

AERFAI-2026 Summer School, July-2026

Error Prediction Toolkit

Error Estimation in Pattern Recognition



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA



Pablo Gandía

With Enrique Vidal and Alejandro H. Toselli

Toolkit presentation

<https://github.com/PRHLT/unsupervisedErrorPrediction>

Running One Experiment

Installation

```
git clone https://github.com/PRHLT/\
unsupervisedErrorPrediction
cd unsupervisedErrorPrediction

chmod +x install.sh toolkit.sh
./install.sh

estimator --help
```

Installer: checks Python, creates a virtual environment, installs NumPy/SciPy/Matplotlib, and makes the estimator command available in the shell.

income example

```
estimator lm data/income/calib.out \
data/income/eval.out results 10
```

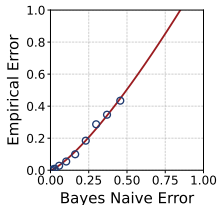
optional flags

```
estimator lm calib eval results 10 \
--remain --trim:1.5
```

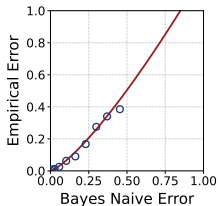
- lm: Levenberg-Marquardt optimizer.
- sx: Simplex/Nelder-Mead optimizer.
- ts: temperature scaling, optimizing T .
- -remain: keeps grouping leftovers as a final group for fit and evaluation.
- -trim: refits after removing calibration groups with large residuals.

Reading a Toolkit Result

Calibration



Evaluation



income, 1m, $K = 10$

Metric	Toolkit print	Interpretation
$E = 13.67$	E	empirical task error
$\hat{E} = 14.61$	Ecal	predicted task error after calibration
$E_e = 17.17$	Ee	raw Bayes Naive Error
$\Delta E = 0.95$	DE	global absolute prediction error
$r\Delta E = 6.92\%$	rDE	relative global prediction error
$\delta E = 1.35$	dE	local average prediction error
$r\delta E = 9.85\%$	rdE	relative local average error

Read the pair together

Upper: calibration learns the curve.

Lower: evaluation checks the same curve on unseen groups.

Input and Output

Input format

Calibration and evaluation files format (.out, .txt, ...)

Each line contains one sample

<E> <1-Pmax> <y_true> <y_pred>

<P0> <P1> ... <Pn>

Example

0 0.0527 9 9 0.0105 ... 0.9473

1 0.0481 9 8 0.0001 ... 0.0479

`lm/sx`: can also run with only <E> <1-Pmax>.

<E>	0 correct, 1 wrong.
<1-Pmax>	estimated error, or Bayes Naive Error.
<y_true>	true class index.
<y_pred>	predicted class index.
<P0>...<Pn>	softmax probabilities, sum to 1.

Terminal print

```
Results for experiment with K=10:
-----
Calibration:
Grouped into blocks of size M=753 with remainder r=0

Evaluation:
Grouped into blocks of size M=753 with remainder r=0

Fit results:
Fit parameters --> a=1.2359, b=1.2928
WSSR --> 0.3279

Ecal: 14.6108
Ee: 17.1661
E: 13.6653
DE: 0.9455
dE: 1.3461
rDE: 6.9188
rdE: 9.8505
```

Fit model: $\hat{E} = f(\mathcal{E}) = a\mathcal{E}^b$

$\mathcal{E}$: Bayes Naive Error; a : scale, b : exponent.

The full run is saved in an experiment folder.

Files Written

```
results/lm_K10_t0_tr0_lim0.0/  
  grouped/  
    calibrationK10  
    evaluationK10  
  predictions/  
    predictionsK10  
  plots/  
    SimplePlotCalibK10.pdf  
    SimplePlotEvalK10.pdf  
  scripts/  
    Calib_plot_scriptK10.py  
    Eval_plot_scriptK10.py  
  metadata.json
```

- grouped/: points used in the plots.
- predictions/: one calibrated prediction per evaluation group.
- plots/: generated PDF figures.
- scripts/: editable scripts to regenerate the plots.
- metadata.json: inputs, options, parameters and final metrics.