_R}\frac{\partial s_R}{\partial\theta}\right)^2 * \frac{n_L n_R}{n_L+n_R}$

- Partition events using optimal binning variables (→ two examples)
  - For cross-sections $\left(\frac{1}{s_i}\frac{\partial s_i}{\partial\sigma_s} = \frac{1}{\sigma_s}\right)$: *separate bins with different ρ_i* (→"FIP2")
  - For a generic parameter θ : *separate bins with different $\rho_i \frac{1}{s_i}\frac{\partial s_i}{\partial\theta}$* (→"FIP3")

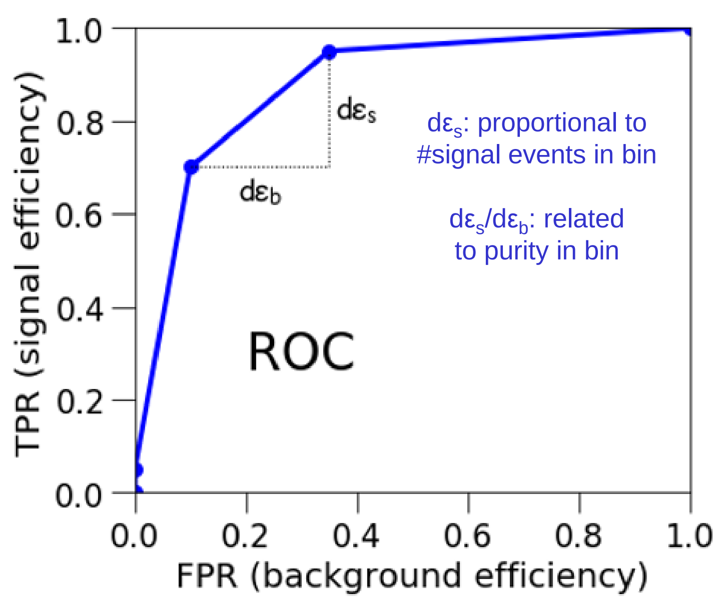
- Practical ML consequences (focus on cross-section example):
  - *Use the scoring classifier (i.e. ~ρ !) to partition events, not to reject them*
  - *Train the scoring classifier to maximize the total Fisher information* of the histogram binning, i.e. train it to maximize its partitioning power
    - Use Fisher Information as a node splitting criterion for decision tree training
    - *Use the decision tree more as a regression tree than as a classification tree*

[FIP2] cross-section fit on the 1-D scoring classifier distribution – evaluation

- FIP special case
  - Cross-section: constant $\frac{1}{s_i}\frac{\partial s_i}{\partial \sigma_s} = \frac{1}{\sigma_s}$
  - Fit on all events: ε_i=1 in all bins
  - Fit scoring classifier: use ROC *and prevalence* to determine purity ρ_i
    - Region of constant ROC slope is a region of constant signal purity

$$\mathrm{FIP2} = \int_0^1 \frac{d\epsilon_s}{1 + \frac{1-\pi_s}{\pi_s}\frac{d\epsilon_b}{d\epsilon_s}}$$

Compare FIP2 to AUC: $\mathrm{AUC} = \int_0^1 \epsilon_s d\epsilon_b = 1 - \int_0^1 \epsilon_b d\epsilon_s$

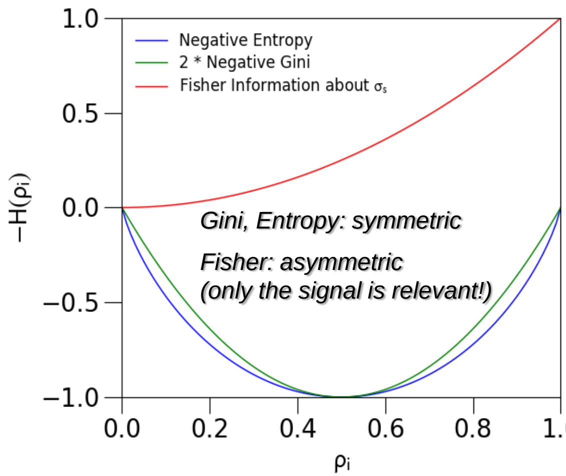


[FIP2] cross-section fit on the 1-D scoring classifier distribution – training

- Is there a gain if I split a node into two ($n \rightarrow n_L+n_R$)?
  - Same question as in optimal partitioning: do I gain by splitting a bin?
- Gain depends on "impurity" function H(ρ): $\Delta = -n_L H(\rho_L) - n_R H(\rho_R) + n H(\rho)$
  - two standard choices: Shannon information (entropy) and Gini impurity
  - *I suggest a third option: Fisher information I_{σ_s} about the cross-section σ_s*
- Surprise: different functions, but *Gini and Fisher gains are equal!*

$$\Delta_{\text{Fisher}} = \frac{(s_L n_R - s_R n_L)^2}{n_L n_R (n_L + n_R)} = \frac{\Delta_{\text{Gini}}}{2}$$

  - So, Gini is OK for cross-sections (or searches?)
  - *But more intuitive physics interpretation for Fisher*
  - No practical gain here, but important principle
    - And proof-of-concept for generic parameter θ



[FIP3] generic parameter fits including the scoring classifier distribution – *work in progress*

- Not a cross-section, e.g. a coupling fit: signal events not all equal
  - –[FIP2] Fit for $\sigma_s \rightarrow$ should partition events into bins with different ρ_i
  - –[FIP3] Fit for θ $\rightarrow$ should partition events into bins with different $\rho_i \frac{1}{s_i}\frac{\partial s_i}{\partial\theta}$

- Example: 2-D fit for θ of the ρ and $\frac{1}{s}\frac{\partial s}{\partial\theta}$ distributions
  - –Train a *regression tree for* $\frac{1}{s}\frac{\partial s}{\partial\theta}$ (on MC weight derivative) using signal alone
  - –Train a *regression tree for ρ* using signal (weighted by $\frac{1}{s}\frac{\partial s}{\partial\theta}$) and background
  - –Use Fisher Information about θ as the gain function in both cases

Boundary between classification and regression even more blurred

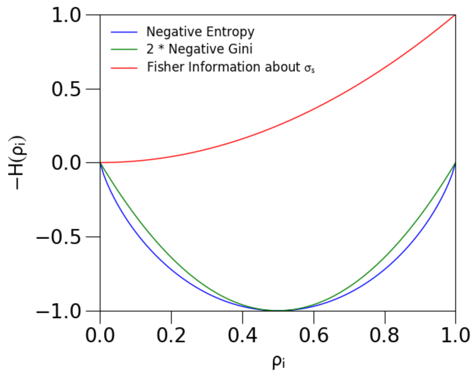
Software technicalities

- I use Python (SciPy, iminuit, bits of rootpy) on SWAN at CERN
  - Thanks to all involved in these projects!
- Custom impurity not available in sklearn DecisionTree's
  - Planned for future sklearn releases (issue #10251 and MR #10325)?
  - I implemented a very simple DecisionTree from scratch, starting from the excellent iCSC notebooks by Thomas Keck (thanks!)
- I plan to make the software available when I find the time…

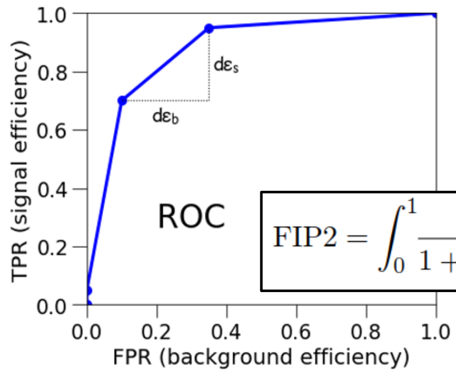
Conclusions and outlook

A paper will be on arxiv soon with all details

- Fisher Information: one metric to bind them all

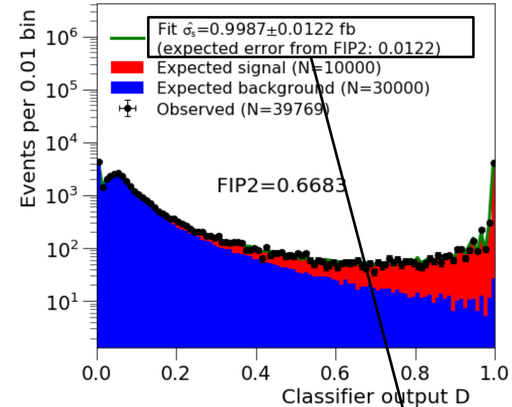


TRAINING
- Fisher Information
  = *measurement error*



$$\mathrm{FIP2} = \int_0^1 \frac{d\epsilon_s}{1 + \frac{1-\pi_s}{\pi_s}\frac{d\epsilon_b}{d\epsilon_s}}$$

EVALUATION
- Fisher Information
  = *measurement error*



PHYSICS
- Fisher Information
  = *measurement error*

- Use scoring classifiers to partition events, not to reject them
  - The boundary between classification and regression is blurred

- We must and can define our own HEP specific metrics
  - I described one case, there are others (searches, systematics, tracking...)
  - Focus on signal. Describe distribution fits. Signal events are not all equal.
  - *Can we please stop using the AUC now?* ☺

