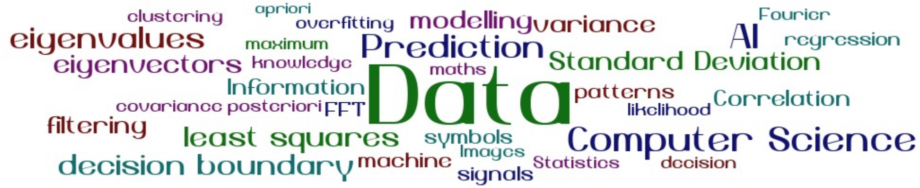
COMS20011 – Data-Driven Computer Science



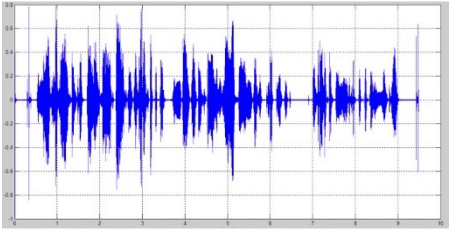
February 2022

Majid Mirmehdi

Some slides in this lecture are adapted from those authored by **Dima Damen** and **Andrew Calway**

Lecture Video #2

Ex2. Speech Recognition

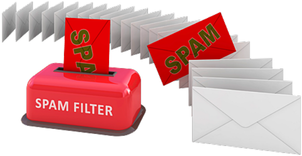


Data: Analogue speech signals (time series numerical data)

Aim: Convert audio into text (think Echo/Siri...)

1. Pre-processing Digitisation
2. Feature Selection Wave amplitude, frequencies
3. Inference Hidden Markov Models (Viterbi algorithm) [or Deep learning]

Ex3. Spam Filter



Data: Email texts

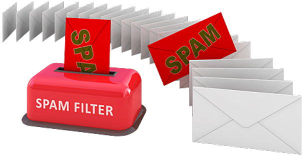
Aim: Determine whether the email is spam

1. Pre-processing - Normalise words
2. Feature Selection - Presence of words

Select subset of words w_i and determine $P(w_i \mid spam)$ and $P(w_i \mid \neg spam)$ from frequencies in training data.

Image from https://www.kdnuggets.com/

Ex3. Spam Filter



Data: Email texts

Aim: Determine whether the email is spam

1. Pre-processing - Normalise words
2. Feature Selection - Presence of words
3. Classification - Naive Bayes classifier

Select subset of words w_i and determine $\boxed{P(w_i \mid spam)}$ and $\boxed{P(w_i \mid \neg spam)}$ from frequencies in training data.

For an Email that contains $w_1, w_2, .., w_n$ of the subset of words, assume

$$P(email \mid spam) = P(w_1 \mid spam)P(w_2 \mid spam)..P(w_n \mid spam) \qquad (1)$$

and

$$P(email \mid \neg spam) = P(w_1 \mid \neg spam)P(w_2 \mid \neg spam)..P(w_n \mid \neg spam) \qquad (2)$$

A new Email is spam if

$$\underline{P(email \mid spam)} > \underline{P(email \mid \neg spam)} \qquad (3)$$

