

Introduction to Audio Content Analysis

Module 6.0: Evaluation and Metrics

alexander lerch

introduction

overview

corresponding textbook section

Section 6

■ lecture content

- evaluation methodology
- good practices
- metrics

■ learning objectives

- design proper evaluation setups for machine learning algorithms
- list relevant metrics for different machine learning models



introduction

overview

corresponding textbook section

Section 6

■ lecture content

- evaluation methodology
- good practices
- metrics

■ learning objectives

- design proper evaluation setups for machine learning algorithms
- list relevant metrics for different machine learning models



evaluation

introduction

- without proper evaluation, there is no way to say whether a system works
- typical mistakes in evaluation
 - 1 non-representative test set
 - ① small, too homogeneous, ...
 - 2 tuning system parameters with the test set (explicitly or implicitly)
 - 3 using misleading evaluation procedures and metrics

evaluation

good practices

- evaluation **method unrelated** to the specific implementation
 - has to be task driven, not algorithm driven
 - metrics should be unrelated to loss function
- **expectations** clearly defined
 - worst case performance (random)
 - best case performance (oracle)
 - realistic performance \Rightarrow baseline system
 - ▶ Zero-R classifier
 - ▶ standard approach
- **comparability** to state-of-the-art
 - use of established datasets and identical data splits
 - running existing systems on your data
- increase **reproducibility**
 - automate evaluation
 - publish source code
- test for **statistical significance**

evaluation

good practices

- evaluation **method unrelated** to the specific implementation
 - has to be task driven, not algorithm driven
 - metrics should be unrelated to loss function
- **expectations** clearly defined
 - worst case performance (random)
 - best case performance (oracle)
 - realistic performance \Rightarrow baseline system
 - ▶ Zero-R classifier
 - ▶ standard approach
- **comparability** to state-of-the-art
 - use of established datasets and identical data splits
 - running existing systems on your data
- increase **reproducibility**
 - automate evaluation
 - publish source code
- test for **statistical significance**

evaluation

good practices

- evaluation **method unrelated** to the specific implementation
 - has to be task driven, not algorithm driven
 - metrics should be unrelated to loss function
- **expectations** clearly defined
 - worst case performance (random)
 - best case performance (oracle)
 - realistic performance \Rightarrow baseline system
 - ▶ Zero-R classifier
 - ▶ standard approach
- **comparability** to state-of-the-art
 - use of established datasets and identical data splits
 - running existing systems on your data
- increase **reproducibility**
 - automate evaluation
 - publish source code
- test for **statistical significance**

evaluation

good practices

- evaluation **method unrelated** to the specific implementation
 - has to be task driven, not algorithm driven
 - metrics should be unrelated to loss function
- **expectations** clearly defined
 - worst case performance (random)
 - best case performance (oracle)
 - realistic performance \Rightarrow baseline system
 - ▶ Zero-R classifier
 - ▶ standard approach
- **comparability** to state-of-the-art
 - use of established datasets and identical data splits
 - running existing systems on your data
- increase **reproducibility**
 - automate evaluation
 - publish source code
- test for **statistical significance**

evaluation

good practices

- evaluation **method unrelated** to the specific implementation
 - has to be task driven, not algorithm driven
 - metrics should be unrelated to loss function
- **expectations** clearly defined
 - worst case performance (random)
 - best case performance (oracle)
 - realistic performance \Rightarrow baseline system
 - ▶ Zero-R classifier
 - ▶ standard approach
- **comparability** to state-of-the-art
 - use of established datasets and identical data splits
 - running existing systems on your data
- increase **reproducibility**
 - automate evaluation
 - publish source code
- test for **statistical significance**

classification metrics

introduction

- possible outcomes of two class problem (positive and negative):
 - TP: Positives correctly identified as Positives,
 - TN: Negatives correctly identified Negatives,
 - FP: Negatives incorrectly identified Positives, and
 - FN: Positives incorrectly identified Negatives.
- visualization: confusion matrix

	Pred. Pos.	Pred. Neg.	Σ
GT Pos.	True Pos. (TP)	False Neg. (FN)	# of GT Pos. (TP+FN)
GT Neg.	False Pos. (FP)	True Neg. (TN)	# of GT Neg. (FP+FN)
Σ	# of Pred. Pos.	# of Pred. Neg.	# of True Pred. (TP+TN)

classification metrics

accuracy and f-measure

- **accuracy**: how many predictions are accurate
- **macro accuracy**: averaged over classes (not observations)
- **precision**: how many predicted positives are correct
- **recall**: how many ground truth positives correctly predicted
- **f-measure**: combines precision and recall

