Machine Learning in Bioinformatics

Model Benchmarking

Philipp Benner
philipp.benner@bam.de

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

February 8, 2026

Motivation

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

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

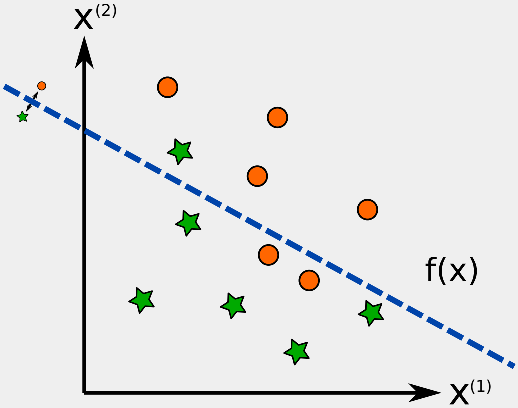
  Depends on the variance of y

- 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



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

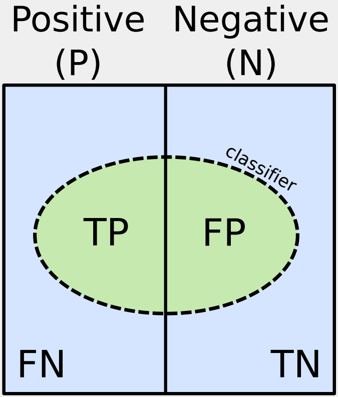
- We have several options for multiclass problems:

  - one class vs. all other classes
  - one class vs. another class
  - use multi-class losses such as cross-entropy

- 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:

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

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



- True positive rate:

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

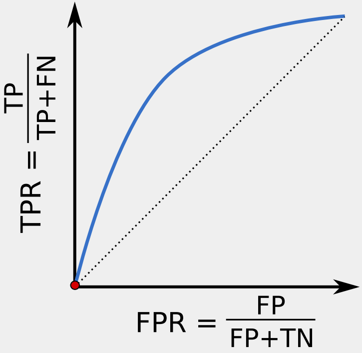
also called: sensitivity or recall (How well are positives recognized)

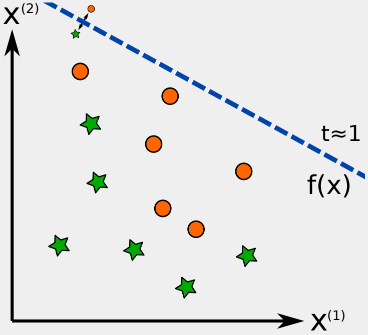
- False positive rate:

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

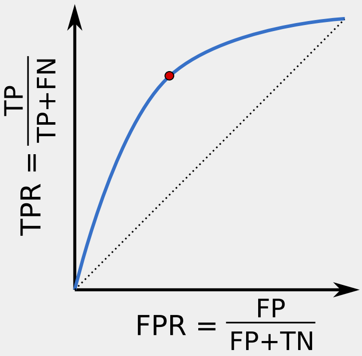
(How well are negatives recognized)

