



AARHUS
UNIVERSITY

Class 14: Exam and Q&A

Computational Analysis of Text, Audio, and Images, Fall 2023

Aarhus University

Mathias Rask (mathiasrask@ps.au.dk)

Aarhus University

Table of Contents

Q&A

Eksamen

Lab

Q1: Metrics

What's the difference between loss functions and evaluation metrics?

Q1: Metrics

What's the difference between loss functions and evaluation metrics?

Loss functions are used in optimization (i.e. in training)

Evaluation of the training is done using metrics:

Q1: Metrics

What's the difference between loss functions and evaluation metrics?

Loss functions are used in optimization (i.e. in training)

Evaluation of the training is done using metrics:

- Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$
 - Proportion of correct predictions

Q1: Metrics

What's the difference between loss functions and evaluation metrics?

Loss functions are used in optimization (i.e. in training)

Evaluation of the training is done using metrics:

- Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$
 - Proportion of correct predictions
- Precision: $\frac{TP}{TP+FP}$
 - Ratio of correctly predicted positives to all positives in the class
 - How certain are we that the predictions are correct?

Q1: Metrics

What's the difference between loss functions and evaluation metrics?

Loss functions are used in optimization (i.e. in training)

Evaluation of the training is done using metrics:

- Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$
 - Proportion of correct predictions
- Precision: $\frac{TP}{TP+FP}$
 - Ratio of correctly predicted positives to all positives in the class
 - How certain are we that the predictions are correct?
- Recall: $\frac{TP}{TP+FN}$
 - Ratio of correctly predicted positives to all observations in actual class
 - How good are we at *recalling* a certain class?

Q1: Metrics

What's the difference between loss functions and evaluation metrics?

Loss functions are used in optimization (i.e. in training)

Evaluation of the training is done using metrics:

- Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$
 - Proportion of correct predictions
- Precision: $\frac{TP}{TP+FP}$
 - Ratio of correctly predicted positives to all positives in the class
 - How certain are we that the predictions are correct?
- Recall: $\frac{TP}{TP+FN}$
 - Ratio of correctly predicted positives to all observations in actual class
 - How good are we at *recalling* a certain class?

Note: Used to evaluate algorithms, but generalizes to more general problems such as dictionaries.

Q5: Losses and Metrics for Continuous Data

Discuss how loss functions and metrics relate to binary vs continuous outcomes.

Q2: Text Preprocessing

Q2: Text Preprocessing

When and which preprocessing steps should we take?

Q2: Text Preprocessing

When and which preprocessing steps should we take?

1. Common sense (do not remove numbers if you are interested in numbers...)

Q2: Text Preprocessing

When and which preprocessing steps should we take?

1. Common sense (do not remove numbers if you are interested in numbers...)
2. What do I think works for my application? (Denny and Spirling, 2018)

Q2: Text Preprocessing

When and which preprocessing steps should we take?

1. Common sense (do not remove numbers if you are interested in numbers...)
2. What do I think works for my application? (Denny and Spirling, 2018)
3. Try different configurations (Rodriguez and Spirling, 2022)

Q2: Text Preprocessing

When and which preprocessing steps should we take?

1. Common sense (do not remove numbers if you are interested in numbers...)
2. What do I think works for my application? (Denny and Spirling, 2018)
3. Try different configurations (Rodriguez and Spirling, 2022)

↪ For exam: Argue for your choices and you will be fine (if you are not totally ridiculous)

Q2: Text Preprocessing

When and which preprocessing steps should we take?

1. Common sense (do not remove numbers if you are interested in numbers...)
2. What do I think works for my application? (Denny and Spirling, 2018)
3. Try different configurations (Rodriguez and Spirling, 2022)

↪ For exam: Argue for your choices and you will be fine (if you are not totally ridiculous)

A good argument uses the underlying method as a reference point (but no global best)

Q4: Embeddings

Q4: Embeddings

What's the idea and intuition with embeddings?

Q4: Embeddings

What's the idea and intuition with embeddings?

- In general: to represent discrete objects (e.g. speakers, words, letters, movies, football players, ...) using a dense and efficient numerical representation

Q4: Embeddings

What's the idea and intuition with embeddings?

