Evaluating Classifier Performance

Tilmann Gneiting
Karlsruhe Institute of Technology (KIT) and Heidelberg Institute for Theoretical Studies (HITS)

Peter Vogel
CSL Behring, Marburg

GridKa School 2019 — The Art of Data

26 August 2019

Evaluating classifier performance

1. Predicting a binary event
2. Receiver operating characteristic (ROC) curves:
   What are they, and what are they good for?
3. Tools for evaluating probabilistic classifiers: A triptych

Evaluating classifier performance

1. Predicting a binary event
2. Receiver operating characteristic (ROC) curves:
   What are they, and what are they good for?
3. Tools for evaluating probabilistic classifiers: A triptych

Predicting a binary event

Predicting a binary event

consider the (presumably) simplest task in forecasting, namely, the prediction of a binary event, such as

- precipitation tomorrow
- recession
- Champions League final
- medical diagnosis
- recidivism
- 2020 presidential election

Predicting a binary event

consider the (presumably) simplest task in forecasting, namely, the prediction of a binary event, such as

- precipitation tomorrow
- recession
- Champions League final
- medical diagnosis
- recidivism
- 2020 presidential election

we let $Y = 1$ denote a positive and $Y = 0$ a negative outcome

Predicting a binary event

consider the (presumably) simplest task in forecasting, namely, the prediction of a binary event, such as

- precipitation tomorrow
- recession
- Champions League final
- medical diagnosis
- recidivism
- 2020 presidential election

we let $Y = 1$ denote a positive and $Y = 0$ a negative outcome

a forecast for Y might take any of the following forms:

Predicting a binary event

consider the (presumably) simplest task in forecasting, namely, the prediction of a binary event, such as

- precipitation tomorrow
- recession
- Champions League final
- medical diagnosis
- recidivism
- 2020 presidential election

we let $Y = 1$ denote a positive and $Y = 0$ a negative outcome

a forecast for Y might take any of the following forms:

- a hard classifier predicts a positive (1) or negative (0) outcome

Predicting a binary event

consider the (presumably) simplest task in forecasting, namely, the prediction of a binary event, such as

- precipitation tomorrow
- recession
- Champions League final
- medical diagnosis
- recidivism
- 2020 presidential election

we let $Y = 1$ denote a positive and $Y = 0$ a negative outcome

a forecast for Y might take any of the following forms:

- a hard classifier predicts a positive (1) or negative (0) outcome
- a probabilistic classifier specifies a probability $p \in [0, 1]$ for a positive outcome

Predicting a binary event

consider the (presumably) simplest task in forecasting, namely, the prediction of a binary event, such as

- precipitation tomorrow
- recession
- Champions League final
- medical diagnosis
- recidivism
- 2020 presidential election

we let $Y = 1$ denote a positive and $Y = 0$ a negative outcome

a forecast for Y might take any of the following forms:

- a hard classifier predicts a positive (1) or negative (0) outcome
- a probabilistic classifier specifies a probability $p \in [0, 1]$ for a positive outcome
- a feature provides a value $x \in \mathbb{R}$, with the understanding that the larger the feature, the more likely a positive outcome

Example: Precipitation forecasts over Africa

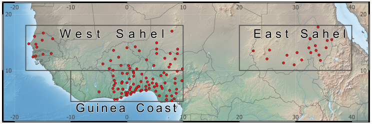
Example: Precipitation forecasts over Africa

key example in this presentation

Example: Precipitation forecasts over Africa

key example in this presentation

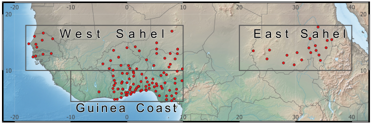
- 24-hour ahead precipitation forecasts over northern tropical Africa (Vogel et al. 2018)



Example: Precipitation forecasts over Africa

key example in this presentation

- 24-hour ahead precipitation forecasts over northern tropical Africa (Vogel et al. 2018)

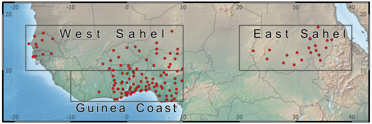


- based on the numerical weather prediction (NWP) ensemble operated by the European Centre for Medium-Range Weather Forecasts (ECMWF)

Example: Precipitation forecasts over Africa

key example in this presentation

- 24-hour ahead precipitation forecasts over northern tropical Africa (Vogel et al. 2018)

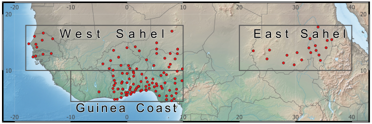


- based on the numerical weather prediction (NWP) ensemble operated by the European Centre for Medium-Range Weather Forecasts (ECMWF)
- the ensemble has 50 members, each of which provides a hard classifier

Example: Precipitation forecasts over Africa

key example in this presentation

- 24-hour ahead precipitation forecasts over northern tropical Africa (Vogel et al. 2018)

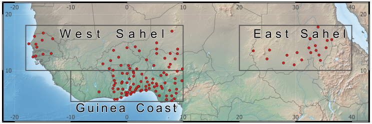


- based on the numerical weather prediction (NWP) ensemble operated by the European Centre for Medium-Range Weather Forecasts (ECMWF)
- the ensemble has 50 members, each of which provides a hard classifier
- the resulting probability of precipitation (PoP) forecast equals the fraction of members that predict precipitation

Example: Precipitation forecasts over Africa

key example in this presentation

- 24-hour ahead precipitation forecasts over northern tropical Africa (Vogel et al. 2018)

