Introduction to **Audio Content Analysis**

Module 6.0: Evaluation and Metrics

alexander lerch

Georgia Tech Center for Music Technology
College of Design

introduction
overview

corresponding textbook section

chapter 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

chapter 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 1/2

- 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 ⇒ baseline system
    - Zero-R classifier
    - standard approach

evaluation

good practices 1/2

- 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

evaluation
good practices 2/2

- **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 2/2

- **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 2/2

Georgia Tech | Center for Music Technology
College of Design

- **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

		Predicted		
		Positive	**Negative**	$\mathbf{\Sigma}$
GT	**Positive**	TP True Positives	FN False Negatives	TP+FN # of GT Positives
	Negative	FP False Positives	TN True Negatives	FP+TN # of GT Negatives
	$\mathbf{\Sigma}$	TP+FP # of Predicted Positives	TN+FN # of Predicted Negatives	TP+TN # of True Predictions

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

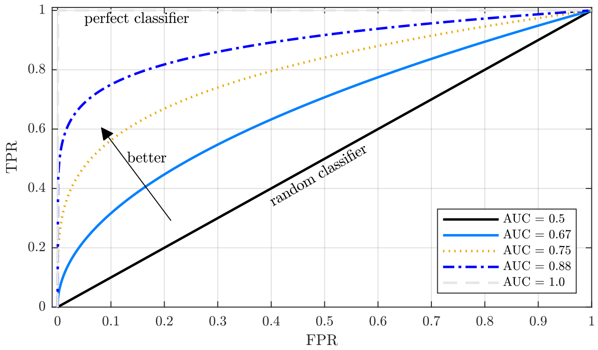
$$P = \frac{TP}{TP + FP}$$

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

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

classification metrics

area under curve



matlab source: plotROC.m

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

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

- mean squared error

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

- coefficient of determination

regression metrics

mae, mse, R^2

goal: measure deviation

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

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

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

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

summary
lecture content

- **evaluation**
  - system development without evaluation is meaningless
  - data and method need to be carefully selected
  - metrics need to reflect the sucess of the system
- **classification metrics**
  - accuracy and macro accuracy
  - precision, recall, and f-measure
  - AUC
- **regression metrics**
  - MAE and MSE
  - coefficient of determination