- In general: to represent discrete objects (e.g. speakers, words, letters, movies, football players, ...) using a dense and efficient numerical representation
- Word embeddings: To represent words efficiently in a dense vector while preserving meaning – we shall know a word by the company it keeps

Q4: Embeddings

What's the idea and intuition with embeddings?

- In general: to represent discrete objects (e.g. speakers, words, letters, movies, football players, ...) using a dense and efficient numerical representation
- Word embeddings: To represent words efficiently in a dense vector while preserving meaning – we shall know a word by the company it keeps
- The context hypothesis usually means that we use less preprocessing

Q4: Embeddings

What's the idea and intuition with embeddings?

- In general: to represent discrete objects (e.g. speakers, words, letters, movies, football players, ...) using a dense and efficient numerical representation
- Word embeddings: To represent words efficiently in a dense vector while preserving meaning – we shall know a word by the company it keeps
- The context hypothesis usually means that we use less preprocessing
- We can take steps in the vector space: $\text{king} + \text{woman} - \text{man} = \text{queen}$. Why?

Q3: Audio Preprocessing

Q3: Audio Preprocessing

What are the intuitions guiding converting audio from analog to digital format?

Q3: Audio Preprocessing

What are the intuitions guiding converting audio from analog to digital format?

- We do not control the conversion – but it influences what we can get out of it (e.g. sampling rate vs signal frequency).

Q3: Audio Preprocessing

What are the intuitions guiding converting audio from analog to digital format?

- We do not control the conversion – but it influences what we can get out of it (e.g. sampling rate vs signal frequency).

The question is also about speaker diarization, speaker recognition, and speaker identification. Logic is the same, but here it is more general: bad quality (on average) gives bad results.

Table of Contents

Q&A

Eksamen

Lab

Opgavesættet består af tre dele:

1. Part I (teoretisk)
2. Part II (kodning)
3. Part III (kodning)

Opgavesættet består af tre dele:

1. Part I (teoretisk)
2. Part II (kodning)
3. Part III (kodning)

→ Indgår med forskellige vægte (fx 25%, 50%, 25%)

Opgavesættet består af tre dele:

1. Part I (teoretisk)
2. Part II (kodning)
3. Part III (kodning)

↪ Indgår med forskellige vægte (fx 25%, 50%, 25%)

Dækker hver af de tre temaer vi har været igennem (minus billeder)

Need-to-know:

Need-to-know:

En gyldig besvarelse indeholder:

Need-to-know:

En gyldig besvarelse indeholder:

- en `.pdf`-fil
- en eller flere `.ipynb`-fil(er)

Need-to-know:

En gyldig besvarelse indeholder:

- en **.pdf**-fil
- en eller flere **.ipynb**-fil(er)

↪ opgaven bedømmes kun, hvis begge indgår

Nice-to-know:

- lav hele besvarelsen i **Jupyter Notebook** – brug **.pdf**'en til at forklare din struktur, hvis du fx. bruger flere notebooks.
- brug en lokal notebook (hvis muligt!!) - data er nemmere at indlæse
- tjek at din(e) notebook(s) kører uden fejlmeddelser. Bedømmelsen gives ikke ud fra pæn og efficient kode, men at koden kører uden fejl er en forudsætning for at vurdere jeres besvarelse
- brug ikke tid på at lave kode pæn, men brug tid på at sikre jer, at den er læsbar for andre – beskriv hvad I gør.
- vær eksplicit omkring, hvornår hver enkelt spørgsmål besvares (fx 1.1. eller 2.1.1)
- undlag at inkludere opgavebeskrivelser i din besvarelse. Inkluder kun nummeringen og/eller overskriften.
- i kan både beskrive på dansk og engelsk – hvis det første er engelske begreber helt fint.

- Eksamen indeholder to praktiske dele, en med tekst og en med lyd.
- I eksamensættet har jeg vedlagt informationer om datasættene (fx hvilke variable der er), som burde være alt I har brug for.
- Der kan sagtens være fejl (fx stavefejl), men det burde stadig være til at gå til.
- **Tjek at I kan tilgå og indlæse data som det første, når eksamen bliver tilgængelig.**
 - ↪ jeg er tilgængelig på Mail de første 2 timer af eksamen.