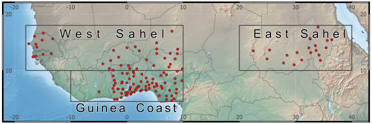


- based on the numerical weather prediction (NWP) ensemble operated by the European Centre for Medium-Range Weather Forecasts (ECMWF)
- the ensemble has 50 members, each of which provides a hard classifier
- the resulting probability of precipitation (PoP) forecast equals the fraction of members that predict precipitation
- our interest is in comparing the ECMWF PoP forecast to statistically postprocessed and climatological (i.e., purely observation based) forecasts

Example: Precipitation forecasts over Africa

key example in this presentation

- 24-hour ahead precipitation forecasts over northern tropical Africa (Vogel et al. 2018)



- based on the numerical weather prediction (NWP) ensemble operated by the European Centre for Medium-Range Weather Forecasts (ECMWF)
- the ensemble has 50 members, each of which provides a hard classifier
- the resulting probability of precipitation (PoP) forecast equals the fraction of members that predict precipitation
- our interest is in comparing the ECMWF PoP forecast to statistically postprocessed and climatological (i.e., purely observation based) forecasts

Vogel, P., Knippertz, P., Fink, A. H., Schlueter, A. and Gneiting, T. (2018). **Skill of global raw and postprocessed ensemble predictions of rainfall over northern tropical Africa.** *Weather and Forecasting*, 33, 369–388.

Evaluating classifier performance

1. Predicting a binary event
2. Receiver operating characteristic (ROC) curves:
   What are they, and what are they good for?
3. Tools for evaluating probabilistic classifiers: A triptych

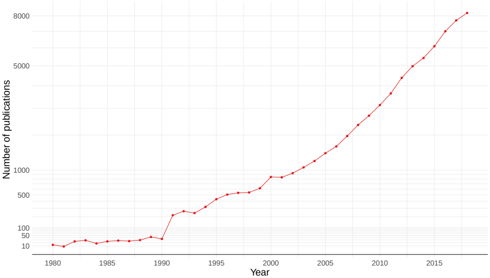
Receiver operating characteristic (ROC) curves

Receiver operating characteristic (ROC) curves

receiver operating characteristic (ROC) curves are outrageously popular tools for evaluating features (in general) and probabilistic classifiers (in particular)

Receiver operating characteristic (ROC) curves

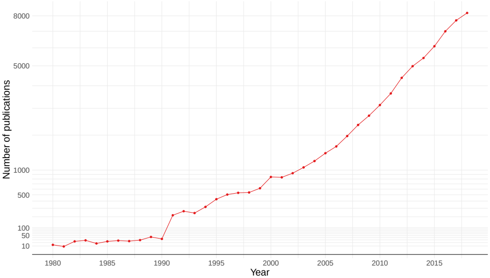
receiver operating characteristic (ROC) curves are outrageously popular tools for evaluating features (in general) and probabilistic classifiers (in particular)



results from a Web of Science `topic` search for `‘‘receiver operating characteristic’’ or ‘‘ROC’’`

Receiver operating characteristic (ROC) curves

receiver operating characteristic (ROC) curves are outrageously popular tools for evaluating features (in general) and probabilistic classifiers (in particular)

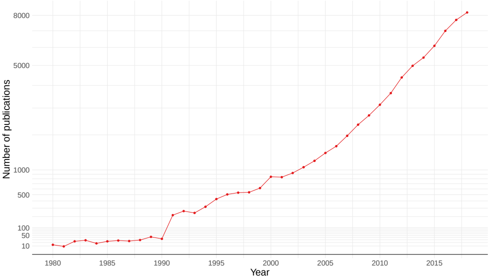


results from a Web of Science `topic` search for `‘‘receiver operating characteristic’’ or ‘‘ROC’’`

proposed in signal detection (Egan, Greenberg and Schulman 1961) and cognitive psychology (Swets 1973)

Receiver operating characteristic (ROC) curves

receiver operating characteristic (ROC) curves are outrageously popular tools for evaluating features (in general) and probabilistic classifiers (in particular)



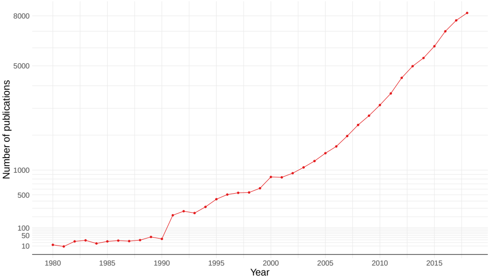
results from a Web of Science `topic` search for `‘‘receiver operating characteristic’’ or ‘‘ROC’’`

proposed in signal detection (Egan, Greenberg and Schulman 1961) and cognitive psychology (Swets 1973)

vastly popular ever since Swets (*Science*, 1988),

Receiver operating characteristic (ROC) curves

receiver operating characteristic (ROC) curves are outrageously popular tools for evaluating features (in general) and probabilistic classifiers (in particular)



results from a Web of Science `topic` search for `‘‘receiver operating characteristic’’ or ‘‘ROC’’`

proposed in signal detection (Egan, Greenberg and Schulman 1961) and cognitive psychology (Swets 1973)

vastly popular ever since Swets (*Science*, 1988), particularly in the life sciences and machine learning

Receiver operating characteristic (ROC) curves

Receiver operating characteristic (ROC) curves

generally, we can use any threshold value x to construct a hard classifier for the binary event Y from the feature X

Receiver operating characteristic (ROC) curves

generally, we can use any threshold value x to construct a hard classifier for the binary event Y from the feature X

- if $X > x$ we predict a positive outcome ($Y = 1$)

Receiver operating characteristic (ROC) curves

generally, we can use any threshold value x to construct a hard classifier for the binary event Y from the feature X

- if $X > x$ we predict a positive outcome ($Y = 1$)
- if $X \leq x$ we predict a negative outcome ($Y = 0$)