Backup slides

Backup – statistical error in binned fits

- Data: *observed event counts n_i* in m bins of a (multi-D) distribution f(x)
  - *expected event counts y_i* = $f(x_i,\theta)dx$ depend on a parameter θ that we want to fit
  - [NB here f is a differential cross section, it is not normalized to 1 like a pdf]

- Fitting θ is like combining the independent measurements in the m bins
  - expected error on n_i in bin x_i is $\Delta n_i = \sqrt{y_i} = \sqrt{f(xi,\theta)\,dx}$
  - expected error on $f(x_i,\theta)$ in bin x_i is $\Delta f = f * \Delta n_i/n_i = \sqrt{f / dx}$
  - expected error on estimated $\hat{\theta}_i$ in bin x_i is $\frac{1}{(\Delta\hat{\theta})^2_{(\mathrm{bin}\ dx)}} = \left(\frac{\partial f}{\partial\theta}\right)^2 \frac{1}{(\Delta f)^2} = \left(\frac{\partial f}{\partial\theta}\right)^2 \left(\frac{\sqrt{dx}}{\sqrt{f}}\right)^2 = \left(\frac{\partial f}{\partial\theta}\right)^2 \frac{dx}{f}$
  - expected error on estimated $\hat{\theta}$ by combining the m bins is $\left(\frac{1}{\Delta\hat{\theta}}\right)^2 = \int \frac{1}{f}\left(\frac{\partial f}{\partial\theta}\right)^2 dx$

- A bit more formally, joint probability for observing the n_i is $P(\mathbf{n};\theta) = \prod_{i=1}^{m} \frac{e^{-y_i} y_i^{n_i}}{n_i!}$
  - Fisher information on θ from the data available is then

$$\mathcal{I}_\theta = E\left[\frac{\partial \log P(\mathbf{n};\theta)}{\partial\theta}\right]^2 \quad \text{i.e.} \quad \mathcal{I}_\theta = \sum_{i=1}^{m} \frac{1}{y_i}\left(\frac{\partial y_i}{\partial\theta}\right)^2 = \int \frac{1}{f}\left(\frac{\partial f}{\partial\theta}\right)^2 dx$$

  - The minimum variance achievable (Cramer-Rao lower bound) is $(\Delta\hat{\theta})^2 = \mathrm{var}(\hat{\theta}) \geq \frac{1}{\mathcal{I}_\theta}$

Optimal partitioning – information inflow

- Information about θ in a binned fit → $\mathcal{I}_\theta = \sum_{i=1}^{m} \frac{1}{y_i} \left(\frac{\partial y_i}{\partial \theta} \right)^2$
- Can I reduce $\Delta\hat{\theta}$ by splitting bin y_i into two bins? $y_i = w_i + z_i$
  - Is the "information inflow" positive? $\frac{1}{w_i}\left(\frac{\partial w_i}{\partial \theta}\right)^2 + \frac{1}{z_i}\left(\frac{\partial z_i}{\partial \theta}\right)^2 - \frac{1}{w_i+z_i}\left(\frac{\partial (w_i+z_i)}{\partial \theta}\right)^2 = \frac{\left(w_i \frac{\partial z_i}{\partial \theta} - z_i \frac{\partial w_i}{\partial \theta}\right)^2}{w_i z_i (w_i+z_i)} \geq 0$
  - information increases (error $\Delta\hat{\theta}$ decreases) if $\frac{1}{w_i}\frac{\partial w_i}{\partial \theta} \neq \frac{1}{z_i}\frac{\partial z_i}{\partial \theta}$
- In the presence of background: $\frac{1}{y_i}\frac{\partial y_i}{\partial \theta} = \rho_i \frac{1}{S_i}\frac{\partial S_i}{\partial \theta}$
  - information increases if $\rho_w \frac{1}{s_w}\frac{\partial s_w}{\partial \theta} \neq \rho_z \frac{1}{s_z}\frac{\partial s_z}{\partial \theta}$
  - therefore: **try to partition the data into bins of different $\rho_i \frac{1}{s_i}\frac{\partial s_i}{\partial \theta}$**
    - for cross-section measurements, $\frac{1}{S_i}\frac{\partial S_i}{\partial \sigma_s} = \frac{1}{\sigma_s}$: *split into bins of different ρ_i*
- Two important practical consequences:
  - *1. use scoring classifiers to partition the data, not to reject events*
  - *2. information can be used also for training classifiers like decision trees*

More detailed slides

(Draft uploaded on July 2nd)

CHEP 2018
Sofia, Bulgaria

Draft uploaded on July 2nd

Fisher information metrics for binary classifier evaluation and training

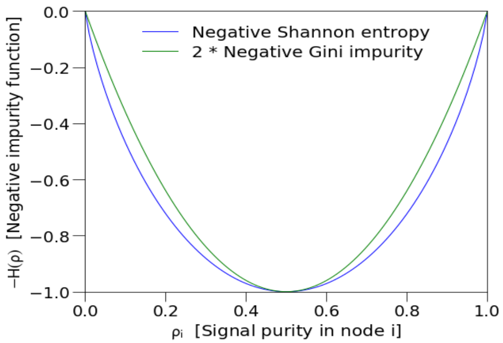
Event selection for HEP precision measurements

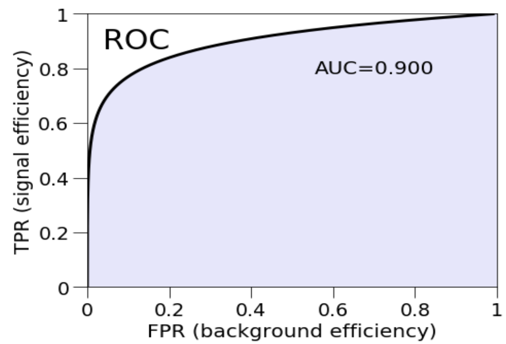
Andrea Valassi
(CERN IT-DI-LCG)

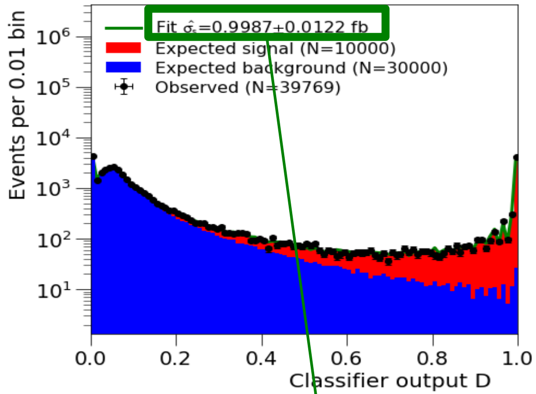
CHEP 2018, Sofia – Machine Learning and Physics Analysis session

Training, Evaluation, Physics: one metric to bind them all?

An <u>oversimplified example</u>: Decision Tree for a cross-section fit







TRAINING
- (either) **Gini impurity**
  - *Economics: inequality*
  - *Ecology: diversity*
- (or) **Shannon information**
  - *Information theory: entropy*

EVALUATION
- **ROC Curve (Receiver Operating Characteristic)**
  - *Signal detection: radar detection*
  - *Psychophysics: sensory detection*
- **AUC (Area Under the ROC Curve)**
  - *Radiology, Medicine: diagnostic accuracy*

PHYSICS
- **Precision**
  - *Parameter estimation:* ***measurement error***

Different problems need different metrics → Always keep the final goal in mind!

<u>Main idea of this talk: use physics precision (Fisher information) also for evaluation and training: MINIMIZE MEASUREMENT ERRORS!</u>

Limited scope of this talk

- Different problems *also within HEP* require different metrics

- In this talk, I will focus on one specific problem:
  - Optimize event selection to minimize statistical errors in point estimation

- Three specific examples (I will focus on the second one)
  - [FIP1] *Total cross-section* measurement in a *counting experiment*
  - [FIP2] *Total cross-section* measurement by *distribution fit*
  - [FIP3] *Generic model parameter fit* (e.g. mass/coupling) by *distribution fit*
    - Even more specific: FIP2 and FIP3 use fits of the scoring classifier distribution

Binary classifier evaluation – reminder

Discrete classifiers: the confusion matrix

Binary decision: signal or background

	true class: **P**ositives (HEP: **signal Stot**)	*true class*: **N**egatives (HEP: **background Btot**)
classified as **P**ositives (HEP: **selected**)	**True Positives (TP)** (HEP: selected signal **Ssel**)	**False Positives (FP)** (HEP: selected bkg **Bsel**)
classified as **N**egatives (HEP: **rejected**)	**False Negatives (FN)** (HEP: rejected signal **Srej**)	**True Negatives (TN)** (HEP: rejected bkg **Brej**)