- Eksamen indeholder to **requirements-exam-XX.txt** (XX: colab eller local) filer, der indeholder nogle pakker I kan bruge. Det er kun tænkt som en hjælp, men I kan sagtens løse den med andre pakker.
- Hvis du bruger en lokal notebook SKAL du lave et virtual environment (for din egen skyld). Brug version **python=3.8**, da det vil sikre kompatibilitet med **gensim**, som bruges til at implementere embeddings.
- Der ligger en ny version for **class08** på GitHub'en, som implementer en colab-version af vores lokale kode. Forskellen er, at colab bruger **python=3.10** (mener jeg) og en anden version af **gensim**. Det betyder til gengæld, at vi ikke kan bruge **danlp** i colab (men den pakke er heller ikke nødvendig til eksamen).
- Dokumenter, dokumenter, dokumenter!

Den gode eksamensbesvarelse gør brug af “transfer learning”.
Fx: Tag begrebet overfittig og oversæt det til en ny kontekst

Den gode besvarelse

Den gode eksamensbesvarelse gør brug af “transfer learning”.

Fx: Tag begrebet overfittig og oversæt det til en ny kontekst

Den gode besvarelse formår at:

Den gode eksamensbesvarelse gør brug af “transfer learning”.

Fx: Tag begrebet overfittig og oversæt det til en ny kontekst

Den gode besvarelse formår at:

1. implementere kode

Den gode eksamensbesvarelse gør brug af “transfer learning”.

Fx: Tag begrebet overfittig og oversæt det til en ny kontekst

Den gode besvarelse formår at:

1. implementere kode
2. fortolke substantielt på resultatet ift. den underliggende teori (e.g. hvorfor standardiserer vi, hvad er formålet egentlig. Med hvad standardiserer vi, etc.?)

Den gode eksamensbesvarelse gør brug af “transfer learning”.

Fx: Tag begrebet overfittig og oversæt det til en ny kontekst

Den gode besvarelse formår at:

1. implementere kode
2. fortolke substantielt på resultatet ift. den underliggende teori (e.g. hvorfor standardiserer vi, hvad er formålet egentlig. Med hvad standardiserer vi, etc.?)
3. oversætte, hvad I har lært til nye scenarier – det viser dybereliggende forståelse

Den gode eksamensbesvarelse gør brug af “transfer learning”.

Fx: Tag begrebet overfittig og oversæt det til en ny kontekst

Den gode besvarelse formår at:

1. implementere kode
2. fortolke substantielt på resultatet ift. den underliggende teori (e.g. hvorfor standardiserer vi, hvad er formålet egentlig. Med hvad standardiserer vi, etc.?)
3. oversætte, hvad I har lært til nye scenarier – det viser dybereliggende forståelse
4. definere det underliggende mål (hvad forsøger vi egentligt at måle?)

Den gode eksamensbesvarelse gør brug af “transfer learning”.

Fx: Tag begrebet overfittig og oversæt det til en ny kontekst

Den gode besvarelse formår at:

1. implementere kode
2. fortolke substantielt på resultatet ift. den underliggende teori (e.g. hvorfor standardiserer vi, hvad er formålet egentlig. Med hvad standardiserer vi, etc.?)
3. oversætte, hvad I har lært til nye scenarier – det viser dybereliggende forståelse
4. definere det underliggende mål (hvad forsøger vi egentligt at måle?)

Table of Contents

Q&A

Eksamen

Lab

See you next week!

Computational Analysis of Text, Audio, and Images, Fall 2023
Aarhus University

- [1] M. J. Denny and A. Spirling, “Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it,” *Political Analysis*, vol. 26, no. 2, pp. 168–189, 2018.
- [2] P. L. Rodriguez and A. Spirling, “Word embeddings: What works, what doesn’t, and how to tell the difference for applied research,” *The Journal of Politics*, vol. 84, no. 1, pp. 101–115, 2022.