The BOSARIS Toolkit

Theory, Algorithms and Code for Surviving the New DCF

Niko Brümmer and Edward de Villiers

AGNITIO Research, South Africa

SRE'11 Analysis Workshop, Atlanta, 6–7 December 2011

Agenda

The purpose of this presentation is to:

1. Introduce the BOSARIS Toolkit.
2. Discuss database size requirements.
3. Introduce the normalized Bayes error-rate plot, a tool for evaluating calibrated likelihood-ratios.
4. Discuss interesting relationships between DET-curves, minDCF and EER.

Outline

The BOSARIS Toolkit

Introduction

The BOSARIS Toolkit is a freely available MATLAB Toolkit, for processing binary classifier scores. The emphasis is on:

- Efficient processing of large trial lists.
- Coverage of a wide range of operating points, including the challenging 'new DCF'.
- Evaluation of the goodness of both uncalibrated scores and calibrated likelihood-ratios.
- Fusion and calibration.

The BOSARIS Toolkit
Collaborators

The Toolkit was created during SRE'10 and the subsequent BOSARIS Workshop.

- Core implementation by AGNITIO
- Collaboration with BUT, CRIM, SRI, Politecnico Torino, SVOX and University of Zaragoza
- Notable contributions from Lukáš Burget, Oldřich Plchot and Nicolas Scheffer

The BOSARIS Toolkit

Practical details

`http://sites.google.com/site/bosaristoolkit`

Documentation:

- User Guide (contents similar to this presentation)
- User Manual (with practical coding details).

MATLAB version: R2008a or later (toolkit has object-oriented API).

Platform-independent interface with other tools:

- (large) text files
- (more efficient) binary HDF5 files

Outline

How many trials do we need?
Counting errors

We need error counts for everything:

- All of the evaluation criteria that we consider below, depend explicitly or implicitly on the counting of errors.
- Training of fusion and calibration and more generally all kinds of system optimization, depend on evaluation criteria and therefore on error counts.

How many trials do we need?
The problem

The problem with error counting is:

- For any system (no matter how accurate), and for any supervised database (no matter how large) there are operating points where false-alarm or miss counts become small and vanish.
- Small, or zero error-counts give unreliable estimates of the error-rates that are to be expected on unseen data.

How many trials do we need?
Analysis

This problem can be analysed with various

- frequentist (e.g. confidence interval) or,
- Bayesian (e.g. credible interval) methods.

The answers will differ, because different analyses depend on different assumptions to make them tractable.

How many trials do we need?
Analysis

One such analysis, Doddington's Rule of 30, uses the assumption of independent Bernoulli trials, to recommend:

> *You need at least 30 errors, for the count frequency to give a probably, approximately correct error-rate estimate.*

We found in SRE'10 and afterwards that this is a useful, practical rule of thumb.

How many trials do we need?
Rule of thumb

Rule of thumb:

> *When training, fusing, calibrating, testing, evaluating, etc. you need the supervised database to be large enough so that, for the system under consideration, you get:*
>
> - *at least 30 false alarms and*
> - *at least 30 misses*
> - *at all operating points of interest.*

The BOSARIS Toolkit makes provision for indicating on various plots where the error counts drop below 30.

Outline

Normalized Bayes error-rate plot

The Normalized Bayes error-rate plot is a new tool for:

- calibration-sensitive evaluation of
- the decision-making ability of system likelihood-ratios,
- over a representative range of operating points.

We show how to generalize the familiar DCF to construct such plots.

DCF

DCF evaluates submitted hard decisions, made by the evaluee:

$$\mathrm{DCF} = \pi C_{\mathrm{miss}} P_{\mathrm{miss}} + (1-\pi) C_{\mathrm{fa}} P_{\mathrm{fa}}$$

where

$0 < \pi < 1$ is the target prior,

$C_{\mathrm{miss}}, C_{\mathrm{fa}} > 0$ are costs,

$P_{\mathrm{miss}}, P_{\mathrm{fa}}$ are the empirical miss and false-alarm rates at some score decision threshold set by the evaluee.

DCF evaluates at a fixed, known operating point.

Bayes risk

Bayes risk evaluates submitted likelihood ratios:

$$\mathcal{R}(\pi, C_{\text{miss}}, C_{\text{fa}}) = \pi C_{\text{miss}} P_{\text{miss}}(\eta) + (1-\pi) C_{\text{fa}} P_{\text{fa}}(\eta)$$

where decisions are made by the evaluator by comparing the likelihood-ratios against the Bayes decision threshold:

$$\eta = \frac{(1-\pi) C_{\text{fa}}}{\pi C_{\text{miss}}}$$

Bayes risk can evaluate over a range of operating points, by varying the parameters, π, C_{miss} and C_{fa}.

Bayes error-rate

From 3 to 1 dimensions

How can we cover all of the values of π, C_{miss} and C_{fa} in a single evaluation procedure?

We show:

- We can combine these 3 parameters into a one-dimensional operating point,
- without loss of generality.