TP (S_{sel}) FP (B_{sel}) **FN** (S_{rej}) TN (B_{rej})	**TP** (S_{sel}) **FP** (B_{sel}) FN (S_{rej}) TN (B_{rej})	TP (S_{sel}) **FP** (B_{sel}) FN (S_{rej}) **TN** (B_{rej})
$\mathbf{TPR} = \frac{\mathbf{TP}}{\mathbf{TP}+\mathbf{FN}}$	$\mathbf{PPV} = \frac{\mathbf{TP}}{\mathbf{TP}+\mathbf{FP}}$	$\mathbf{TNR} = \frac{\mathbf{TN}}{\mathbf{TN}+\mathbf{FP}} = \mathbf{1}-\mathbf{FPR}$
HEP: "efficiency" $\epsilon_s = \frac{S_{sel}}{S_{tot}}$	HEP: "purity" $\rho = \frac{S_{sel}}{S_{sel}+B_{sel}}$	HEP: "background rejection" $1-\epsilon_b = 1-\frac{B_{sel}}{B_{tot}}$
IR: "recall"	IR: "precision"	—
MED: "sensitivity"	—	MED: "specificity"

Different domains
→ Focus on different concepts
→ Use different terminologies

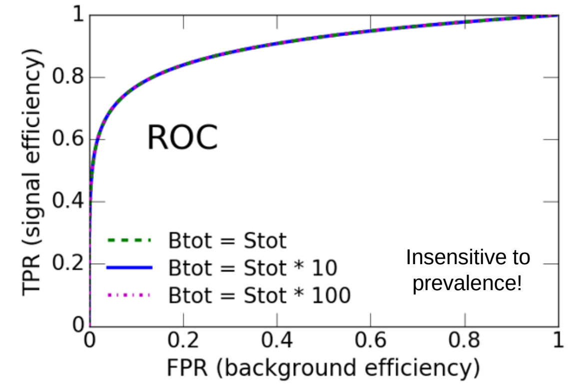
Examples from three domains:
- **Medical Diagnostics (MED)**
  does Mr. A. have cancer?
- **Information Retrieval (IR)**
  Google documents about "ROC"
- **HEP event selection (HEP)**
  select Higgs event candidates

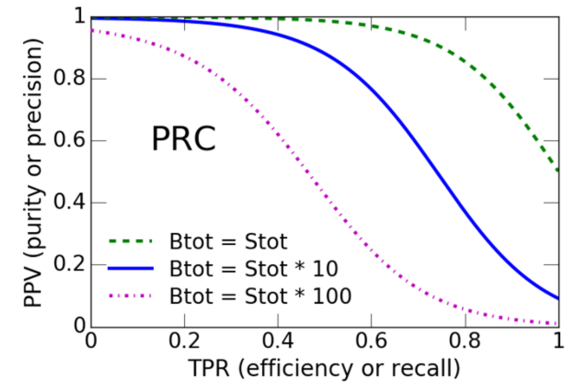
MED: prevalence
$$\pi_s = \frac{S_{tot}}{S_{tot}+B_{tot}}$$

Scoring classifiers: ROC and PRC curves

Continuous output: probability to be signal

Vary the binary decision by varying the cut on the scoring classifier





Binary classifier evaluation in other domains

Medical Diagnostics (MD) $\rightarrow$ diagnostic accuracy

- <u>*Symmetric: all patients important, both truly ill (TP) and truly healthy (TN)*</u>
- Traditional $\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$ was too sensitive to prevalence: moved to ROC
  - But now ROC is questioned as too insensitive to prevalence (imbalanced data)
- ROC-based analysis: sensitivity and specificity
  - Accuracy metric: e.g. *AUC = probability that diagnosis gives greater suspicion to a randomly chosen sick subject than to a randomly chosen healthy subject*

Information Retrieval (IR)

- <u>*Asymmetric: distinction between relevant and non-relevant documents*</u>
- PRC-based evaluation: precision and recall
  - Single metric: e.g. Mean Average Precision ~ area under PRC (AUCPR)

Oversimplification: $\mathrm{AUC} = \int_0^1 \epsilon_s d\epsilon_b = 1 - \int_0^1 \epsilon_b d\epsilon_s$ (MD) vs. (IR) $\mathrm{AUCPR} = \int_0^1 \rho \, d\epsilon_s$

Evaluation: (main) specificities of HEP

1. Qualitative asymmetry: only the signal has interesting physics

- HEP event selection is like Information Retrieval: background is irrelevant
  - *True Negatives and the AUC are irrelevant in HEP event selection*
- Classical evaluation metrics: signal efficiency and purity (the PRC in IR!)
  - *ROC alone is not enough – also need prevalence to interpret the ROC*

2. Distribution fits: several disjoint bins, not just a global selection

- Analyze *local signal efficiency and purity in each bin*, not just global ones
- Counting experiments (e.g. FIP1) vs. distribution fits (e.g. FIP2, FIP3)
  - Special case: fits involving distributions of the scoring classifiers

3. Signal events not all equal: they *may* have different sensitivities

- Example: only events close to a mass peak are sensitive to the mass
- Total cross-section (e.g. FIP1, FIP2) vs. generic parameter fit (e.g. FIP3)

Fisher Information Part (FIP)

- Consider a measurement $\hat{\theta}$ of one physics parameter θ
  - Fisher Information about θ is $1/\Delta\hat{\theta}^2$ (keep this simple, not formal)

- Evaluate an event selection from the effect on the error $\Delta\hat{\theta}$
  - *Compare to an "ideal" case where there is no background*

- FIP: fraction of "ideal" FI that is retained by the real classifier
  - Range in [0,1] $\rightarrow$ 0 if no signal, 1 if select all signal and no background
  - Qualitatively relevant: higher is better $\rightarrow$ maximize FIP to minimize $\Delta\hat{\theta}$
  - Numerically meaningful: related to $\Delta\hat{\theta}$ $\rightarrow$ $(\Delta\hat{\theta}^{(\text{real classifier})})^2 = \frac{1}{\text{FIP}}(\Delta\hat{\theta}^{(\text{ideal classifier})})^2$

- For a binned fit of θ from a (1-D or multi-D) histogram:
  - With expected event counts in i^{th} bin $y_i = \epsilon_i{*}S_i + b_i = \epsilon_i{*}S_i / \rho_i$
  - Consider only statistical errors $\rightarrow$ sum information from the different bins

$$\text{FIP} = \frac{\mathcal{I}_\theta^{(\text{real classifier})}}{\mathcal{I}_\theta^{(\text{ideal classifier})}} = \frac{\sum_{i=1}^{m} \epsilon_i \rho_i \times \frac{1}{S_i}\left(\frac{\partial S_i}{\partial \theta}\right)^2}{\sum_{i=1}^{m} \frac{1}{S_i}\left(\frac{\partial S_i}{\partial \theta}\right)^2}$$

Remember from the previous slide:
1. Only signal is interesting: background appears via ρ_i
2. Distribution fit: need local ϵ_i and ρ_i
3. Signal events are not all equal: need sensitivity $\frac{1}{S_i}\frac{\partial S_i}{\partial \theta}$

[FIP1] Cross-section in counting experiment

- Counting experiment: measure a single number N_{meas}
  - Well-known since decades: maximize ε_s*ρ to minimize statistical errors

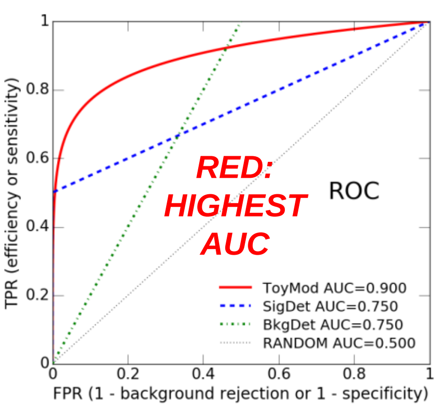
$$(\sigma_s)_{\rm meas} = \frac{N_{\rm meas} - \mathcal{L}\epsilon_b\sigma_b}{\mathcal{L}\epsilon_s} \quad \rightarrow \quad \frac{1}{(\Delta\sigma_s)^2} = \frac{1}{\sigma_s}\mathcal{L}\epsilon_s\rho = \frac{1}{\sigma_s^2}S_{\rm tot}\epsilon_s\rho = \frac{1}{\sigma_s^2}\times\frac{S_{\rm sel}^2}{S_{\rm sel}+B_{\rm sel}}$$

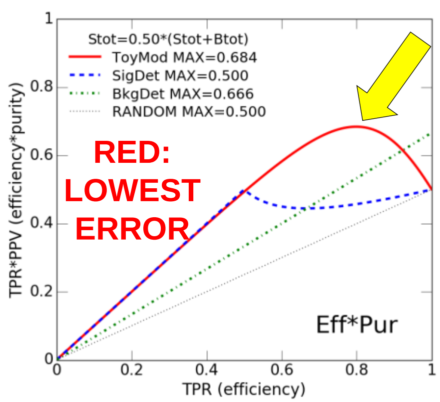
- FIP special case: FIP1 = ε_s*ρ
  - Counting experiment → global signal efficiency and purity
  - Cross-section fit $\theta=\sigma_s$ → all events have equal sensitivity $\frac{1}{S_i}\frac{\partial S_i}{\partial\sigma_s} = \frac{1}{\sigma_s}$