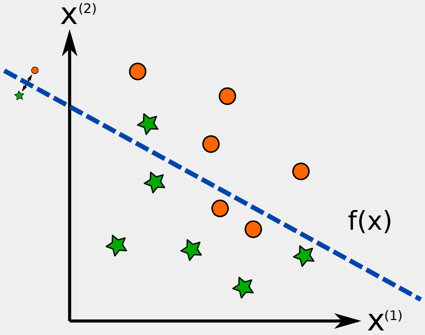
Benchmarking Classifiers: ROC-Curves



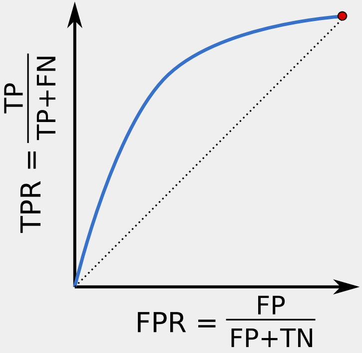


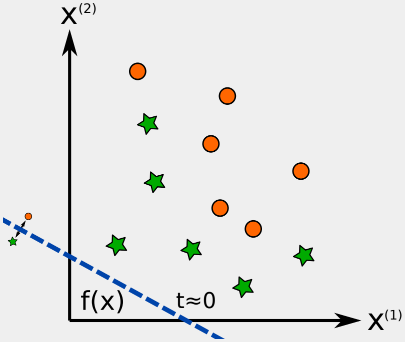
Benchmarking Classifiers: ROC-Curves



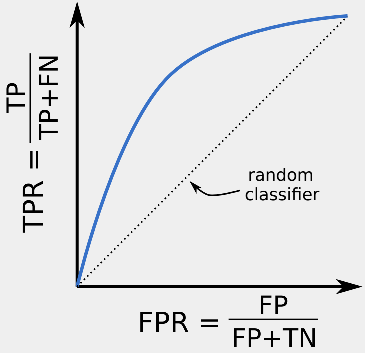


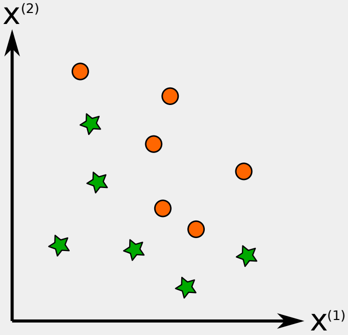
Benchmarking Classifiers: ROC-Curves



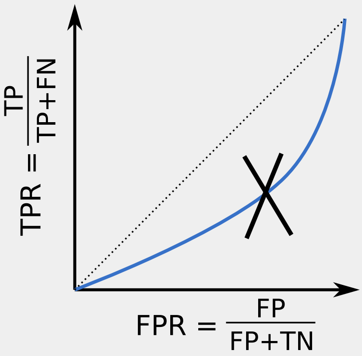


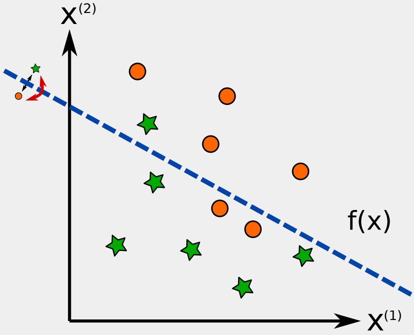
Benchmarking Classifiers: ROC-Curves

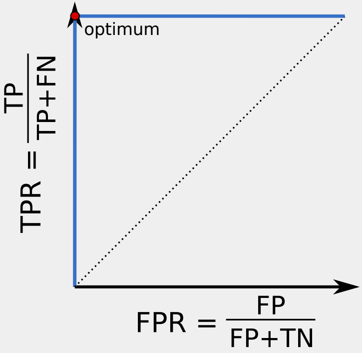


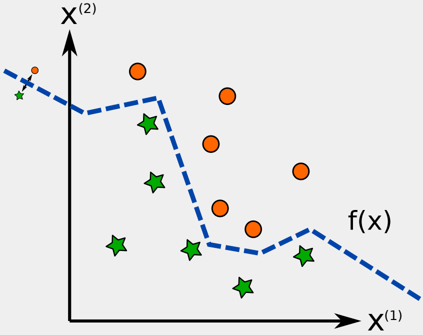


Benchmarking Classifiers: ROC-Curves






optimum
$TPR = \frac{TP}{TP+FN}$
$FPR = \frac{FP}{FP+TN}$

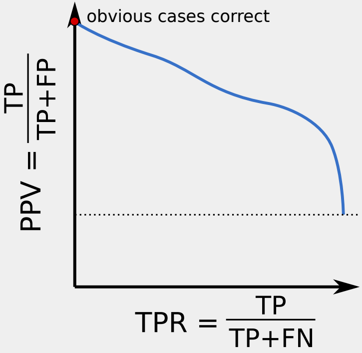

$x^{(2)}$
$x^{(1)}$
f(x)

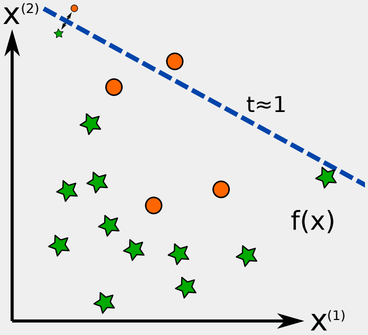
- In practice we often deal with imbalanced data sets, i.e. $P \ll N$
- Example: Genome wide identification of enhancers
- Remark: If $N \ll P$ then we flip labels!
- With ROC curves we never compare P and N:

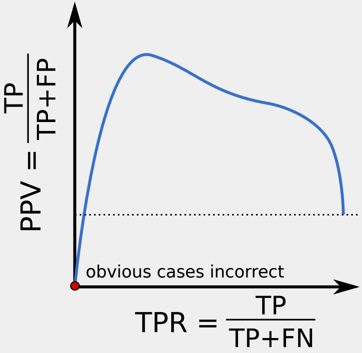
$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \quad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{N}} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