Bayes error-rate

From 3 to 1 dimensions

Define the effective prior:

$$\tilde{\pi} = \frac{\pi C_{\text{miss}}}{\pi C_{\text{miss}} + (1-\pi) C_{\text{fa}}}$$

and the Bayes error-rate:

$$\mathcal{E}(\tilde{\pi}) = \mathcal{R}(\tilde{\pi}, 1, 1) = \frac{\mathcal{R}(\pi, C_{\text{miss}}, C_{\text{fa}})}{\pi C_{\text{miss}} + (1-\pi) C_{\text{fa}}}$$

Bayes error-rate

Error-rate and risk are equivalent

So that:

- Bayes error-rate and Bayes risk are proportional:

$$\mathcal{E}(\tilde{\pi}) = k\mathcal{R}(\pi, C_{\text{miss}}, C_{\text{fa}})$$

  where $k > 0$ does not depend on the system under evaluation.

- For evaluation purposes, these two criteria are equivalent.

Bayes error-rate
Operating point

We have constructed the Bayes error rate criterion:

$$\mathcal{E}(\tilde{\pi}) = \tilde{\pi} P_{\text{miss}}\left(\frac{1-\tilde{\pi}}{\tilde{\pi}}\right) + (1-\tilde{\pi}) P_{\text{fa}}\left(\frac{1-\tilde{\pi}}{\tilde{\pi}}\right)$$

which is parametrized by a single, scalar 'operating point' parameter: $0 < \tilde{\pi} < 1$.

Note, $\tilde{\pi}$ plays two roles:

- it weights the error-rates
- it determines the decision threshold

Bayes error-rate
Operating point

NIST's operating points were:

- 'Old DCF': $\tilde{\pi} \approx 0.092$
- 'New DCF': $\tilde{\pi} = 0.001$

But now we generalize. We construct an evaluation recipe that sweeps the operating point, just like the DET curve does.

Bayes error-rate

Reference values

Default error-rate: Decisions based on $\tilde{\pi}$ alone (i.e. LR $= 1$) give:

$$\mathcal{E}_0(\tilde{\pi}) = \min(\tilde{\pi}, 1 - \tilde{\pi})$$

minDCF: Evaluator optimizes the decision threshold (γ):

$$\mathcal{E}_{\min}(\tilde{\pi}) = \text{minDCF}(\tilde{\pi}, 1, 1) = \min_{\gamma} \tilde{\pi} P_{\text{miss}}(\gamma) + (1 - \tilde{\pi}) P_{\text{fa}}(\gamma)$$

At operating point $\tilde{\pi}$, a system is

well-calibrated, if $\mathcal{E}(\tilde{\pi}) \approx \mathcal{E}_{\min}(\tilde{\pi})$.

badly calibrated, if $\mathcal{E}(\tilde{\pi}) > \mathcal{E}_0(\tilde{\pi})$.

Normalized Bayes error-rate plot

Definition

To turn Bayes error-rate into a graphical analysis tool, we plot Bayes error-rate as a function of the operating point:

- The vertical axis is the normalized Bayes error-rate:

$$y = \frac{\mathcal{E}(\tilde{\pi})}{\mathcal{E}_0(\tilde{\pi})}$$

- The horizontal axis is the (effective) prior log odds:

$$x = \text{logit}(\tilde{\pi}) = \log \frac{\tilde{\pi}}{1-\tilde{\pi}} = -\log \eta$$

where η is just the Bayes decision threshold.

Normalized Bayes error-rate plot

Axis amplification

Reasons for axis amplifications:

- Un-normalized $\mathcal{E}(\tilde{\pi})$ becomes small to the point of invisibility, near $\tilde{\pi} = 0$ and $\tilde{\pi} = 1$.
- $x = \text{logit}(\tilde{\pi})$ spreads out operating points that would otherwise be compressed against $\tilde{\pi} = 0$ and $\tilde{\pi} = 1$.

Normalized Bayes error-rate plot

Examples

Below we show some examples of normalized Bayes error-rate plots, for:

- Synthetic data, to show the effects of deliberate miscalibration.
- LRs from three real automatic speaker recognition systems, from NIST SRE 2010.

