

MACHINE LEARNING IN BIOINFORMATICS

MODEL BENCHMARKING

Philipp Benner
philipp.benner@bam.de

VP.1 - eScience
Federal Institute of Materials Research and Testing (BAM)

April 25, 2024

MOTIVATION

- Assume we have developed a machine learning model f :

$$f(X) = \hat{y}$$

- X are the predictors or independent variables, e.g.
 - ▶ DNA sequences, motif scores
- \hat{y} are the predictions
 - ▶ Gene expression levels (regression)
 - ▶ Enhancer active/inactive (classification)
- Suppose we have a test data set (X, y) . How can we evaluate the performance of our model f ?

MODEL BENCHMARKING

BENCHMARKING REGRESSORS

- Residual sum of squares

$$\sum_{i=1}^n (y_i - f(x_i))^2$$

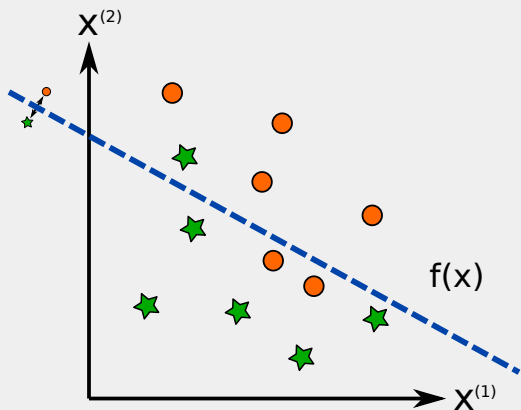
Depends on the variance of y

- Coefficient of determination

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - f(x_i))^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\text{residual sum of squares}}{\text{total sum of squares}}$$

- $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$, the mean, can be interpreted as a reference or baseline regressor
- R^2 compares the predictions of f to the baseline

BENCHMARKING CLASSIFIERS



True positive (TP): 4

True negative (TN): 5

False positive (FP): 1

False negative (FN): 2

BENCHMARKING CLASSIFIERS

- We discuss here binary classification problems, i.e. data with two classes
- We have several options for multiclass problems:
 - ▶ one class vs. all other classes
 - ▶ one class vs. another class
 - ▶ use multi-class losses such as cross-entropy

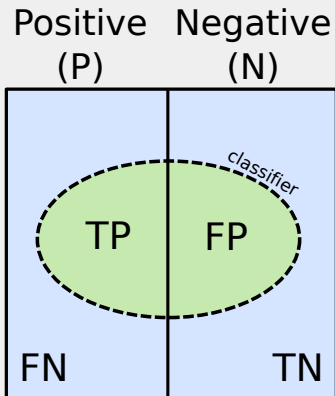
BENCHMARKING CLASSIFIERS

- We discuss here binary classification problems, i.e. data with two classes
- We have several options for multiclass problems:
 - ▶ one class vs. all other classes
 - ▶ one class vs. another class
 - ▶ use multi-class losses such as cross-entropy
- Classifiers typically return a score, or better, a probability:

$$f(x) = P(\text{positive class} | x) > t \Rightarrow \hat{y} = 1$$

- t is a threshold that we can vary
- If the model f is a simple linear function, then t determines the y-intercept

BENCHMARKING CLASSIFIERS



- True positive rate:

$$\text{TPR} = \frac{\text{TP}}{P} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

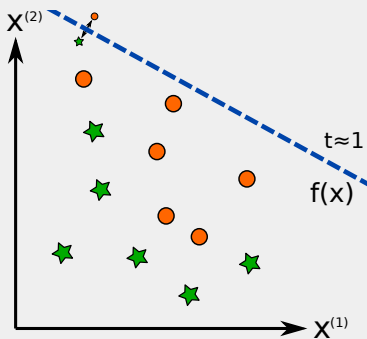
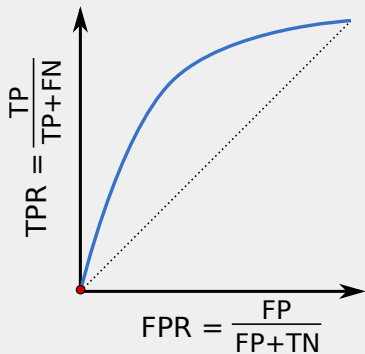
also called: sensitivity or recall
(How well are positives recognized)

- False positive rate:

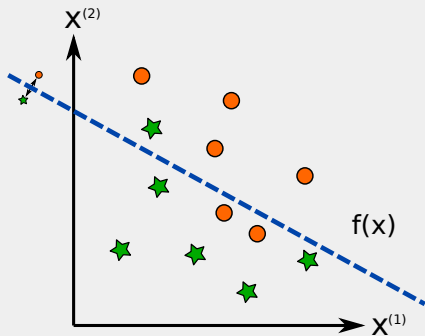
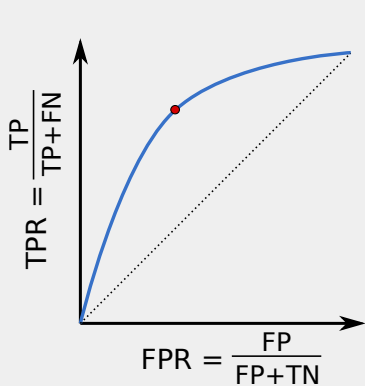
$$\text{FPR} = \frac{\text{FP}}{N} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

(How well are negatives recognized)

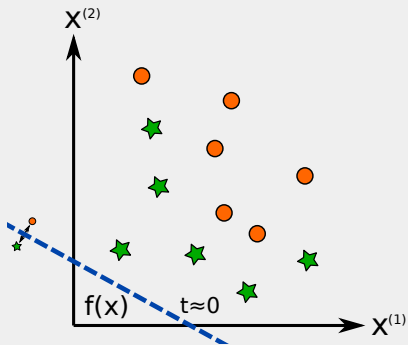
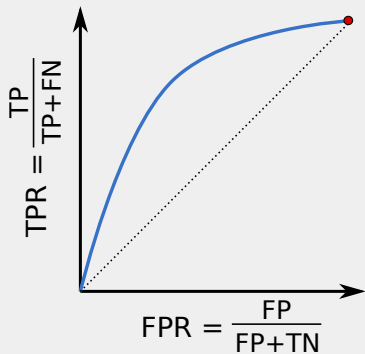
BENCHMARKING CLASSIFIERS: ROC-CURVES



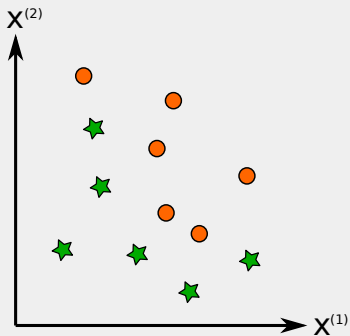
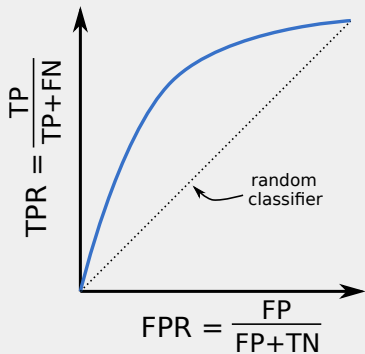
BENCHMARKING CLASSIFIERS: ROC-CURVES



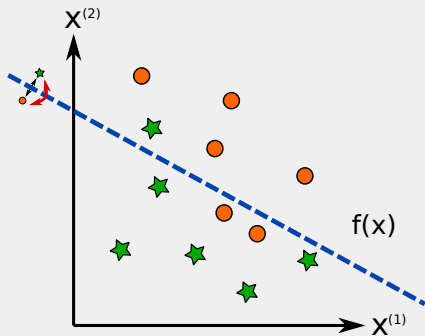
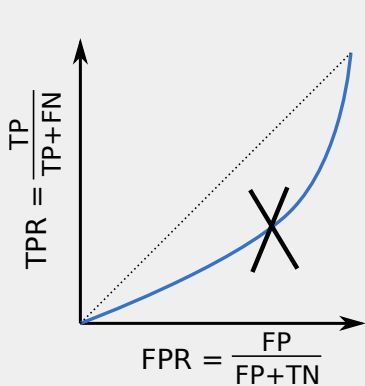
BENCHMARKING CLASSIFIERS: ROC-CURVES



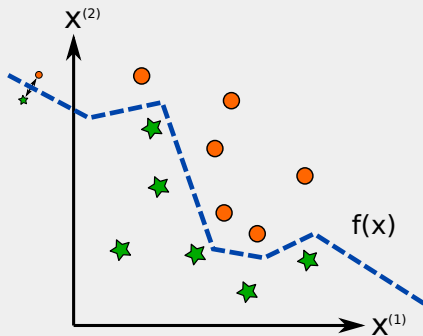
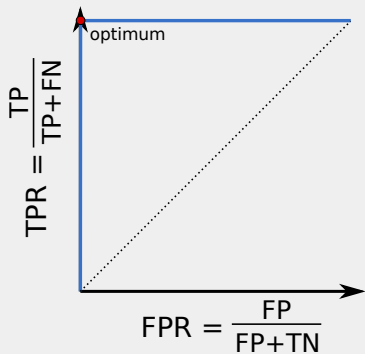
BENCHMARKING CLASSIFIERS: ROC-CURVES