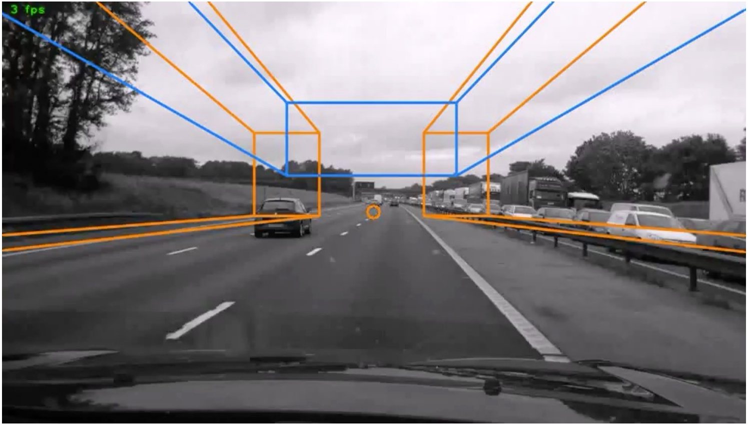
Image from https://www.kdnuggets.com/

Ex4.1 – Towards Autonomous Driving

Data: Video

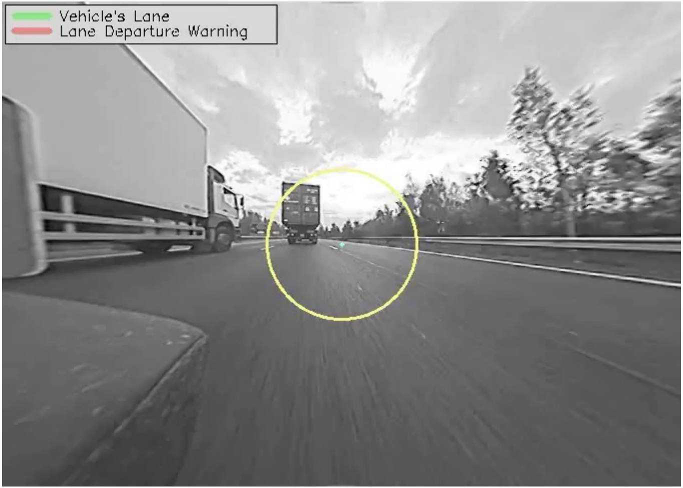
Aim: Determine knowledge from the road or inside the vehicle

1. Pre-processing (Detect vanishing point)
2. Feature Selection (Use constraints to reduce number and dimensionality)
3. Recognition (Perspective transformations and OCR)



Ex4.2 – Towards Autonomous Driving

1. Pre-processing (Detect vanishing point)
2. Feature Selection (Straight lines)
3. Model Building (Detecting, predicting, decision making)



Ex4.3 – Towards Autonomous Driving

1. Pre-processing (Detect vanishing point)
2. Feature Selection (MSERs, Histogram of Gradients)
3. Classification (Support Vector Machines)

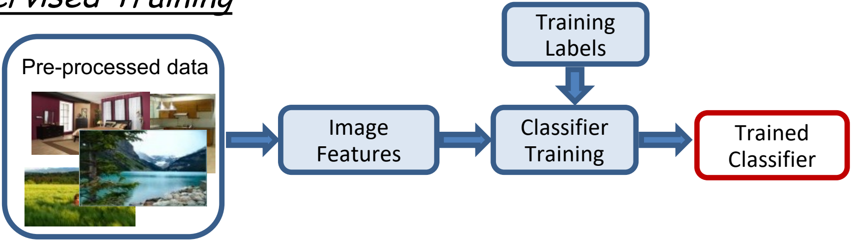
Ex4.4 – Towards Autonomous Driving

1. Pre-processing (Background subtraction)
2. Feature Selection (hand shapes)
3. Classification (Random Forest classifier)

Summary: Typical Data Analysis Problem

1. Pre-processing
2. Feature Selection
3. Modelling & Classification

Supervised Training



Pre-processed data

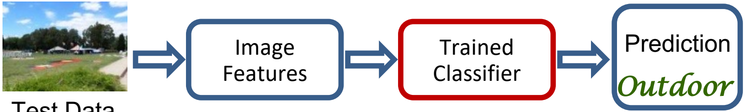
Image Features

Training Labels

Classifier Training

Trained Classifier

Testing



Test Data

Image Features

Trained Classifier

Prediction

Outdoor

COMS20011 Unit Online

Labs

- Thursdays 13:00 - 14:00 [by timetable]: Group 1
- Thursdays 14:00 - 15:00 [by timetable]: Group 2
- Lab Environment [Jupyter + Python]
- TA support in Teams: **grp-COMS20011_2021**
- Labs are essential for learning unit content!

Unit pages : https://github.com/LaurenceA/COMS20011_2021