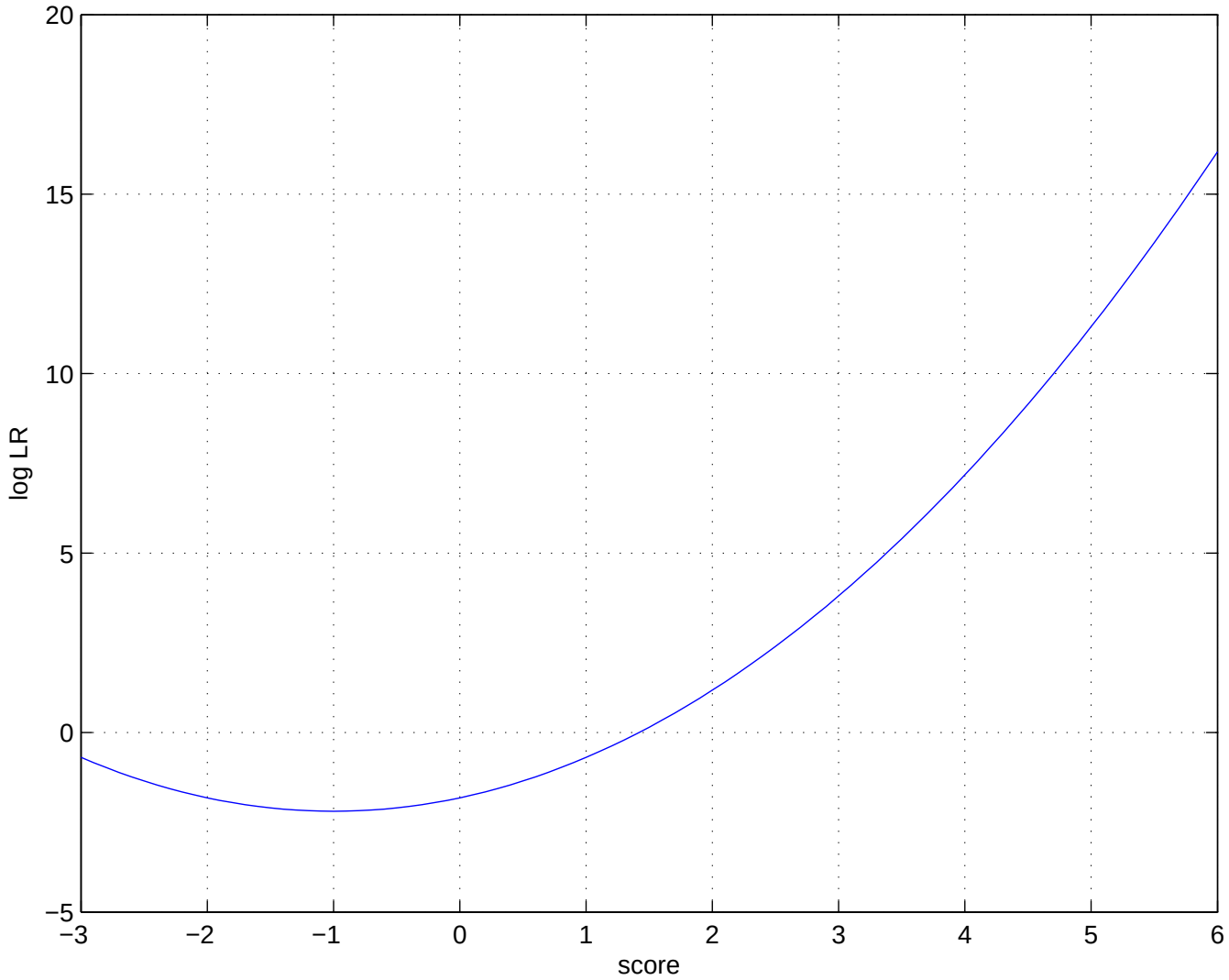
Synthetic example

We generate Gaussian scores, with likelihoods:

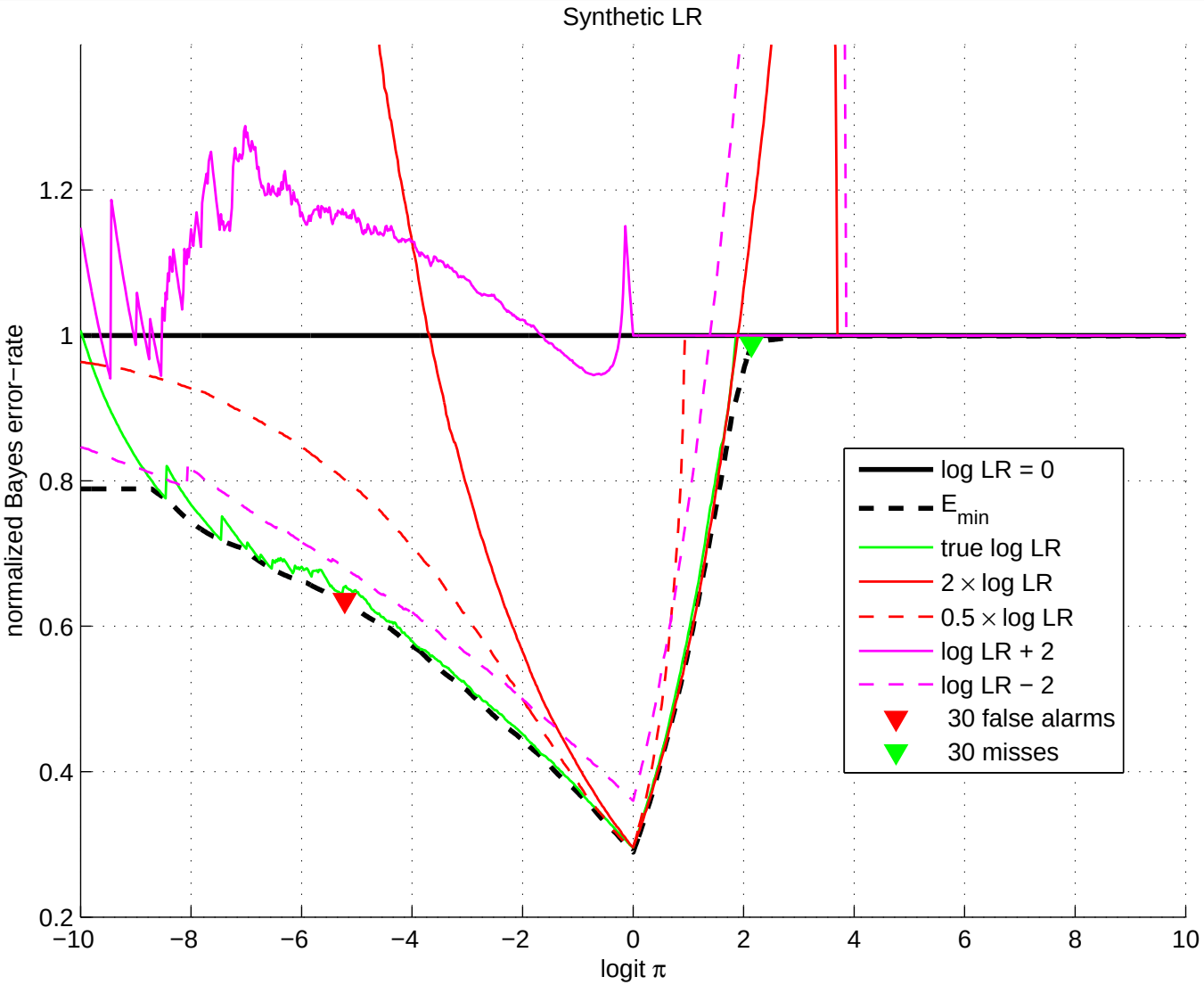
$$P(s|\text{target}) = \mathcal{N}(s|3,2), \qquad P(s|\text{non-target}) = \mathcal{N}(s|0,1)$$

The true $\log \text{LR}(s)$ is a parabola, with turning point at $\log \text{LR} \approx -2$. See plot.