BENCHMARKING CLASSIFIERS: ROC-CURVES



BENCHMARKING CLASSIFIERS: ROC-CURVES



BENCHMARKING CLASSIFIERS

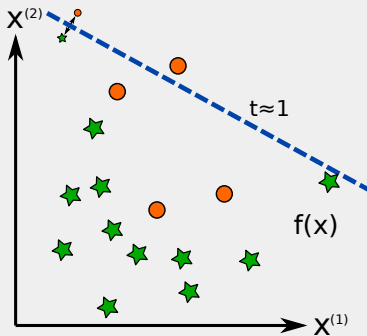
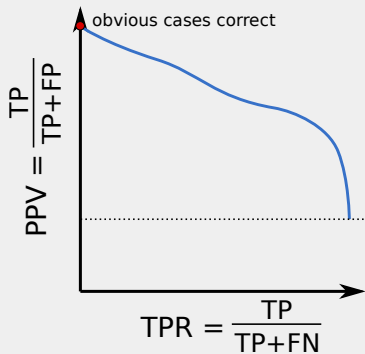
- In practice we often deal with imbalanced data sets, i.e. $P \ll N$
- Example: Genome wide identification of enhancers
- Remark: If $N \ll P$ then we flip labels!
- With ROC curves we never compare P and N :

$$\text{TPR} = \frac{\text{TP}}{P} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{FPR} = \frac{\text{FP}}{N} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

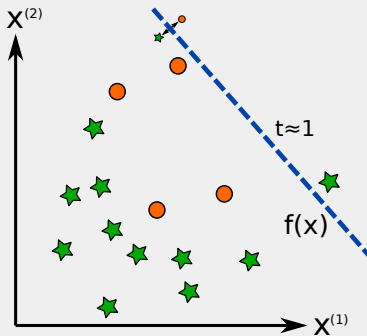
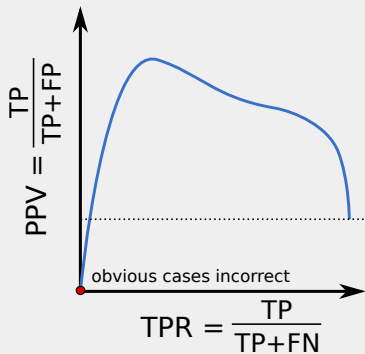
- Positive predictive value (PPV, or precision):

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

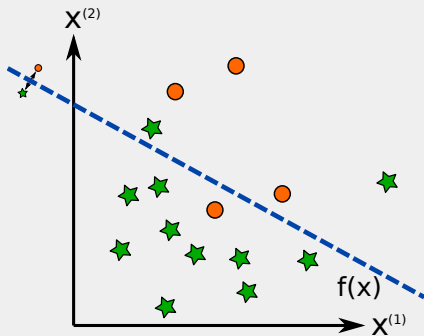
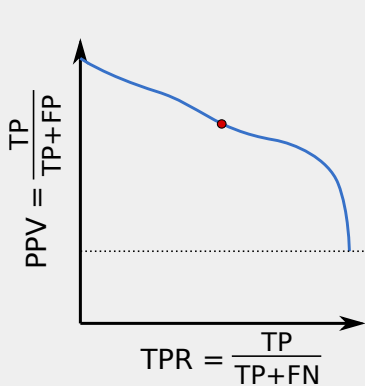
BENCHMARKING CLASSIFIERS: PR-CURVES



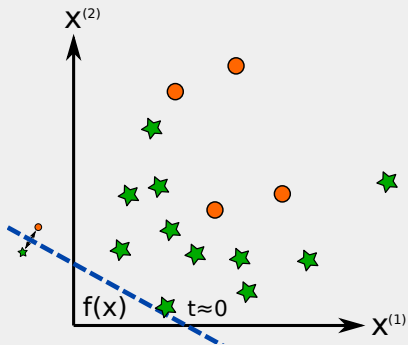
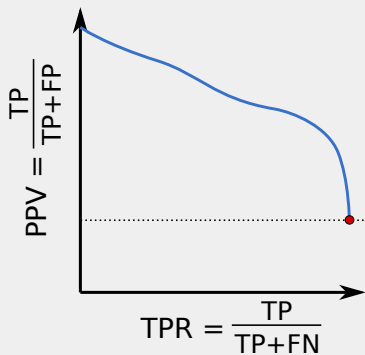
BENCHMARKING CLASSIFIERS: PR-CURVES