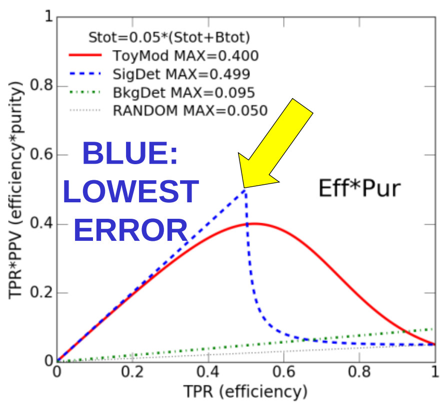
$$\mathrm{FIP} = \frac{\mathcal{I}_\theta^{(\rm real\ classifier)}}{\mathcal{I}_\theta^{(\rm ideal\ classifier)}} = \frac{\sum_{i=1}^{m}\epsilon_i\rho_i \times \frac{1}{S_i}\left(\frac{\partial S_i}{\partial\theta}\right)^2}{\sum_{i=1}^{m}\frac{1}{S_i}\left(\frac{\partial S_i}{\partial\theta}\right)^2} \quad \rightarrow \quad \boxed{\mathrm{FIP1} = \varepsilon_s*\rho}$$

Examples of issues in AUCs – crossing ROCs

- Cross-section measurement by counting experiment
  - maximize FIP1=ε_s*ρ → minimize the statistical error $\Delta\sigma^2$
- Compare two classifiers: red (AUC=0.90) and blue (AUC=0.75)
  - The red and blue ROCs cross (otherwise the choice would be obvious!)
- Choice of classifier achieving minimum $\Delta\sigma^2$ depends on S_{tot}/B_{tot}
  - Signal prevalence 50%: choose classifier with higher AUC (red)
  - Signal prevalence 5%: choose classifier with lower AUC (blue)
  - AUC is irrelevant – and ROC is only useful if you also know prevalence







	FIP1	AUC
Range in [0,1]	YES	YES
Higher is better	YES	NO
Numerically meanigful	YES	NO

Optimal partitioning – information inflow

- Does $\mathcal{I}_\theta = \sum_{i=1}^{m} \frac{1}{y_i} \left(\frac{\partial y_i}{\partial \theta} \right)^2$ increase if I split y_i into two bins? $y_i = w_i + z_i$
  - Information increases and error $\Delta\hat{\theta}$ decreases if $\frac{1}{w_i}\frac{\partial w_i}{\partial \theta} \neq \frac{1}{z_i}\frac{\partial z_i}{\partial \theta}$
  - In the presence of background, $\Delta\hat{\theta}$ decreases if $\rho_w \frac{1}{s_w}\frac{\partial s_w}{\partial \theta} \neq \rho_z \frac{1}{s_z}\frac{\partial s_z}{\partial \theta}$

- Hence: **try to partition the data into bins of different $\rho_i \frac{1}{s_i}\frac{\partial s_i}{\partial \theta}$**
  - For cross-section measurements, $\frac{1}{S_i}\frac{\partial S_i}{\partial \sigma_s} = \frac{1}{\sigma_s}$: *split into bins of different ρ_i*
    - As the scoring classifier represents ρ, fit its distribution! (next slide: FIP2)

- Two important practical consequences:
  - *1. use scoring classifiers to partition the data, not to reject events*
  - *2. information can be used also for training classifiers like decision trees*

[FIP2] cross-section measurement by fitting the 1-D scoring classifier distribution

$$\text{FIP} = \frac{\mathcal{I}_\theta^{\text{(real classifier)}}}{\mathcal{I}_\theta^{\text{(ideal classifier)}}} = \frac{\sum_{i=1}^{m} \epsilon_i \rho_i \times \frac{1}{S_i}\left(\frac{\partial S_i}{\partial \theta}\right)^2}{\sum_{i=1}^{m} \frac{1}{S_i}\left(\frac{\partial S_i}{\partial \theta}\right)^2}$$

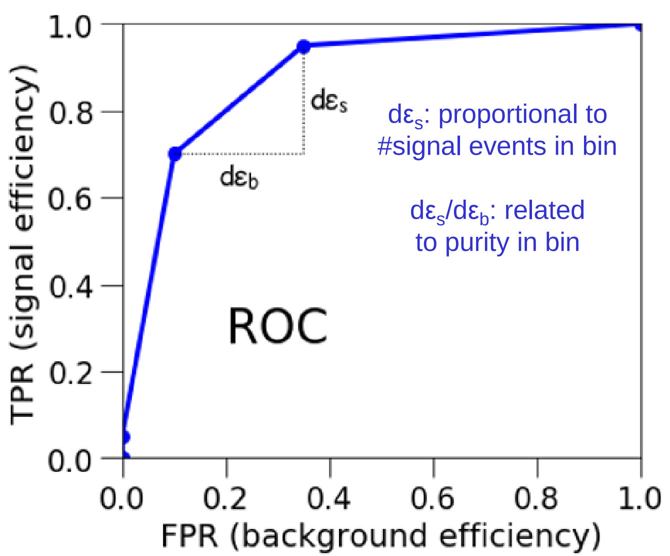
- FIP special case
  - Cross-section: constant $\frac{1}{s_i}\frac{\partial s_i}{\partial \sigma_s} = \frac{1}{\sigma_s}$
  - Fit on all events: ϵ_i=1 in all bins

  $\rightarrow$ $\text{FIP2} = \frac{\sum_{i=1}^m s_i^2/n_i}{\sum_{i=1}^m s_i} = \frac{\sum_{i=1}^m \rho_i s_i}{\sum_{i=1}^m s_i}$

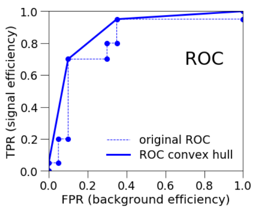
  - Fit the scoring classifier: use ROC* *and* prevalence to determine the local purity $\rho_i = \frac{1}{1+\frac{B_{\text{tot}}}{S_{\text{tot}}}\frac{d\epsilon_b}{d\epsilon_s}}$ in a bin with $s_i = S_{\text{tot}}\, d\epsilon_s$
    - Region of constant ROC slope is a region of constant signal purity

$$\text{FIP2} = \int_0^1 \frac{d\epsilon_s}{1+\frac{1-\pi_s}{\pi_s}\frac{d\epsilon_b}{d\epsilon_s}}$$

Compare FIP2 to AUC: $\text{AUC} = \int_0^1 \epsilon_s d\epsilon_b = 1 - \int_0^1 \epsilon_b d\epsilon_s$

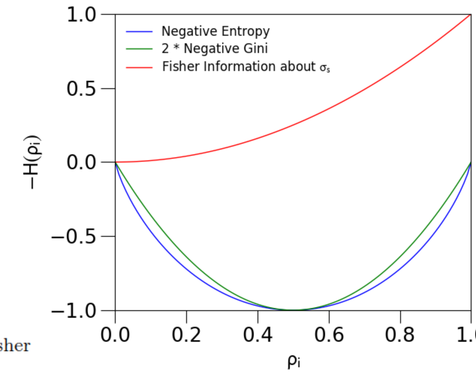


*Technicality (my Python code): convert ROC to *convex hull*
- ensure decreasing slope, i.e. decreasing purity
- avoid staircase effect that would artificially inflate FIP2 (bins of 100% purity: only signal or only background)



FIP2 for training decision trees

- Decision Tree → partition training set *into nodes of different* ρ_i
  - The best split $(n,s)=(n_L,s_L)+(n_R,s_R)$ maximizes $\Delta = -n_L H(\rho_L) - n_R H(\rho_R) + nH(\rho)$
- Current metrics are Gini and entropy: add Fisher information!
  - negative Gini impurity → $-n_i H(\rho_i) = n_i \times [-2\rho_i(1-\rho_i)]$
  - Shannon information → $-n_i H(\rho_i) = n_i \times [\rho_i \log_2 \rho_i + (1-\rho_i)\log_2(1-\rho_i)]$
  - Fisher information on σ_s → $-n_i H(\rho_i) = n_i \times [\rho_i^2]$



- Functions look different, but (modulo a constant factor)…
  - … information gain is the same for Fisher and Gini!

$$\Delta_{\mathrm{Fisher}} = \frac{s_L^2}{n_L} + \frac{s_R^2}{n_R} - \frac{(s_L+s_R)^2}{n_L+n_R} = \frac{(s_L n_R - s_R n_L)^2}{n_L n_R (n_L+n_R)}$$

$$\frac{\Delta_{\mathrm{Gini}}}{2} = -s_L\left(1-\frac{s_L}{n_L}\right) - s_R\left(1-\frac{s_R}{n_R}\right) + (s_L+s_R)\left(1-\frac{s_L+s_R}{n_L+n_R}\right) = \Delta_{\mathrm{Fisher}}$$

- But interpretation is clearer for Fisher: reduce the error on the fit
  - And this is a proof-of-concept for FIP3: split *into nodes of different* $\rho_i \, \frac{1}{s_i}\frac{\partial si}{\partial \theta}$