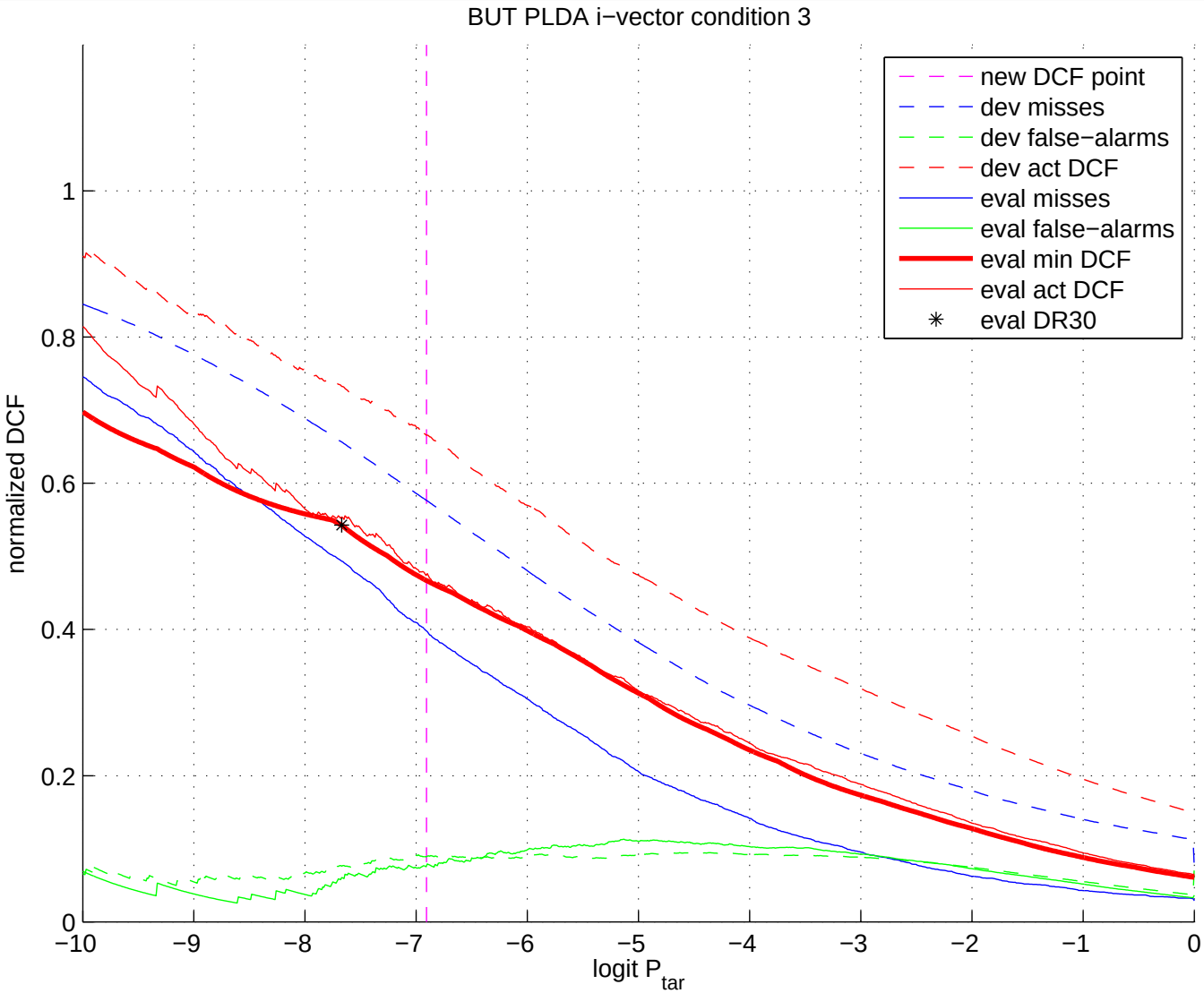
Log LR for synthetic Gaussian scores



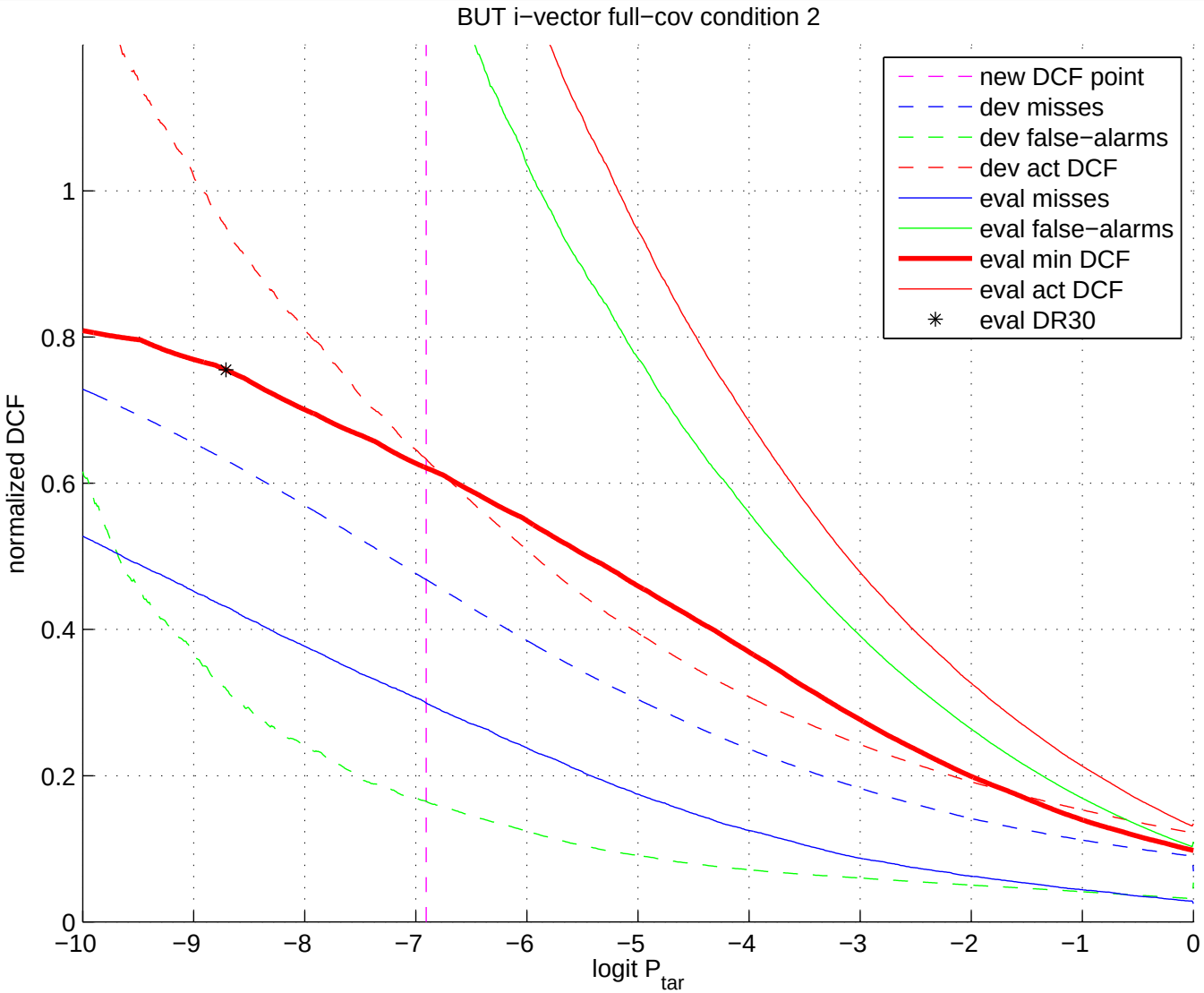
Normalized Bayes error-rate plot



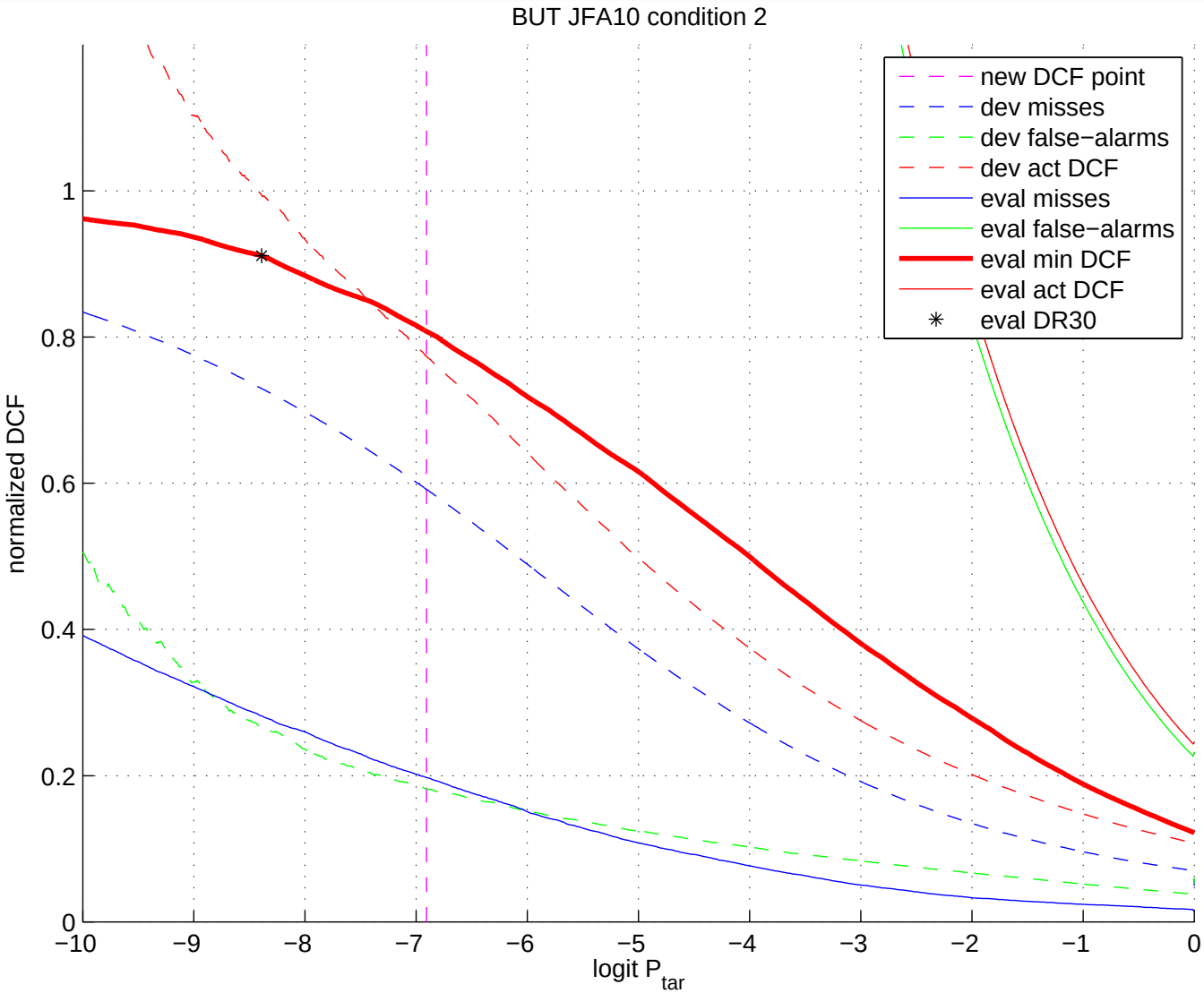
NIST SRE10: Example of Excellent Calibration



NIST SRE10: Example of Bad Calibration



NIST SRE10: Example of Worse Calibration



Outline

DET/ROC, EER and minDCF

We've discussed DCF and its generalization, Bayes error-rate, which are calibration-sensitive evaluation criteria.

Here we are interested in criteria that evaluate the optimal decision-making potential of uncalibrated scores, when calibration is not of immediate interest.

DET/ROC, EER and minDCF are well-known, but we discuss some interesting relationships between them.

DET vs ROC

For the purpose of this discussion DET and ROC are equivalent:

- ROC has: $x = P_{\text{fa}}$, $y = P_{\text{miss}}$
- DET has: $x = \text{probit}(P_{\text{fa}})$, $y = \text{probit}(P_{\text{miss}})$

where the probit function is the inverse of the normal CDF.