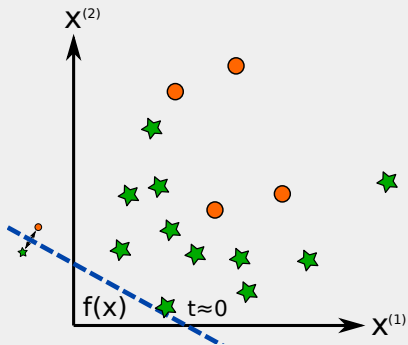
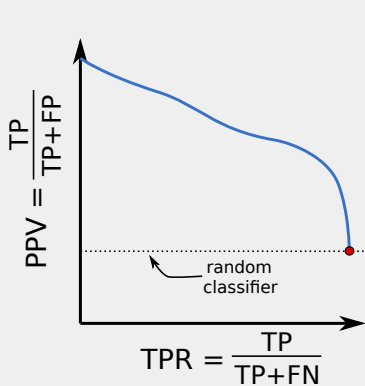
BENCHMARKING CLASSIFIERS: PR-CURVES



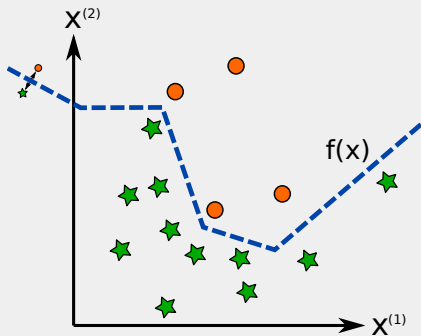
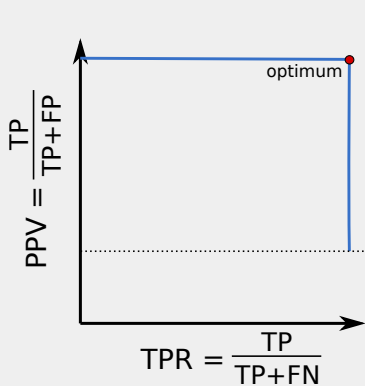
BENCHMARKING CLASSIFIERS: PR-CURVES



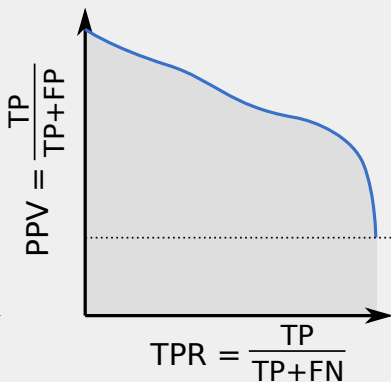
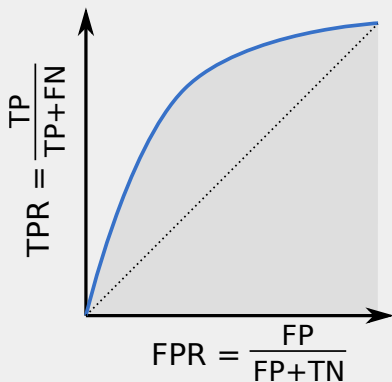
BENCHMARKING CLASSIFIERS: PR-CURVES



BENCHMARKING CLASSIFIERS: PR-CURVES



BENCHMARKING CLASSIFIERS: ROC/PR-AUC



$$ROC-AUC = \int_t TPR(t) dFPR(t)$$

$$PR-AUC = \int_t PPV(t) dTPR(t)$$

BENCHMARKING CLASSIFIERS: CLASSICAL STATISTICS

- Probability of a type I error:

$$\begin{aligned}\alpha &= P(f(x) = 1 | y = 0) \\ &\approx \frac{\text{FP}}{\text{FP} + \text{TN}}\end{aligned}$$

- Probability of a type II error:

$$\begin{aligned}\beta &= P(f(x) = 0 | y = 1) \\ &\approx \frac{\text{FN}}{\text{TP} + \text{FN}}\end{aligned}$$

- Power of a statistical test:

$$\begin{aligned}\gamma &= P(f(x) = 1 | y = 1) \\ &= 1 - \beta\end{aligned}$$

BENCHMARKING CLASSIFIERS: ADVANCED MEASURES

- All measures so far were likelihood based, which ignore prevalences
- There are also posterior or "Bayesian" measures
- False discovery rate (FDR):

$$P(y = 1 | f(x) = 0) = \frac{\alpha\pi_0}{\alpha\pi_0 + \gamma\pi_1}$$

- False omission rate (FOR):

$$P(y = 0 | f(x) = 1) = \frac{\beta\pi_1}{(1 - \alpha)\pi_0 + \beta\pi_1}$$

- Where: $\pi_0 = N/(P + N)$ and $\pi_1 = P/(P + N)$
- Both the FDR and FOR require an estimate of prevalences

- Section 5.7.2 [Murphy, 2012]

REFERENCES



MURPHY, K. P. (2012).

MACHINE LEARNING: A PROBABILISTIC PERSPECTIVE.

MIT press.