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN}$$

classification metrics

accuracy and f-measure

- **accuracy**: how many predictions are accurate
- **macro accuracy**: averaged over classes (not observations)
- **precision**: how many predicted positives are correct
- **recall**: how many ground truth positives correctly predicted
- **f-measure**: combines precision and recall

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Acc}_{\text{Macro}} = \frac{\frac{TP}{TP+FN} + \frac{TN}{TN+FP}}{2} = \frac{TPR + TNR}{2}$$

classification metrics

accuracy and f-measure

- **accuracy**: how many predictions are accurate
- **macro accuracy**: averaged over classes (not observations)
- **precision**: how many predicted positives are correct
- **recall**: how many ground truth positives correctly predicted
- **f-measure**: combines precision and recall

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Acc}_{\text{Macro}} = \frac{\frac{TP}{TP+FN} + \frac{TN}{TN+FP}}{2} = \frac{\text{TPR} + \text{TNR}}{2}$$

$$P = \frac{TP}{TP + FP}$$

classification metrics

accuracy and f-measure

- **accuracy**: how many predictions are accurate
- **macro accuracy**: averaged over classes (not observations)
- **precision**: how many predicted positives are correct
- **recall**: how many ground truth positives correctly predicted
- **f-measure**: combines precision and recall

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Acc}_{\text{Macro}} = \frac{\frac{TP}{TP+FN} + \frac{TN}{TN+FP}}{2} = \frac{\text{TPR} + \text{TNR}}{2}$$

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

classification metrics

accuracy and f-measure

- **accuracy**: how many predictions are accurate
- **macro accuracy**: averaged over classes (not observations)
- **precision**: how many predicted positives are correct
- **recall**: how many ground truth positives correctly predicted
- **f-measure**: combines precision and recall

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Acc}_{\text{Macro}} = \frac{\frac{TP}{TP+FN} + \frac{TN}{TN+FP}}{2} = \frac{\text{TPR} + \text{TNR}}{2}$$

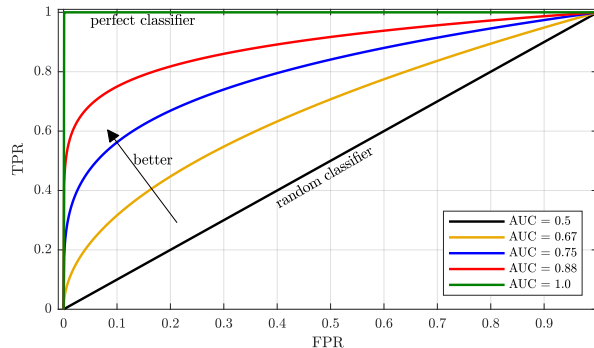
$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

$$F = 2 \cdot \frac{P \cdot R}{P + R}$$

classification metrics

area under curve



regression metrics

mae, mse, R^2

goal: measure deviation

■ mean absolute error

■ mean squared error

■ coefficient of determination

$$MAE = \frac{1}{\mathcal{R}} \sum_{\forall r} |y(r) - \hat{y}(r)|$$

regression metrics

mae, mse, R^2

goal: measure deviation

■ mean absolute error

■ mean squared error

■ coefficient of determination

$$MAE = \frac{1}{\mathcal{R}} \sum_{\forall r} |y(r) - \hat{y}(r)|$$

$$MSE = \frac{1}{\mathcal{R}} \sum_{\forall r} (y(r) - \hat{y}(r))^2$$

regression metrics

mae, mse, R^2

goal: measure deviation

- mean absolute error
- mean squared error
- coefficient of determination

$$MAE = \frac{1}{\mathcal{R}} \sum_{\forall r} |y(r) - \hat{y}(r)|$$

$$MSE = \frac{1}{\mathcal{R}} \sum_{\forall r} (y(r) - \hat{y}(r))^2$$

$$R^2 = 1 - \frac{MSE(y - \hat{y})}{MSE(y - \mu_y)}$$

summary

lecture content

■ evaluation

- system development without evaluation is meaningless
- data and method need to be carefully selected
- metrics need to reflect the success of the system

■ classification metrics

- accuracy and macro accuracy
- precision, recall, and f-measure
- AUC

■ regression metrics

- MAE and MSE
- coefficient of determination