Receiver operating characteristic (ROC) curves

generally, we can use any threshold value x to construct a hard classifier for the binary event Y from the feature X

- if $X > x$ we predict a positive outcome ($Y = 1$)
- if $X \leq x$ we predict a negative outcome ($Y = 0$)

in practice,

Receiver operating characteristic (ROC) curves

generally, we can use any threshold value x to construct a hard classifier for the binary event Y from the feature X

- if $X > x$ we predict a positive outcome ($Y = 1$)
- if $X \leq x$ we predict a negative outcome ($Y = 0$)

in practice,

- we are given a dataset

$$\{(x_i, y_i) : i = 1, \ldots, n\},$$

  where $x_i \in \mathbb{R}$ is the feature and $y_i \in \{0, 1\}$ the associated binary outcome

Receiver operating characteristic (ROC) curves

generally, we can use any threshold value x to construct a hard classifier for the binary event Y from the feature X

- if $X > x$ we predict a positive outcome ($Y = 1$)
- if $X \leq x$ we predict a negative outcome ($Y = 0$)

in practice,

- we are given a dataset

$$\{(x_i, y_i) : i = 1, \ldots, n\},$$

  where $x_i \in \mathbb{R}$ is the feature and $y_i \in \{0, 1\}$ the associated binary outcome
- by thresholding at $x \in \mathbb{R}$, we get a hard classifier

Receiver operating characteristic (ROC) curves

generally, we can use any threshold value x to construct a hard classifier for the binary event Y from the feature X

- if $X > x$ we predict a positive outcome ($Y = 1$)
- if $X \leq x$ we predict a negative outcome ($Y = 0$)

in practice,

- we are given a dataset

  $$\{(x_i, y_i) : i = 1, \ldots, n\},$$

  where $x_i \in \mathbb{R}$ is the feature and $y_i \in \{0, 1\}$ the associated binary outcome
- by thresholding at $x \in \mathbb{R}$, we get a hard classifier
- with true positive ($\mathrm{TP} : x_i > x, \ \ y_i = 1$), true negative ($\mathrm{TN} : x_i < x, \ \ y_i = 0$), false positive ($\mathrm{FP} : x_i > x, \ \ y_i = 0$) and false negative ($\mathrm{FN} : x_i < x, \ \ y_i = 1$) instances in the dataset

Receiver operating characteristic (ROC) curves

Receiver operating characteristic (ROC) curves

let P and N denote the overall numbers of positive and negative cases in the dataset

Receiver operating characteristic (ROC) curves

let P and N denote the overall numbers of positive and negative cases in the dataset

given any threshold value x,

Receiver operating characteristic (ROC) curves

let P and N denote the overall numbers of positive and negative cases in the dataset

given any threshold value x,

- let $\mathrm{TP}(x)$, $\mathrm{FN}(x)$, $\mathrm{FP}(x)$ and $\mathrm{TN}(x)$ denote the respective numbers of true positive, false negative, false positive and true negative cases

Receiver operating characteristic (ROC) curves

let P and N denote the overall numbers of positive and negative cases in the dataset

given any threshold value x,

- let $\mathrm{TP}(x)$, $\mathrm{FN}(x)$, $\mathrm{FP}(x)$ and $\mathrm{TN}(x)$ denote the respective numbers of true positive, false negative, false positive and true negative cases
- the hit rate (HR) at x is

$$\mathrm{HR}(x) = \frac{\mathrm{TP}(x)}{\mathrm{TP}(x) + \mathrm{FN}(x)} = \frac{\mathrm{TP}(x)}{\mathrm{P}}$$

Receiver operating characteristic (ROC) curves

let P and N denote the overall numbers of positive and negative cases in the dataset

given any threshold value x,

- let $\mathrm{TP}(x)$, $\mathrm{FN}(x)$, $\mathrm{FP}(x)$ and $\mathrm{TN}(x)$ denote the respective numbers of true positive, false negative, false positive and true negative cases
- the hit rate (HR) at x is

$$\mathrm{HR}(x) = \frac{\mathrm{TP}(x)}{\mathrm{TP}(x) + \mathrm{FN}(x)} = \frac{\mathrm{TP}(x)}{\mathrm{P}}$$

- the false alarm rate (FAR) at x is

$$\mathrm{FAR}(x) = \frac{\mathrm{FP}(x)}{\mathrm{FP}(x) + \mathrm{TN}(x)} = \frac{\mathrm{FP}(x)}{\mathrm{N}}$$

Receiver operating characteristic (ROC) curves

let P and N denote the overall numbers of positive and negative cases in the dataset

given any threshold value x,

- let $\mathrm{TP}(x)$, $\mathrm{FN}(x)$, $\mathrm{FP}(x)$ and $\mathrm{TN}(x)$ denote the respective numbers of true positive, false negative, false positive and true negative cases
- the hit rate (HR) at x is

$$\mathrm{HR}(x) = \frac{\mathrm{TP}(x)}{\mathrm{TP}(x) + \mathrm{FN}(x)} = \frac{\mathrm{TP}(x)}{\mathrm{P}}$$

- the false alarm rate (FAR) at x is

$$\mathrm{FAR}(x) = \frac{\mathrm{FP}(x)}{\mathrm{FP}(x) + \mathrm{TN}(x)} = \frac{\mathrm{FP}(x)}{\mathrm{N}}$$

the ROC curve plots the hit rate $\mathrm{HR}(x)$ vs. the false alarm rate $\mathrm{FAR}(x)$ as the decision threshold x varies