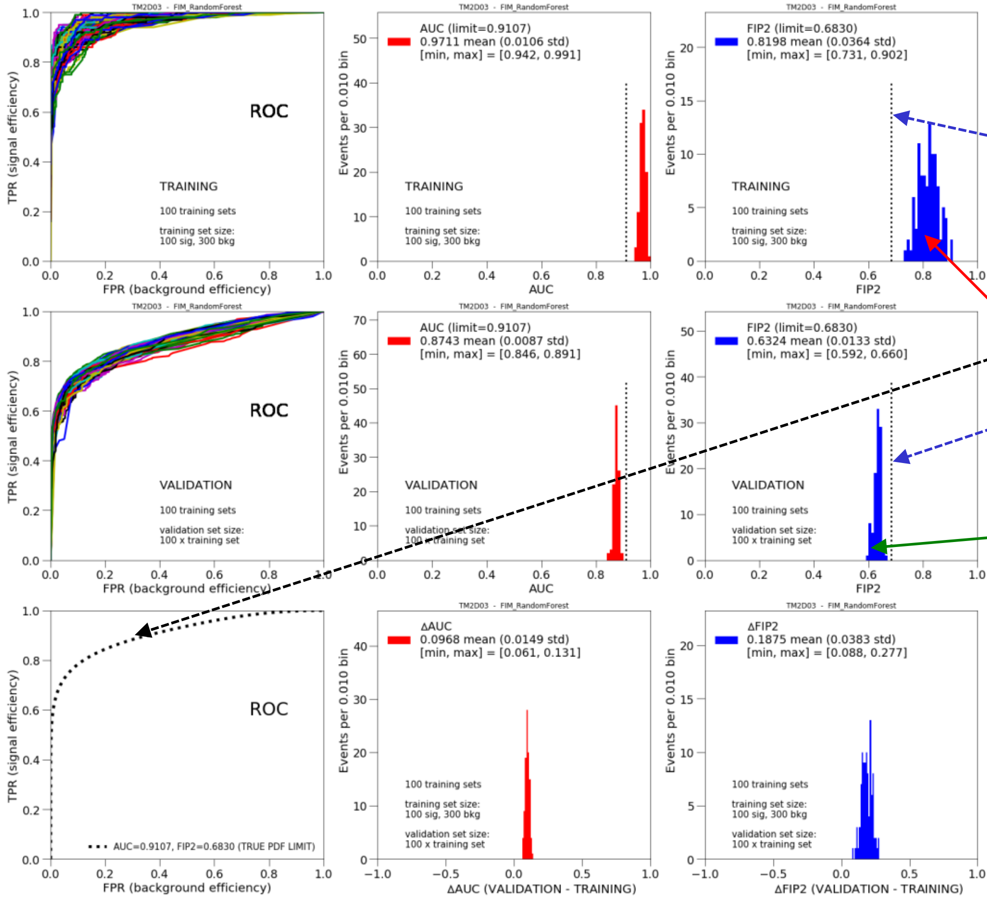
Technicality: user-defined criteria for DecisionTree's will only be available in future sklearn releases
→ I implemented a DecisionTree from scratch, reusing the excellent iCSC notebooks by Thomas Keck (thanks!)

Limits to knowledge

- FIP2 range is [0,1] $\rightarrow$ but it does not mean that 1 is achievable
  - 1 represents the *ideal* case where there is no background
- In some regions of phase space, signal and background events may be undistinguishable based on the available observations
  - There is a limit ROC which depends on the signal and background pdf's
  - There is a limit FIP2 which depends on prevalence and the limit ROC
- Example – toy model, you know the real pdf's and prevalence
  - See next slide about overtraining

Overtraining

- Using the same metric for training and evaluation also simplifies the interpretation of overtraining

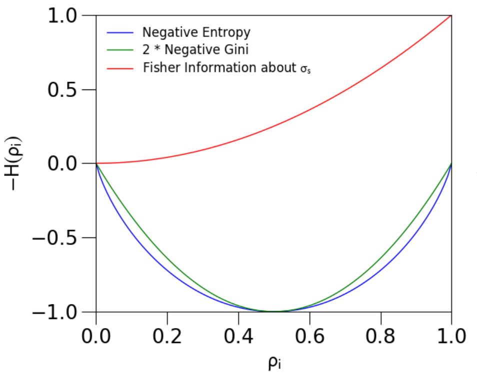


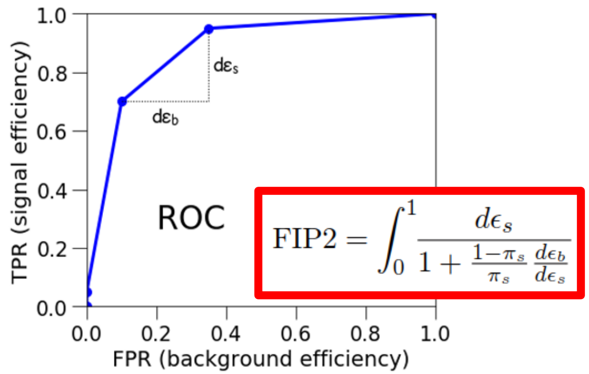
- Example: toy model where you know the real pdf
  - You know the limit ROC
  - You know the limit FIP2
  - You want your validation FIP2 as close as possible to the limit, but it will be lower
  - To get there you maximize your training FIP2, but it will be higher than the real limit
    - You may trace back every increase to one node split
  - You may study the effects of things like min_sample_leaf

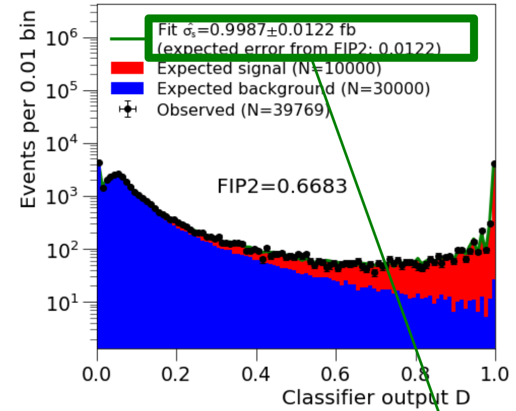
[FIP3] parameter fits including the scoring classifier distribution – *work in progress*

- FIP2 for σ_s fits: one metric for training, evaluation, physics
  - FIP3: one metric for training, evaluation, physics in fits of a generic θ
- Difference with FIP2: include event-by-event sensitivities $\frac{1}{s_i}\frac{\partial s_i}{\partial \theta}$
  - [FIP2] Fit for $\sigma_s \rightarrow$ should partition events into bins of different ρ_i
  - [FIP3] Fit for θ $\rightarrow$ should partition events into bins of different $\rho_i \frac{1}{s_i}\frac{\partial s_i}{\partial \theta}$
    - Example: a 1-D fit on $\rho_i \frac{1}{s_i}\frac{\partial s_i}{\partial \theta}$ or (better) a 2-D fit on ρ_i and $\frac{1}{s_i}\frac{\partial s_i}{\partial \theta}$
- Challenge: what is the value of $\frac{1}{s_i}\frac{\partial s_i}{\partial \theta}$ for real data events?
  - On MC events you can get it from event-by-event MC weight derivatives
  - On data, train a regression tree for $\frac{1}{s_i}\frac{\partial s_i}{\partial \theta}$ on signal only and a decision tree for ρ_i on signal+bkg: use Fisher Information as splitting criterion in both
- *The boundary between classification and regression is blurred!*

Conclusions: one metric to bind them all







TRAINING
- **Fisher Information**
  = measurement error

EVALUATION
- **Fisher Information**
  = measurement error

PHYSICS
- **Fisher Information**
  = measurement error

- One metric for training, evaluation, physics: Fisher Information
- FI meets HEP specificities for evaluation: focuses on signal; describes distribution fits; describes event-by-event sensitivity
  – Different problems need different metrics: HEP needs its own metrics
- The boundary between binary classification and regression is blurred: should partition events into bins of different $\rho_i \frac{1}{s_i}\frac{\partial s_i}{\partial \theta}$

Additional backup slides

Selected slides from my previous IML talks

in April (https://indico.cern.ch/event/668017/contributions/2947015)

and January (https://indico.cern.ch/event/679765/contributions/2814562)

FIP2 from the ROC (+prevalence) or from the PRC

FIP2: integrals on ROC and PRC,
more relevant to HEP than AUC or AUCPR!
(well-defined meaning for distribution fits)

- From the previous slide: $\mathrm{FIP2} = \frac{\sum_{i=1}^{m} \rho_i s_i}{\sum_{i=1}^{m} s_i}$

- FIP2 from the ROC (+prevalence $\pi_s = \frac{S_{\mathrm{tot}}}{S_{\mathrm{tot}} + B_{\mathrm{tot}}}$):

$$S_{\mathrm{sel}} = S_{\mathrm{tot}}\,\epsilon_s, \quad B_{\mathrm{sel}} = B_{\mathrm{tot}}\,\epsilon_b \;\Rightarrow\; s_i = dS_{\mathrm{sel}} = S_{\mathrm{tot}}\,d\epsilon_s, \quad b_i = dB_{\mathrm{sel}} = B_{\mathrm{tot}}\,d\epsilon_b \;\Rightarrow\; \rho_i = \frac{1}{1 + \frac{B_{\mathrm{tot}}}{S_{\mathrm{tot}}}\frac{d\epsilon_b}{d\epsilon_s}} \;\Rightarrow\; \mathrm{FIP2} = \int_0^1 \frac{d\epsilon_s}{1 + \frac{1-\pi_s}{\pi_s}\frac{d\epsilon_b}{d\epsilon_s}}$$