ROC and minDCF

For decision threshold γ, denote:

$$\mathrm{DCF}(\gamma|\pi, C_{\mathrm{miss}}, C_{\mathrm{fa}}) = \pi C_{\mathrm{miss}} P_{\mathrm{miss}}(\gamma) + (1-\pi) C_{\mathrm{fa}} P_{\mathrm{miss}}(\gamma)$$

and

$$\mathrm{minDCF}(\pi, C_{\mathrm{miss}}, C_{\mathrm{fa}}) = \min_{\gamma} \mathrm{DCF}(\gamma|\pi, C_{\mathrm{miss}}, C_{\mathrm{fa}})$$

ROCCH and minDCF

A point, x, y is on the:

Steppy ROC curve, iff

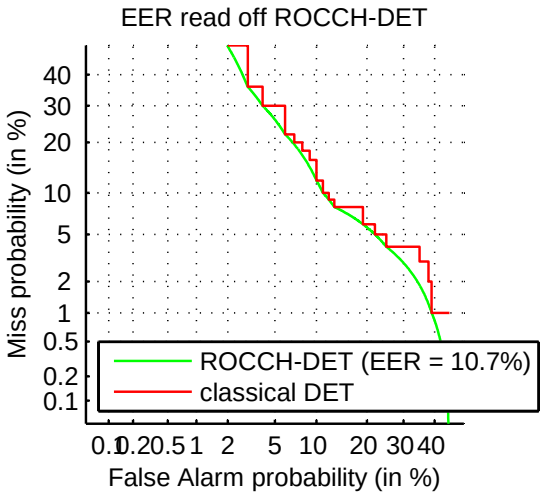
$$y = \frac{C_{\text{miss}}}{C_{\text{fa}}} x = \frac{1}{C_{\text{fa}}} \min_{\gamma} \max_{\pi} \text{DCF}(\gamma|\pi, C_{\text{miss}}, C_{\text{fa}})$$

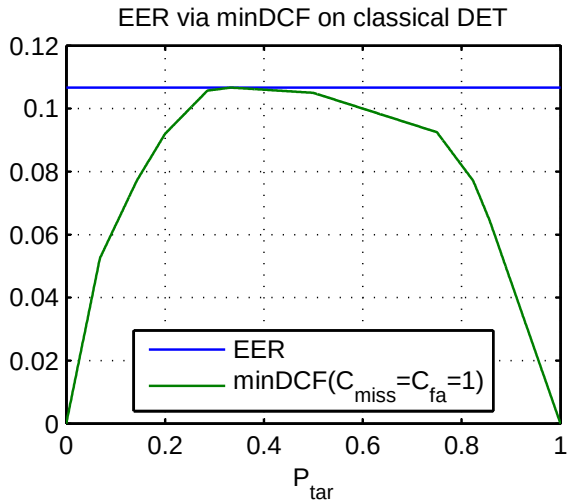
ROCCH curve, iff

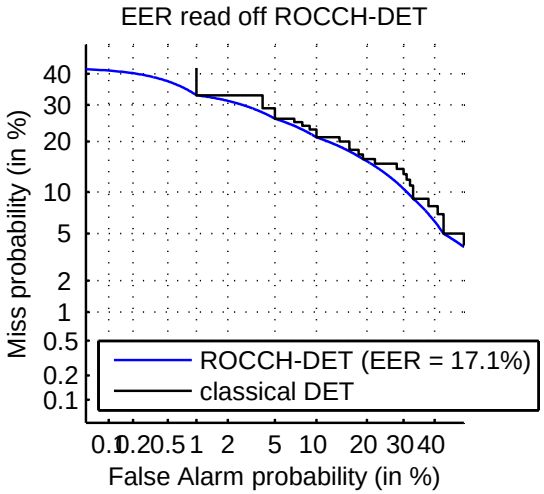
$$y = \frac{C_{\text{miss}}}{C_{\text{fa}}} x = \frac{1}{C_{\text{fa}}} \max_{\pi} \text{minDCF}(\pi, C_{\text{miss}}, C_{\text{fa}})$$

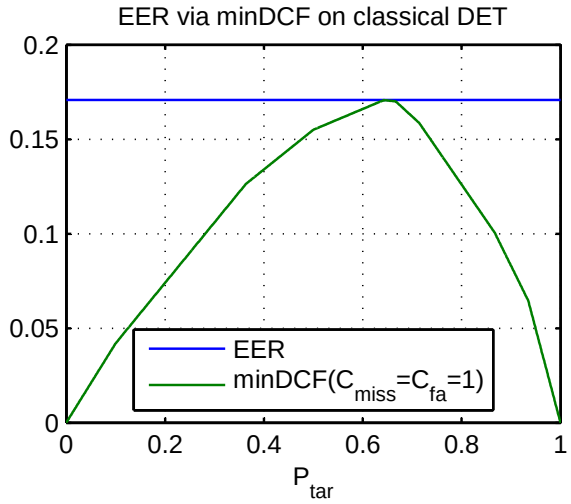
Both curves are generated by sweeping $\frac{C_{\text{miss}}}{C_{\text{fa}}}$ from 0 to ∞.