Receiver operating characteristic (ROC) curves

let P and N denote the overall numbers of positive and negative cases in the dataset

given any threshold value x,

- let $\mathrm{TP}(x)$, $\mathrm{FN}(x)$, $\mathrm{FP}(x)$ and $\mathrm{TN}(x)$ denote the respective numbers of true positive, false negative, false positive and true negative cases
- the hit rate (HR) at x is

$$\mathrm{HR}(x) = \frac{\mathrm{TP}(x)}{\mathrm{TP}(x) + \mathrm{FN}(x)} = \frac{\mathrm{TP}(x)}{\mathrm{P}}$$

- the false alarm rate (FAR) at x is

$$\mathrm{FAR}(x) = \frac{\mathrm{FP}(x)}{\mathrm{FP}(x) + \mathrm{TN}(x)} = \frac{\mathrm{FP}(x)}{\mathrm{N}}$$

the ROC curve plots the hit rate $\mathrm{HR}(x)$ vs. the false alarm rate $\mathrm{FAR}(x)$ as the decision threshold x varies

the area under the ROC curve (AUC) is a positively oriented measure of predictive ability

Area under the ROC curve (AUC)

Area under the ROC curve (AUC)

the area under the ROC curve (AUC) is a positively oriented measure of predictive ability

Area under the ROC curve (AUC)

the area under the ROC curve (AUC) is a positively oriented measure of predictive ability

- appealing interpretation as the probability that a (randomly chosen) feature value under a positive outcome is larger than a (randomly chosen) value under a negative outcome:

$$\mathrm{AUC} = \mathbb{P}(X' > X \mid Y' = 1, Y = 0)$$

Area under the ROC curve (AUC)

the area under the ROC curve (AUC) is a positively oriented measure of predictive ability

- appealing interpretation as the probability that a (randomly chosen) feature value under a positive outcome is larger than a (randomly chosen) value under a negative outcome:

$$\mathrm{AUC} = \mathbb{P}(X' > X \mid Y' = 1, Y = 0)$$

- $\mathrm{AUC} = (D + 1)/2$ in terms of Somers' D (Somers 1962)

Area under the ROC curve (AUC)

the area under the ROC curve (AUC) is a positively oriented measure of predictive ability

- appealing interpretation as the probability that a (randomly chosen) feature value under a positive outcome is larger than a (randomly chosen) value under a negative outcome:

$$\mathrm{AUC} = \mathbb{P}(X' > X \mid Y' = 1, Y = 0)$$

- $\mathrm{AUC} = (D + 1)/2$ in terms of Somers' D (Somers 1962)
- $\mathrm{AUC} = 1/2$ for a useless feature that is independent of the binary outcome

Area under the ROC curve (AUC)

the area under the ROC curve (AUC) is a positively oriented measure of predictive ability

- appealing interpretation as the probability that a (randomly chosen) feature value under a positive outcome is larger than a (randomly chosen) value under a negative outcome:

$$\text{AUC} = \mathbb{P}(X' > X \mid Y' = 1, Y = 0)$$

- $\text{AUC} = (D + 1)/2$ in terms of Somers' D (Somers 1962)
- $\text{AUC} = 1/2$ for a useless feature that is independent of the binary outcome
- $\text{AUC} = 1$ for a perfect feature

Example

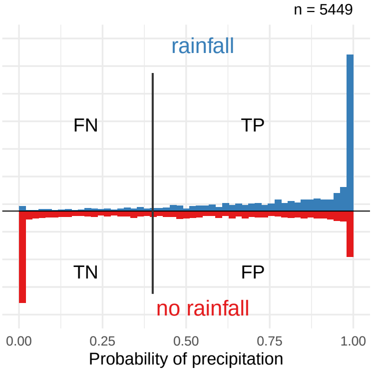
Example

ROC curve and AUC for 24-hour probability of precipitation (PoP) forecasts from the ECMWF ensemble over West Sahel

Example

ROC curve and AUC for 24-hour probability of precipitation (PoP) forecasts from the ECMWF ensemble over West Sahel

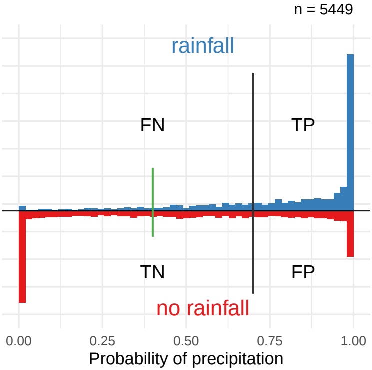
Vogel et al. (2018)



Example

ROC curve and AUC for 24-hour probability of precipitation (PoP) forecasts from the ECMWF ensemble over West Sahel

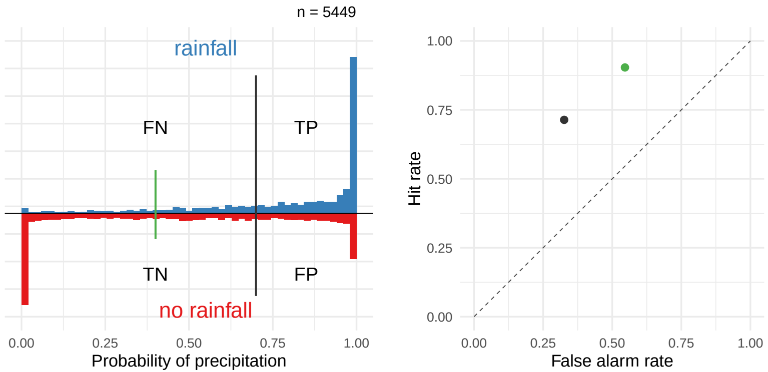
Vogel et al. (2018)



Example

ROC curve and AUC for 24-hour probability of precipitation (PoP) forecasts from the ECMWF ensemble over West Sahel