Compare FIP2(ROC) to AUC

$$\mathrm{AUC} = \int_0^1 \epsilon_s d\epsilon_b = 1 - \int_0^1 \epsilon_b d\epsilon_s$$

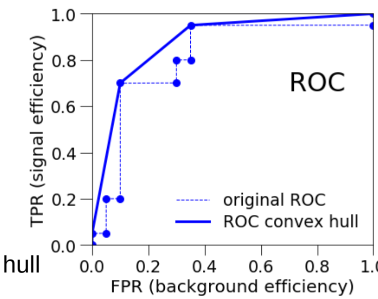
- FIP2 from the PRC:

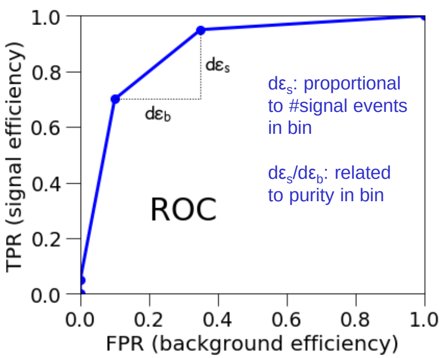
$$S_{\mathrm{sel}} = S_{\mathrm{tot}}\,\epsilon_s, \quad B_{\mathrm{sel}} = S_{\mathrm{sel}}\left(\frac{1}{\rho} - 1\right) \;\Rightarrow\; s_i = dS_{\mathrm{sel}} = S_{\mathrm{tot}}\,d\epsilon_s, \quad b_i = dB_{\mathrm{sel}} = S_{\mathrm{tot}}\left[d\epsilon_s\left(\frac{1}{\rho} - 1\right) - \epsilon_s\frac{d\rho}{\rho^2}\right] \;\Rightarrow\; \rho_i = \frac{\rho}{1 - \frac{\epsilon_s}{\rho}\frac{d\rho}{d\epsilon_s}} \;\Rightarrow\; \mathrm{FIP2} = \int_0^1 \frac{\rho\, d\epsilon_s}{1 - \frac{\epsilon_s}{\rho}\frac{d\rho}{d\epsilon_s}}$$

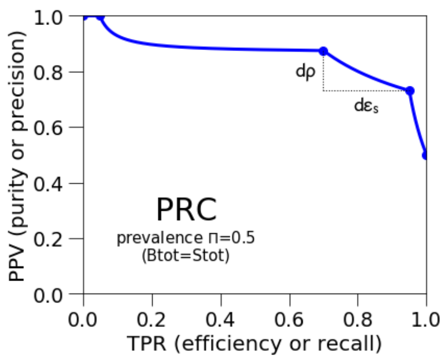
Compare FIP2(PRC) to AUCPR

$$\mathrm{AUCPR} = \int_0^1 \rho\, d\epsilon_s$$

- Easier calculation and interpretation from ROC (+prevalence) than from PRC
  - *region of constant ROC slope** *= region of constant signal purity*
  - decreasing ROC slope = decreasing purity
    - technicality (my Python code): convert ROC to convex hull** first

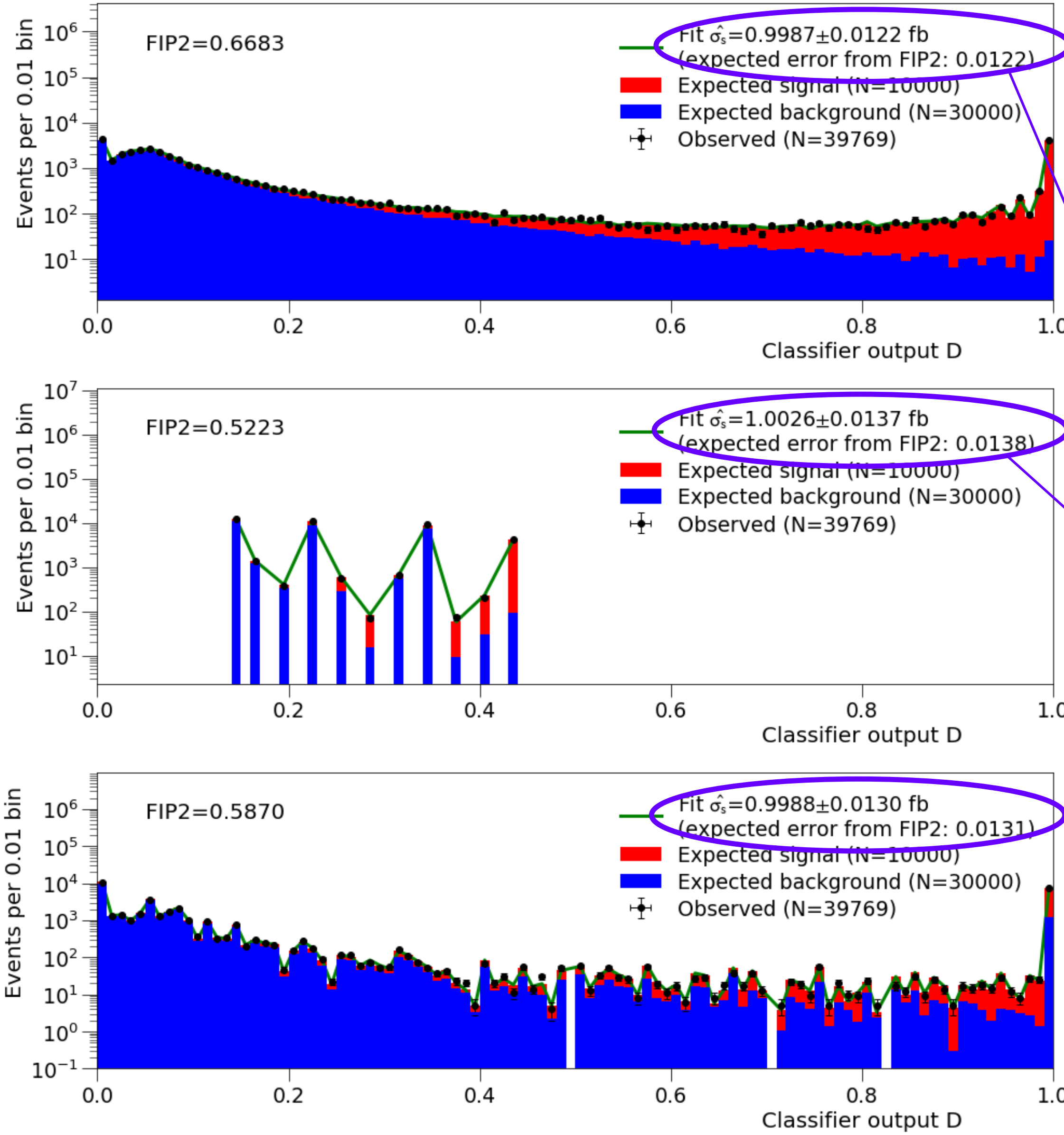






**Convert ROC to convex hull
- ensure decreasing slope
- avoid staircase effect that would artificially inflate FIP2
(bins of 100% purity: only signal or only background)

*ROC slopes are also discussed in medical literature in relation to diagnostic likelihood ratios [Choi 1998], but their use does not seem to be widespread(?)



Sanity check

- Three random forests (on the same 2-D pdf)
  - reasonable
  - undertrained
  - overtrained
- Fit σ_s from the distribution of the classifier output
  - Errors consistent with FIP2

$$(\Delta\hat{\theta}^{(\text{real classifier})})^2 = \frac{1}{\text{FIP}}(\Delta\hat{\theta}^{(\text{ideal classifier})})^2$$

My development environment: SciPy ecosystem, iminuit and bits of rootpy, on SWAN at CERN. Thanks to all involved in these projects!

M by 1D fit to m – visual interpretation

- Information after cuts: $\sum_i \frac{1}{s_i}\left(\frac{\partial si}{\partial M}\right)^2 * \varepsilon_i * \rho_i \rightarrow$ show the 3 terms in each bin i
  - fit = combine N different measurements in N bins $\rightarrow$ local $\varepsilon_{i,}$ ρ_i relevant!
  - important thing is: maximise purity, efficiency in bins with highest sensitivity!