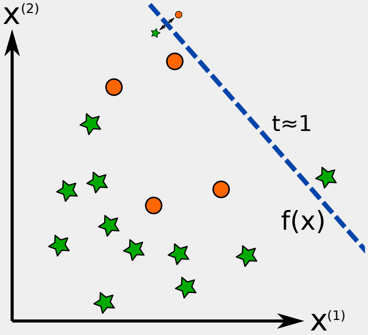
- Positive predictive value (PPV, or precision):

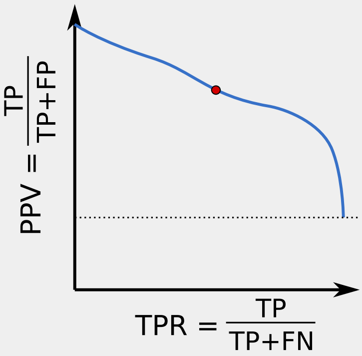
$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

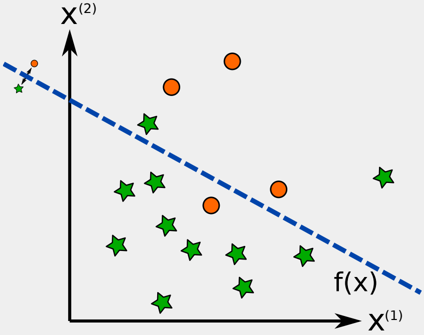

obvious cases correct
$PPV = \frac{TP}{TP+FP}$
$TPR = \frac{TP}{TP+FN}$

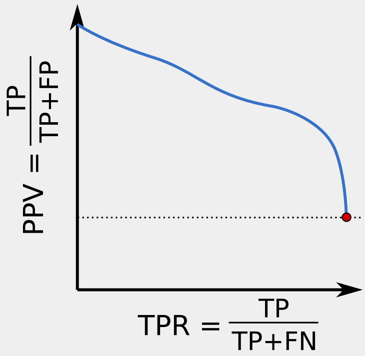

$x^{(2)}$
t≈1
f(x)
$x^{(1)}$


PPV = TP / TP+FP
obvious cases incorrect
TPR = TP / TP+FN

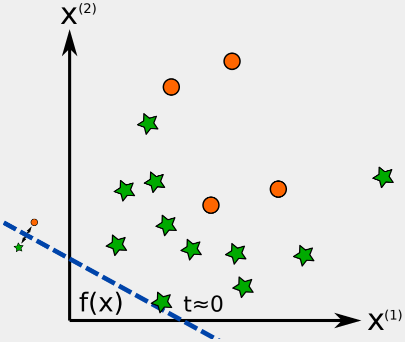

x(2)
t≈1
f(x)
x(1)


PPV = TP / TP+FP
TPR = TP / TP+FN


x(2)
x(1)
f(x)

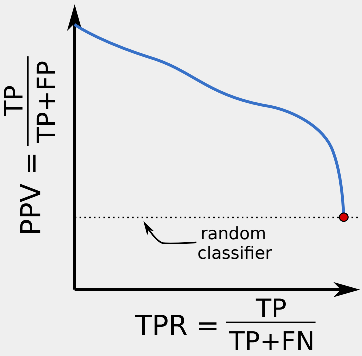


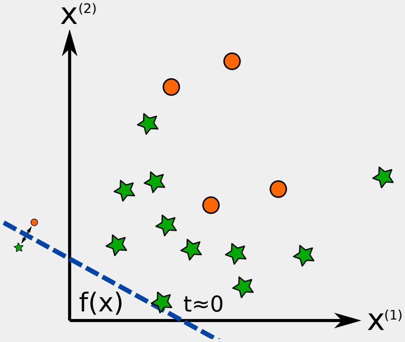
$PPV = \frac{TP}{TP+FP}$
$TPR = \frac{TP}{TP+FN}$