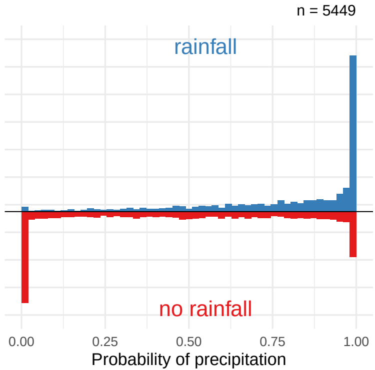
Vogel et al. (2018)



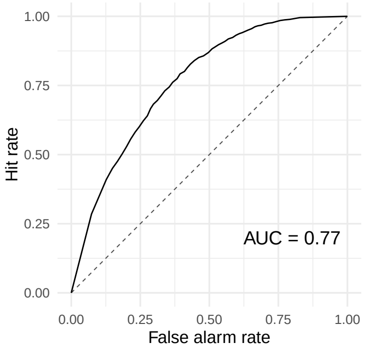
Example

ROC curve and AUC for 24-hour probability of precipitation (PoP) forecasts from the ECMWF ensemble over West Sahel

Vogel et al. (2018)



ROC curve



A formal approach to ROC curves

A formal approach to ROC curves

formal setting

X	real-valued feature
Y	binary outcome
$\mathbb{P}$	joint distribution of (X, Y)

$$F_1(x) = \mathbb{P}(X \leq x \mid Y = 1)$$
$$F_0(x) = \mathbb{P}(X \leq x \mid Y = 0)$$

A formal approach to ROC curves

formal setting

X real-valued feature
Y binary outcome
$\mathbb{P}$ joint distribution of (X, Y)

$$F_1(x) = \mathbb{P}(X \le x \mid Y = 1)$$
$$F_0(x) = \mathbb{P}(X \le x \mid Y = 0)$$

the raw ROC diagnostic is the set of all points of the form

$$(\mathrm{FAR}(x), \mathrm{HR}(x)) \in [0,1] \times [0,1]$$

threshold $x \in \mathbb{R}$, $\mathrm{FAR}(x) = 1 - F_0(x)$, $\mathrm{HR}(x) = 1 - F_1(x)$

A formal approach to ROC curves

formal setting

X real-valued feature
Y binary outcome
$\mathbb{P}$ joint distribution of (X, Y)

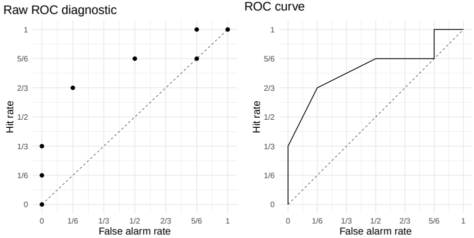
$$F_1(x) = \mathbb{P}(X \leq x \,|\, Y = 1)$$
$$F_0(x) = \mathbb{P}(X \leq x \,|\, Y = 0)$$

the raw ROC diagnostic is the set of all points of the form

$$(\mathrm{FAR}(x), \mathrm{HR}(x)) \in [0,1] \times [0,1]$$

$$\text{threshold } x \in \mathbb{R}, \quad \mathrm{FAR}(x) = 1 - F_0(x), \quad \mathrm{HR}(x) = 1 - F_1(x)$$

the ROC curve is the linearly interpolated raw ROC diagnostic

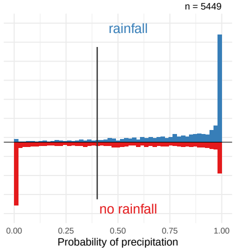


Properties of ROC curves and AUC

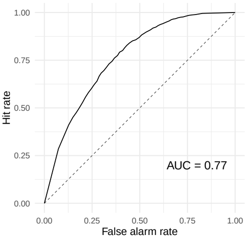
interpretation as function: for continuous, strictly increasing F_0 and F_1,

$$R(\alpha) = 1 - F_1(F_0^{-1}(1-\alpha)), \quad \alpha = \mathrm{FAR}(x) \in (0,1)$$

Vogel et al. (2018)



ROC curve



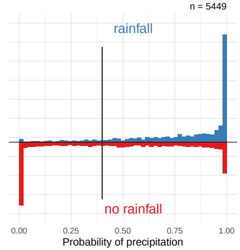
Properties of ROC curves and AUC

interpretation as function: for continuous, strictly increasing F_0 and F_1,

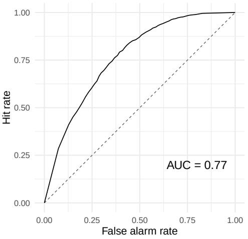
$$R(\alpha) = 1 - F_1(F_0^{-1}(1 - \alpha)), \quad \alpha = \mathrm{FAR}(x) \in (0, 1)$$

ensuing math fact: characterization of ROC curves

Vogel et al. (2018)



ROC curve



Properties of ROC curves and AUC

interpretation as function: for continuous, strictly increasing F_0 and F_1,

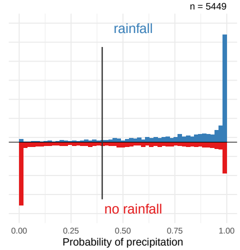
$$R(\alpha) = 1 - F_1(F_0^{-1}(1-\alpha)), \quad \alpha = \mathrm{FAR}(x) \in (0,1)$$

ensuing math fact: characterization of ROC curves

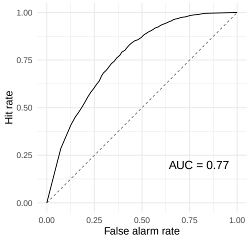
invariance of ROC curves and AUC under

- changes in class proportions
- strictly increasing transformations of the feature X

Vogel et al. (2018)