Ideal case - yellow histogram (after cuts) coincides with and covers red histogram (ideal)

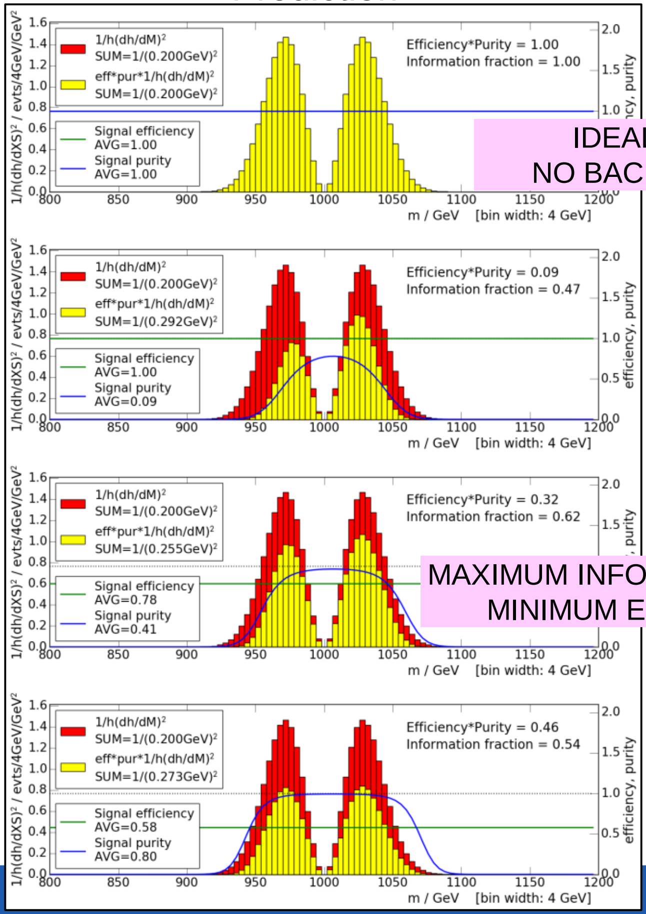
Red histogram: information per bin, ideal case $\frac{1}{s_i}\left(\frac{\partial si}{\partial M}\right)^2$

Blue line: local purity in the bin, ρ_i

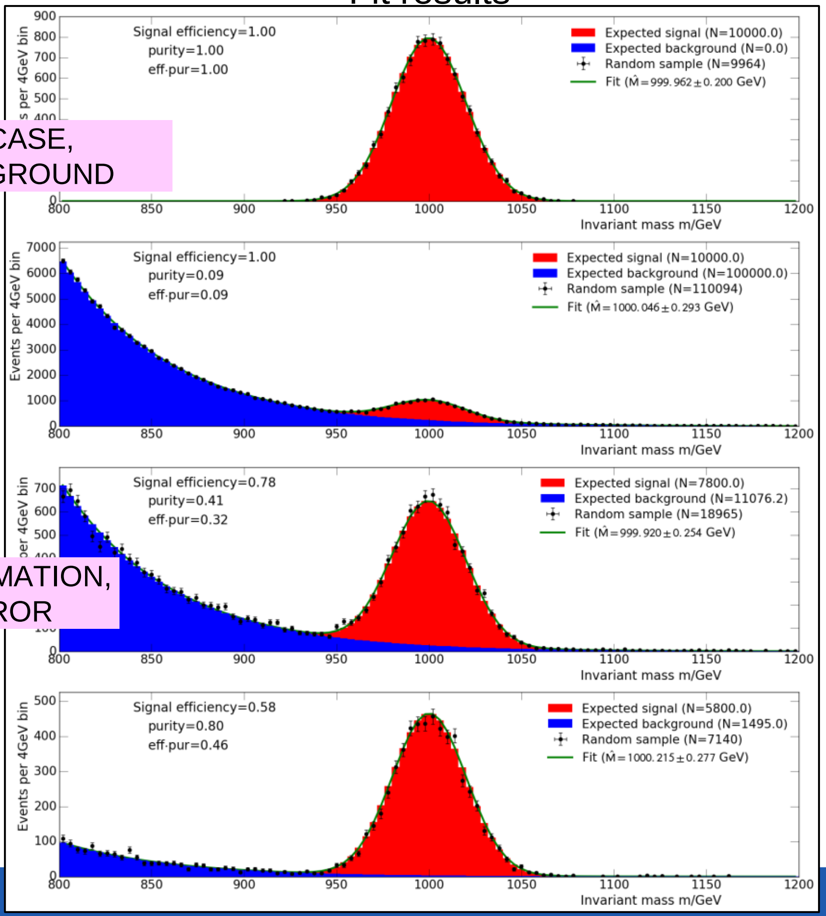
Green line: local efficiency in the bin, ε_i

Yellow histogram: information per bin, after cuts $\varepsilon_i * \rho_i * \frac{1}{s_i}\left(\frac{\partial si}{\partial M}\right)^2$

Prediction



Fit results



IDEAL CASE,
NO BACKGROUND

MAXIMUM INFORMATION,
MINIMUM ERROR

Event selection in HEP searches

- Statistical error in searches by counting experiment → "significance"
  - several metrics → but optimization always involves ϵ_s, ρ alone → TN irrelevant

C. Adam-Bourdarios et al., *The Higgs Machine Learning Challenge*, Proc. NIPS 2014 Workshop on High-Energy Physics and Machine Learning (HEPML2014), Montreal, Canada, PMLR 42 (2015) 19. http://proceedings.mlr.press/v42/cowa14.html

$$Z_0 = \frac{S_\mathrm{sel}}{\sqrt{S_\mathrm{sel}+B_\mathrm{sel}}} \Longrightarrow (Z_0)^2 = S_\mathrm{tot}\epsilon_s\rho$$

Z_0 – Not recommended? (confuses search with measuring σ_s once signal established)

Z_2 – Most appropriate? (also used as "AMS2" in Higgs ML challenge)

$$Z_2 = \sqrt{2\left((S_\mathrm{sel}+B_\mathrm{sel})\log(1+\frac{S_\mathrm{sel}}{B_\mathrm{sel}}) - S_\mathrm{sel}\right)} \Longrightarrow (Z_2)^2 = 2S_\mathrm{tot}\epsilon_s\left(\frac{1}{\rho}\log(\frac{1}{1-\rho})-1\right) = S_\mathrm{tot}\epsilon_s\rho\left(1+\frac{2}{3}\rho+\mathcal{O}(\rho^2)\right)$$

$$Z_3 = \frac{S_\mathrm{sel}}{\sqrt{B_\mathrm{sel}}} \Longrightarrow (Z_3)^2 = S_\mathrm{tot}\epsilon_s\frac{\rho}{1-\rho} = S_\mathrm{tot}\epsilon_s\rho\left(1+\rho+\mathcal{O}(\rho^2)\right)$$

Z_3 ("AMS3" in Higgs ML) – Most widely used, but strictly valid only as an approximation of Z_2 as an expansion in $S_{sel}/B_{sel} \ll 1$?

$$\frac{S_\mathrm{sel}}{B_\mathrm{sel}} = \frac{\rho}{1-\rho} = \rho\left(1+\rho+\mathcal{O}(\rho^2)\right)$$

Expansion in $\rho \ll 1$? – use the expression for Z_2 if anything

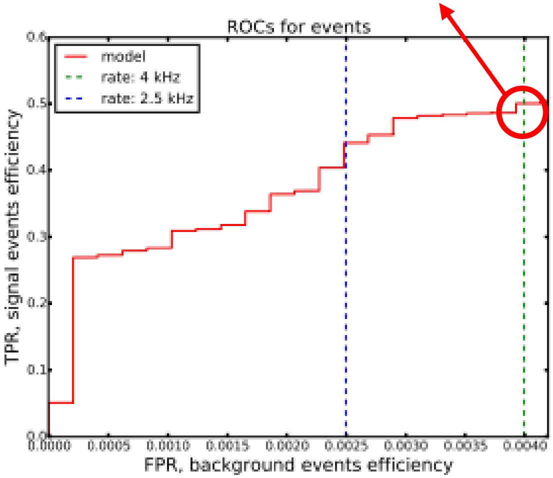
G. Punzi, *Sensitivity of searches for new signals and its optimization*, Proc. PhyStat2003, Stanford, USA (2003). arXiv:physics/0308063v2 [physics.data-an]
G. Cowan, E. Gross, *Discovery significance with statistical uncertainty in the background estimate*, ATLAS Statistics Forum (2008, unpublished). http://www.pp.rhul.ac.uk/~cowan/stat/notes/SigCalcNote.pdf (accessed 15 January 2018)

R. D. Cousins, J. T. Linnemann, J. Tucker, *Evaluation of three methods for calculating statistical significance when incorporating a systematic uncertainty into a test of the background-only hypothesis for a Poisson process*, Nucl. Instr. Meth. Phys. Res. A 595 (2008) 480. doi:10.1016/j.nima.2008.07.086
G. Cowan, K. Cranmer, E. Gross, O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, Eur. Phys. J. C 71 (2011) 15. doi:10.1140/epjc/s10052-011-1554-0

- Several other interesting open questions → *beyond the scope of this talk*
  - optimization of systematics? → e.g. see AMS1 in Higgs ML challenge
  - predict significance in a binned fit? → integral over Z^2 (=sum of log likelihoods)?