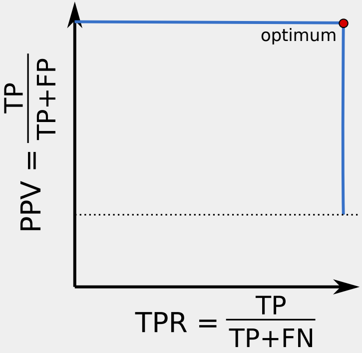
$x^{(2)}$
$x^{(1)}$
f(x)
t≈0

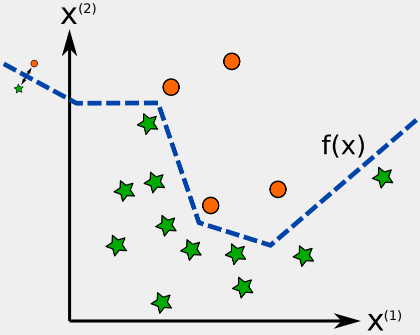
Benchmarking Classifiers: PR-Curves



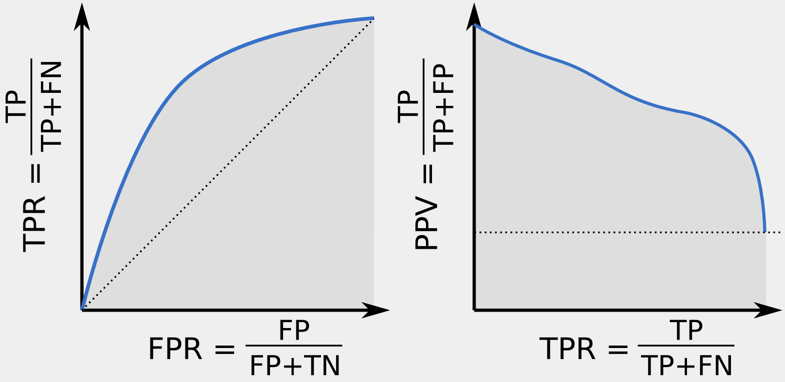


Benchmarking Classifiers: PR-Curves



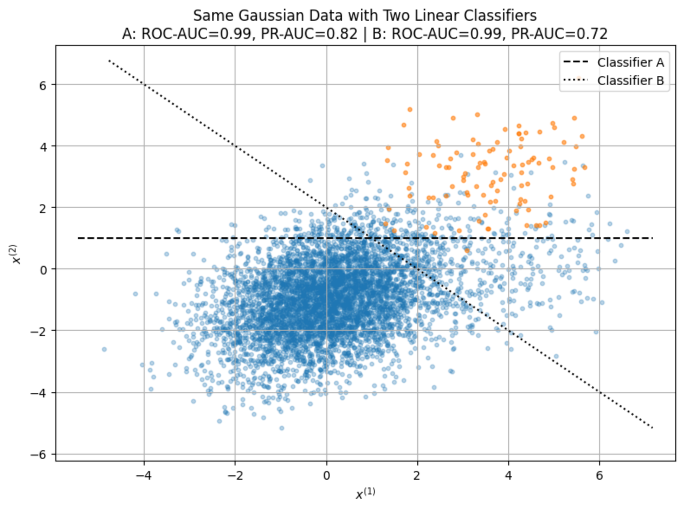


Benchmarking Classifiers: ROC/PR-AUC

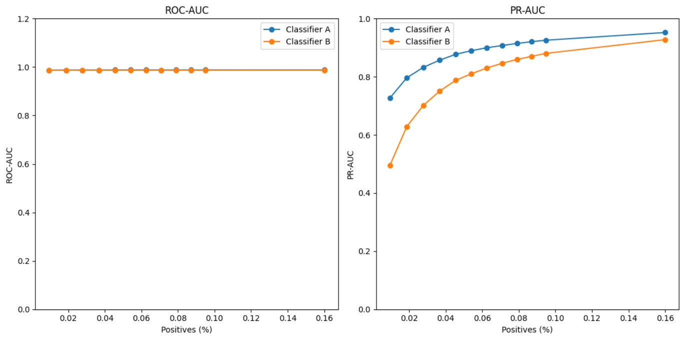


$$\text{ROC-AUC} = \int_t \text{TPR}(t) \mathrm{d}\text{FPR}(t) \qquad \text{PR-AUC} = \int_t \text{PPV}(t) \mathrm{d}\text{TPR}(t)$$

Benchmarking Classifiers: ROC/PR-AUC



Benchmarking Classifiers: ROC/PR-AUC



Benchmarking Classifiers: Classical Statistics

- Probability of a type I error:

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

- Probability of a type II error:

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

- Power of a statistical test:

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

- 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

Reading

- Section 5.7.2 [Murphy, 2012]

References

Murphy, K. P. (2012).
Machine learning: a probabilistic perspective.
MIT press.