ROC curve



Properties of ROC curves and AUC

interpretation as function: for continuous, strictly increasing F_0 and F_1,

$$R(\alpha) = 1 - F_1(F_0^{-1}(1-\alpha)), \quad \alpha = \mathrm{FAR}(x) \in (0,1)$$

ensuing math fact: characterization of ROC curves

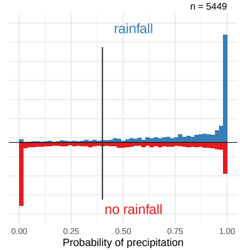
invariance of ROC curves and AUC under

- changes in class proportions
- strictly increasing transformations of the feature X

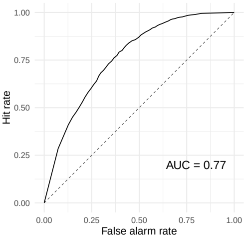
consequence: ROC curves and AUC

- apply to real-valued or ordinal features X on arbitrary scales, but
- do not consider calibration nor economic value for probabilistic classifiers

Vogel et al. (2018)



ROC curve



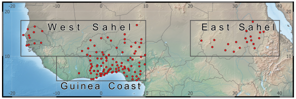
Back to our example

Back to our example

24-hour precipitation forecasts over the West Sahel region in northern tropical Africa in monsoon season 2014 (Vogel et al. 2018)

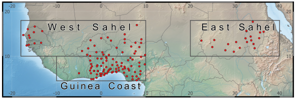
Back to our example

24-hour precipitation forecasts over the West Sahel region in northern tropical Africa in monsoon season 2014 (Vogel et al. 2018)



Back to our example

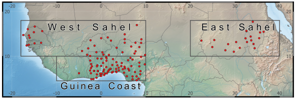
24-hour precipitation forecasts over the West Sahel region in northern tropical Africa in monsoon season 2014 (Vogel et al. 2018)



competing probability of precipitation (PoP) forecasts

Back to our example

24-hour precipitation forecasts over the West Sahel region in northern tropical Africa in monsoon season 2014 (Vogel et al. 2018)

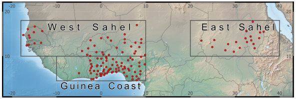


competing probability of precipitation (PoP) forecasts

- ENS ECMWF NWP ensemble

Back to our example

24-hour precipitation forecasts over the West Sahel region in northern tropical Africa in monsoon season 2014 (Vogel et al. 2018)

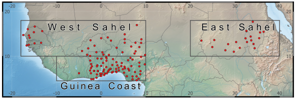


competing probability of precipitation (PoP) forecasts

- ENS ECMWF NWP ensemble
- EMOS postprocessed by ensemble model output statistics

Back to our example

24-hour precipitation forecasts over the West Sahel region in northern tropical Africa in monsoon season 2014 (Vogel et al. 2018)

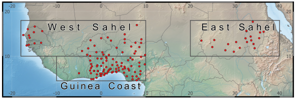


competing probability of precipitation (PoP) forecasts

- ENS ECMWF NWP ensemble
- EMOS postprocessed by ensemble model output statistics
- BMA postprocessed by Bayesian model averaging

Back to our example

24-hour precipitation forecasts over the West Sahel region in northern tropical Africa in monsoon season 2014 (Vogel et al. 2018)

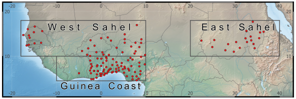


competing probability of precipitation (PoP) forecasts

- ENS ECMWF NWP ensemble
- EMOS postprocessed by ensemble model output statistics
- BMA postprocessed by Bayesian model averaging
- EPC extended probabilistic climatology

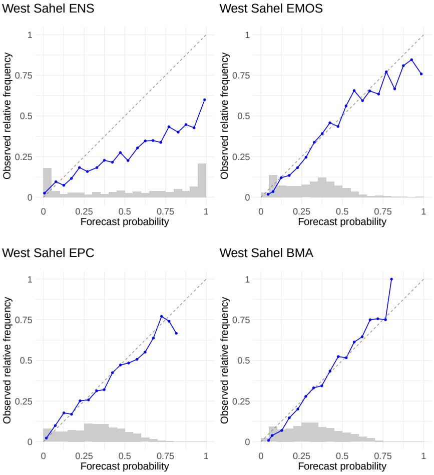
Back to our example

24-hour precipitation forecasts over the West Sahel region in northern tropical Africa in monsoon season 2014 (Vogel et al. 2018)



competing probability of precipitation (PoP) forecasts

- **ENS** ECMWF NWP ensemble
- **EMOS** postprocessed by ensemble model output statistics
- **BMA** postprocessed by Bayesian model averaging
- **EPC** extended probabilistic climatology



Evaluating classifier performance

1. Predicting a binary event
2. Receiver operating characteristic (ROC) curves:
   What are they, and what are they good for?
3. Tools for evaluating probabilistic classifiers: A triptych

ROC curves address potential predictive ability (only)

invariance under strictly monotone transformations has stark implications:

ROC curves address potential predictive ability (only)

invariance under strictly monotone transformations has stark implications:

- ROC curves and AUC can be used to assess the potential predictive ability of any real-valued feature

ROC curves address potential predictive ability (only)

invariance under strictly monotone transformations has stark implications:

- ROC curves and AUC can be used to assess the potential predictive ability of any real-valued feature
- however, in the case of probabilistic classifiers, both calibration and actual value are ignored

ROC curves address potential predictive ability (only)

invariance under strictly monotone transformations has stark implications:

- ROC curves and AUC can be used to assess the potential predictive ability of any real-valued feature
- however, in the case of probabilistic classifiers, both calibration and actual value are ignored
- hence, ROC curves should be used in concert with reliability diagrams and Murphy diagrams

ROC curves address potential predictive ability (only)

invariance under strictly monotone transformations has stark implications:

- ROC curves and AUC can be used to assess the potential predictive ability of any real-valued feature
- however, in the case of probabilistic classifiers, both calibration and actual value are ignored
- hence, ROC curves should be used in concert with reliability diagrams and Murphy diagrams

reliability diagrams

- assess calibration, by plotting the empirical conditional event frequency vs. the predicted probability

ROC curves address potential predictive ability (only)

invariance under strictly monotone transformations has stark implications:

- ROC curves and AUC can be used to assess the potential predictive ability of any real-valued feature
- however, in the case of probabilistic classifiers, both calibration and actual value are ignored
- hence, ROC curves should be used in concert with reliability diagrams and Murphy diagrams

reliability diagrams

- assess calibration, by plotting the empirical conditional event frequency vs. the predicted probability

Murphy diagrams

- assess actual economic value, by considering all (!) proper scoring rules simultaneously

Proper scoring rules

Proper scoring rules

a scoring rule is a function $S(p, y)$ that assigns a numerical score to a probability forecast p with binary outcome y

Proper scoring rules

a scoring rule is a function $S(p, y)$ that assigns a numerical score to a probability forecast p with binary outcome y

we take scores to be negatively oriented: the smaller, the better

Proper scoring rules

a scoring rule is a function $\mathrm{S}(p, y)$ that assigns a numerical score to a probability forecast p with binary outcome y

we take scores to be negatively oriented: the smaller, the better

a proper scoring rule rewards honest and careful forecasting: truth telling is the best strategy in expectation

Proper scoring rules

a scoring rule is a function $\mathrm{S}(p, y)$ that assigns a numerical score to a probability forecast p with binary outcome y

we take scores to be negatively oriented: the smaller, the better

a proper scoring rule rewards honest and careful forecasting: truth telling is the best strategy in expectation

every proper scoring rule can be represented as a weighted mixture over elementary scores

$$\mathrm{S}_\theta(p, y) = \begin{cases} \theta, & y = 0,\ p > \theta, \\ 1 - \theta, & y = 1,\ p \le \theta, \\ 0, & \text{otherwise}, \end{cases}$$

where the index $\theta \in (0, 1)$ reflects a decision maker's cost ratio; e.g.

Proper scoring rules

a scoring rule is a function $S(p, y)$ that assigns a numerical score to a probability forecast p with binary outcome y

we take scores to be negatively oriented: the smaller, the better

a proper scoring rule rewards honest and careful forecasting: truth telling is the best strategy in expectation

every proper scoring rule can be represented as a weighted mixture over elementary scores

$$S_\theta(p, y) = \begin{cases} \theta, & y = 0,\ p > \theta, \\ 1 - \theta, & y = 1,\ p \leq \theta, \\ 0, & \text{otherwise}, \end{cases}$$

where the index $\theta \in (0, 1)$ reflects a decision maker's cost ratio; e.g.

- for the quadratic or Brier score, $S(p, y) = (p - y)^2$, the mixture is (twice) uniform over $\theta \in (0, 1)$

Proper scoring rules

a scoring rule is a function $\mathrm{S}(p, y)$ that assigns a numerical score to a probability forecast p with binary outcome y

we take scores to be negatively oriented: the smaller, the better

a proper scoring rule rewards honest and careful forecasting: truth telling is the best strategy in expectation

every proper scoring rule can be represented as a weighted mixture over elementary scores

$$\mathrm{S}_\theta(p, y) = \begin{cases} \theta, & y = 0,\ p > \theta, \\ 1 - \theta, & y = 1,\ p \le \theta, \\ 0, & \text{otherwise}, \end{cases}$$

where the index $\theta \in (0, 1)$ reflects a decision maker's cost ratio; e.g.

- for the quadratic or Brier score, $\mathrm{S}(p, y) = (p - y)^2$, the mixture is (twice) uniform over $\theta \in (0, 1)$
- for the logarithmic score, $\mathrm{S}(p, y) = -\log p$, the mixture measure has density $\propto (\theta(1 - \theta))^{-1}$

Murphy diagrams

Murphy diagrams

a Murphy curve plots the mean elementary score S_θ of a probabilistic classifier as a function of $\theta \in (0, 1)$

Murphy diagrams

a Murphy curve plots the mean elementary score S_θ of a probabilistic classifier as a function of $\theta \in (0, 1)$

- the Brier score equals (twice) the area under the Murphy curve

Murphy diagrams

a Murphy curve plots the mean elementary score S_θ of a probabilistic classifier as a function of $\theta \in (0, 1)$

- the Brier score equals (twice) the area under the Murphy curve
- the widely reported misclassification rate equals (twice) the height of the Murphy curve at $\theta = 1/2$

Murphy diagrams

a Murphy curve plots the mean elementary score S_θ of a probabilistic classifier as a function of $\theta \in (0, 1)$

- the Brier score equals (twice) the area under the Murphy curve
- the widely reported misclassification rate equals (twice) the height of the Murphy curve at $\theta = 1/2$

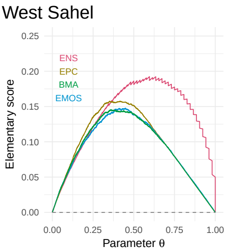
a Murphy diagram plots the curves of competing forecasters in a single graph, as implemented in the R package `murphydiagram`

Murphy diagrams

a Murphy curve plots the mean elementary score S_θ of a probabilistic classifier as a function of $\theta \in (0, 1)$

- the Brier score equals (twice) the area under the Murphy curve
- the widely reported misclassification rate equals (twice) the height of the Murphy curve at $\theta = 1/2$

a Murphy diagram plots the curves of competing forecasters in a single graph, as implemented in the R package `murphydiagram`