Trigger

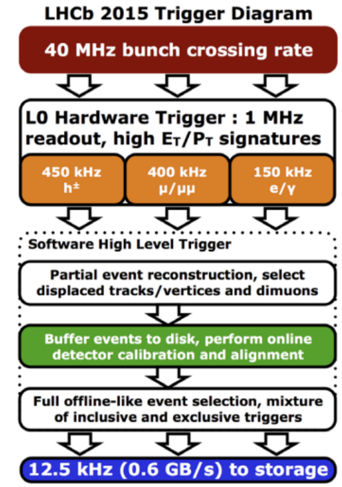
Maximise ε_s at 4 kHz



T. Likhomanenko et al., *LHCb Topological Trigger Reoptimization*, Proc. CHEP 2015, J. Phys. Conf. Series 664 (2015) 082025. doi:10.1088/1742-6596/664/8/082025

Figure 2. Trigger events ROC curve. An output rate of 2.5 kHz corresponds to an FPR of 0.25%, 4 kHz — 0.4%. Thus to find the signal efficiency for a 2.5 kHz output rate, we take 0.25% background efficiency and find the point on the ROC curve that corresponds to this FPR.

IIUC, 4kHz is ε_b (FPR) = 0.4% of 1 MHz L0 hw rate

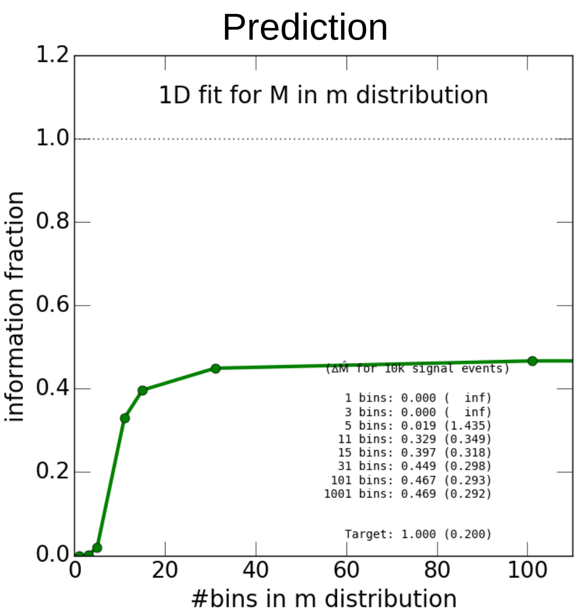


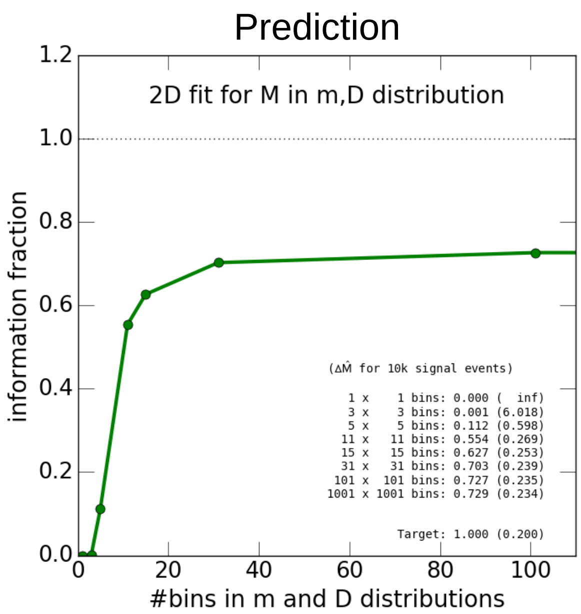
F. Dordei, *LHCb detector and trigger performance in Run II*, Proc. 5th Int. Conf. on New Frontiers in Physics (ICNFP 2016), EPJ Web of Conferences 164, 01016 (2017). doi:10.1051/epjconf/201716401016

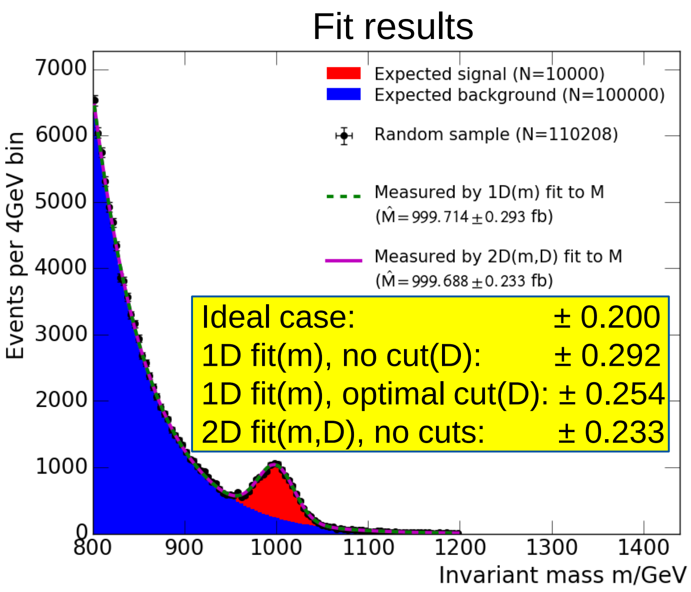
- Different meaning of absolute numbers in the confusion matrix
  - Trigger → events per unit time i.e. trigger rates
  - (Physics analyses → total event sample sizes i.e. total integrated luminosities)
- Binary classifier optimisation goal: maximise ε_s for a given B_{sel} per unit time
  - i.e. maximise TP/(TP+FN) for a given FP → TN irrelevant
- Relevant plot → ε_s vs. B_{sel} per unit time (i.e. *TPR vs FP*)
  - ROC curve (*TPR vs. FPR*) confusing – ***AUC irrelevant***
  - e.g. maximise ε_s for 4 kHz trigger rate, whether L0 rate is 1 MHz or 2MHz

M by 2D fit – use classifier to partition, not to cut

- Showed a fit for M on m, after a cut on D $\rightarrow$ can also fit in 2-D with no cuts
  - again, use the scoring classifier D to partition data, not to reject events
- Why is binning so important, especially using a discriminating variable?
  - next slide...







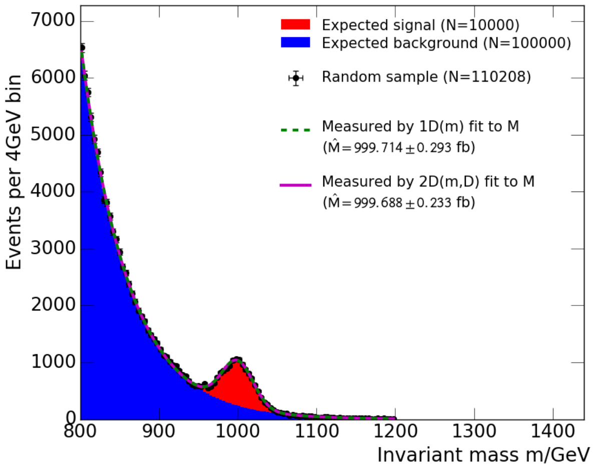
Optimal partitioning – optimal variables

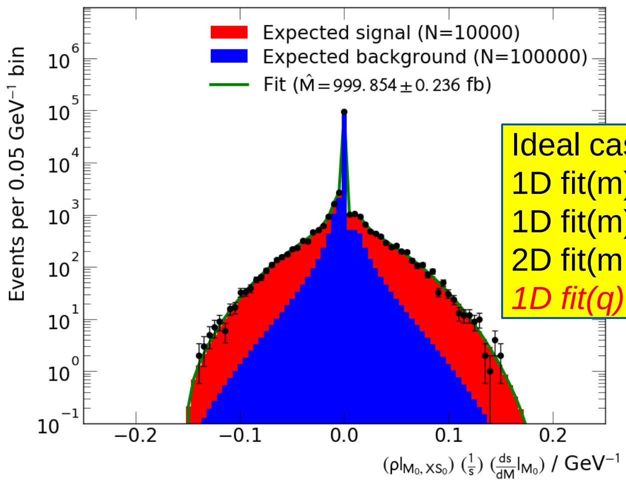
- The previous slide implies that q = $\rho \frac{1}{s}\frac{\partial s}{\partial \theta}$ is an optimal variable to fit for θ
  - proof of concept → 1-D fit of q has the same precision on M as 2-D fit of (m,D)
  - closely related to the "optimal observables" technique

M. Davier, L. Duflot, F. LeDiberder, A. Rougé, *The optimal method for the measurement of tau polarization*, Phys. Lett. B 306 (1993) 411. doi:10.1016/0370-2693(93)90101-M

M. Diel, O. Nachtmann, *Optimal observables for the measurement of three-gauge-boson couplings in $e^+e^- \to W^+W^-$*, Z. Phys. C 62 (1994) 397. doi:10.1007/BF01555899

O. Nachtmann, F. Nagel, *Optimal observables and phase-space ambiguities*, Eur. Phys. J. C40 (2005) 497. doi:10.1140/epjc/s2005-02153-9





Ideal case:	± 0.200
1D fit(m), no cut(D):	± 0.292
1D fit(m), optimal cut(D):	± 0.254
2D fit(m,D), no cuts:	± 0.233
1D fit(q):	*± 0.236*

- In practice: train one ML variable to reproduce $\frac{1}{s}\frac{\partial s}{\partial \theta}$?
  - not needed for cross-sections or searches (this is constant)