Examples: Steppy vs ROCCH DET









ROCCH and minDCF

Note:

- minDCF lives on the smooth ROCCH (ROC convex hull) curve, not exactly on the steppy ROC.
- The ROCCH curve is a tight lower bound for the steppy ROC (examples below).
- For large databases, the steps are tiny and the curves are close.

The ROCCH curve has many attractive theoretical and practical properties (see full paper). Here we conclude with one final comment.

ROCCH and EER

Final comment

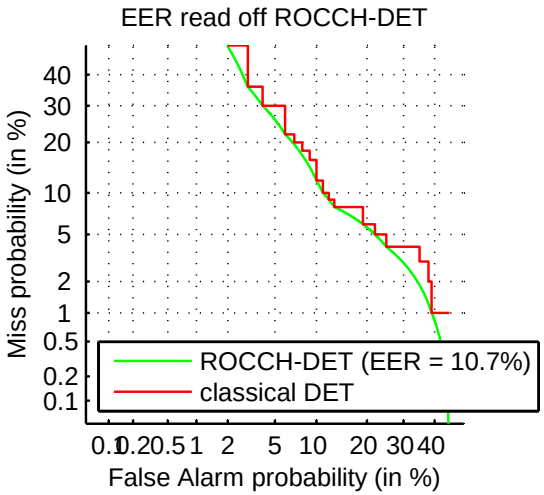
Let the error-rate-ratio be $R = \frac{P_{\text{miss}}}{P_{\text{fa}}}$. Point x, y is on the ROCCH curve, iff

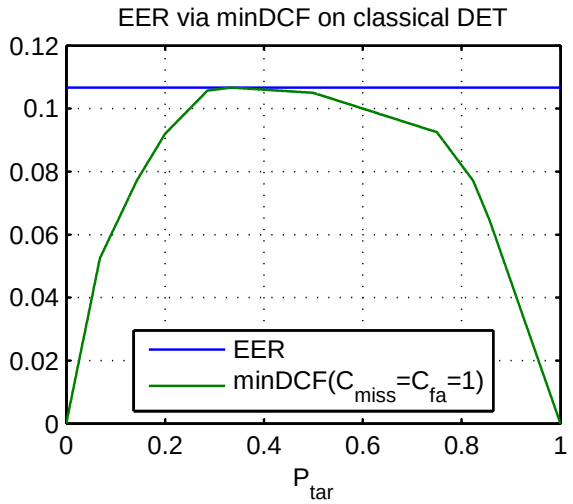
$$y = Rx = \max_{\pi} \text{minDCF}(\pi, C_{\text{miss}} = R, C_{\text{fa}} = 1)$$

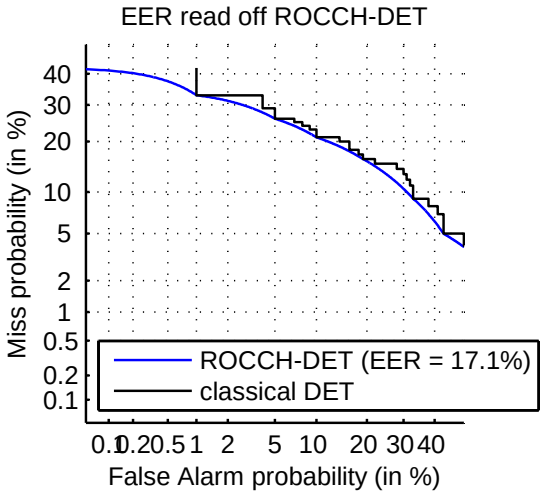
- The location on the ROCCH curve is parametrized by R.
- The special case, $R = 1$, gives the EER (equal-error-rate).
- In general, any point on the ROCCH curve can be interpreted as a tight upper bound for minDCF, over a range of operating points. This makes any such point an attractive objective criterion for system optimization.

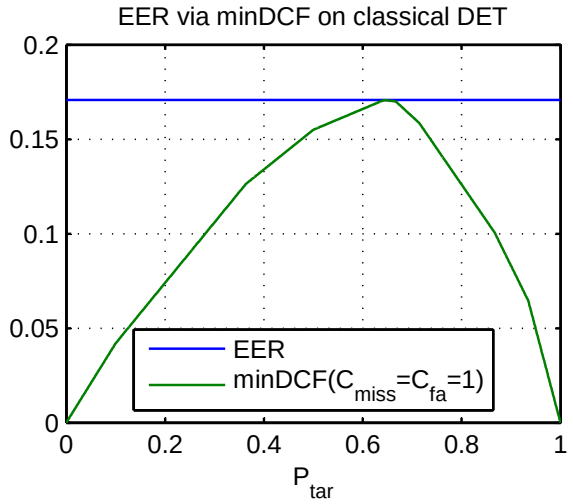
See examples below.

Examples: EER = max minDCF









Conclusion

If you are a MATLAB fan, go and try the BOSARIS Toolkit.

If not, I hope the full paper has enough detail for others to implement some of these ideas in other languages.