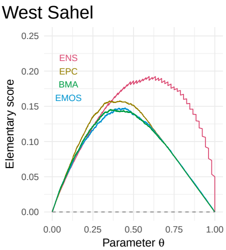


Murphy diagrams

a Murphy curve plots the mean elementary score S_θ of a probabilistic classifier as a function of $\theta \in (0, 1)$

- the Brier score equals (twice) the area under the Murphy curve
- the widely reported misclassification rate equals (twice) the height of the Murphy curve at $\theta = 1/2$

a Murphy diagram plots the curves of competing forecasters in a single graph, as implemented in the R package `murphydiagram`



covers all economic scenarios simultaneously and eliminates the need to choose a proper scoring rule (Murphy 1977; Ehm et al. 2016)

ROC curve, reliability diagram, and Murphy diagram

. . . a triptych

ROC curve, reliability diagram, and Murphy diagram

. . . a triptych

according to Wikipedia, a triptych is "a piece of art [. . .] that is divided into three sections"

ROC curve, reliability diagram, and Murphy diagram

. . . a triptych

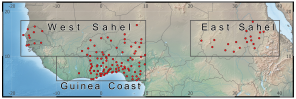
according to Wikipedia, a triptych is “a piece of art [. . .] that is divided into three sections”

Merode Altarpiece by Robert Campin, ca. 1427–32
Metropolitan Museum of Art

Source: Wikimedia
https://commons.wikimedia.org/wiki/File:Robert_Campin_-_L%27_Annonciation_-_1425.jpg

Example

24-hour precipitation forecasts over the West Sahel region in northern tropical Africa in monsoon season 2014 (Vogel et al. 2018)

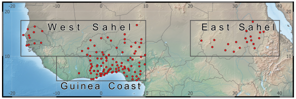


competing probability of precipitation (PoP) forecasts

- **ENS** ECMWF NWP ensemble
- **EMOS** postprocessed by ensemble model output statistics
- **BMA** postprocessed by Bayesian model averaging
- **EPC** extended probabilistic climatology

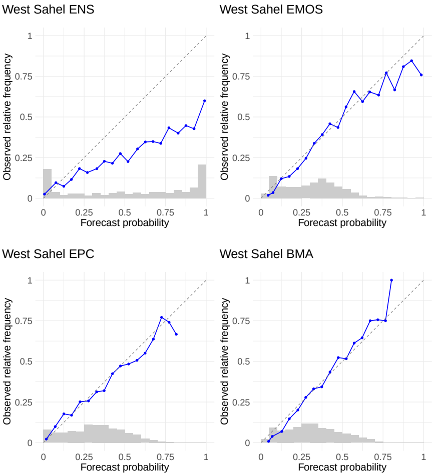
Example

24-hour precipitation forecasts over the West Sahel region in northern tropical Africa in monsoon season 2014 (Vogel et al. 2018)



competing probability of precipitation (PoP) forecasts

- **ENS** ECMWF NWP ensemble
- **EMOS** postprocessed by ensemble model output statistics
- **BMA** postprocessed by Bayesian model averaging
- **EPC** extended probabilistic climatology



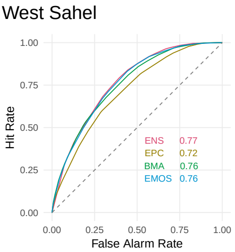
ROC curve, reliability diagram, and Murphy diagram

. . . the triptych in practice

ROC curve, reliability diagram, and Murphy diagram

. . . the triptych in practice

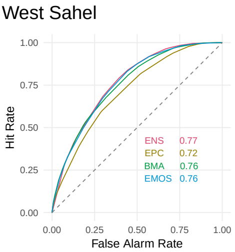
ROC curves assess potential predictive ability

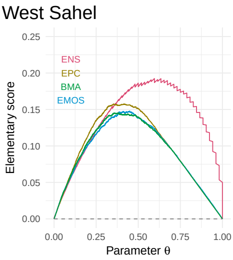


ROC curve, reliability diagram, and Murphy diagram
...the triptych in practice

ROC curves assess potential predictive ability

Murphy diagrams visualize actual (normalized) costs for a decision maker with expense ratio $\theta/(1-\theta)$



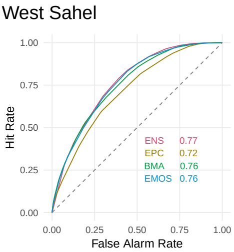


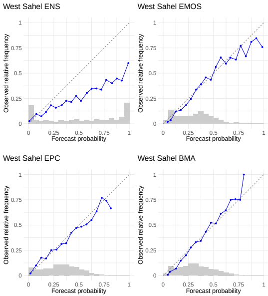
ROC curve, reliability diagram, and Murphy diagram
...the triptych in practice

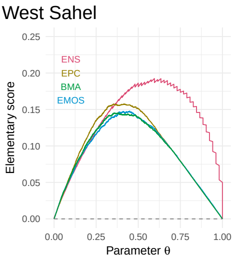
ROC curves assess potential predictive ability

reliability diagrams demonstrate calibration

Murphy diagrams visualize actual (normalized) costs for a decision maker with expense ratio $\theta/(1-\theta)$







Selected references and code

Gneiting, T. and Vogel, P. (2018). **Receiver operating characteristic (ROC) curves.** Preprint, `arXiv:1809.04808`.

Ehm, W., Gneiting, T., Jordan, A. and Krüger, F. (2016), **Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings (with discussion and rejoinder),** *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78, 505–562.

Vogel, P., Knippertz, P., Fink, A. H., Schlueter, A. and Gneiting, T. (2018). **Skill of global raw and postprocessed ensemble predictions of rainfall over northern tropical Africa.** *Weather and Forecasting*, 33, 369–388.

some useful R packages at CRAN: `pROC`, `ROCR`, `verification`, `scoringRules`, `murphydiagram`