Computational Social Science with Images and Audio

Elliott Ash, **Philine Widmer**

17 November 2023

Recap from computer vision: For classical ML, we (often) extract features explicitly.

- Recall the typical pipeline for classical machine learning:

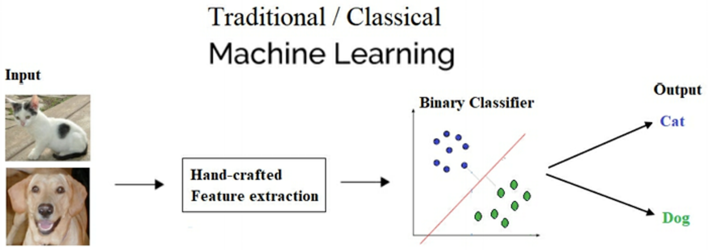

Figure: Dey (2018)[1]

[1] Dey, S. (2018). Hands-On Image Processing with Python: Expert Techniques for Advanced Image analysis and Effective Interpretation of Image Data. Packt Publishing Ltd.

We can use a similar workflow for audio data.

- Recall the typical pipeline for classical machine learning:

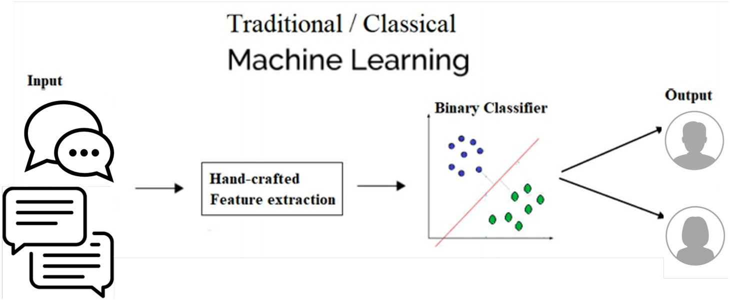

Figure: Own Adaptation of Dey (2018)[2]

[2] Dey, S. (2018). Hands-On Image Processing with Python: Expert Techniques for Advanced Image analysis and Effective Interpretation of Image Data. Packt Publishing Ltd.

Can we use the raw audio input for classification tasks?

- Neural networks, such as CNN, can handle raw audio data $\rightarrow$ extract relevant features automatically (cf. images part)
- For very simple tasks, it may be possible with classical machine learning
  - Can you think of an example?
- More generally, audio waveforms are high-dimensional data, especially for long clips
  - Computationally demanding
  - And often not necessary $\rightarrow$ suitable format depends on task
- For both classical and neural approaches, feature extraction is often useful

Mel-Frequency Cepstral Coefficients (MFCCs) are often used for feature extraction.

- Humans perceive sound frequencies non-linearly (rather, logarithmically)
- Human sensitivity is greater to changes in lower frequencies
- The Mel scale is a way to measure pitch that matches how we hear sounds
  - It is a perceptual scale of pitches judged to be equal in distance from one another
  - Originally derived from experiments with human listeners
- MFCCs represent the power spectrum of an audio signal more in line with human hearing
  - They use the Mel scale

MFCCs capture the short-term power spectrum of sound.

- One typically begins by dividing the audio signal into short (overlapping) frames, for instance 20-40 ms
  - This allows assuming stationarity within each frame
  - Stationarity means that the statistical properties of the signal (like mean, variance) are constant over the frame's duration
- Several processing steps involved
- The bottom line is that we typically end up with 12-13 coefficients per frame
  - Empirically found to capture the most important features

With MFCCs, do we dimension-reduce?

- Assume an audio clip of 0.1 seconds (100 milliseconds)
- What's the dimension of the raw wave?
- How often can we take a 20ms frame with an overlap of, say, 50%?
- Now, each frame comes with 12-13 coefficients...

Spectral features capture important characteristics of a sound's frequency content.

- Spectral features are widely used for audio tasks (e.g., music analysis, speech processing, audio classification)
- Typically, they encompass spectral centroid, spectral bandwidth, spectral flatness, or spectral roll-off
- The spectral centroid shows the frequency spectrum's "center of gravity"
  - Calculated as the weighted mean of the frequencies in the signal, with their amplitudes being the weights
  - Higher values reflect brighter sound
  - Technically, we cut the sound into frames and apply a Fourier Transform per frame
  - Hence, we transform the frame's time-domain signal into the frequency-domain

Some background: Fourier Transforms are omnipresent in audio analysis.

- The Fourier Transform decomposes a time-domain signal into its constituent frequencies
- Let us discrete time-domain signal $y[n]$, the Discrete Fourier Transform (DFT) is defined as:

$$Y[k] = \sum_{n=0}^{N-1} y[n] \cdot e^{-j\frac{2\pi}{N}kn}$$

- Definitions:
  - $y[n]$ is the audio wave amplitude at time n
  - N is the total number of samples
  - $Y[k]$ is the amplitude of the frequency component at frequency k
  - $e^{-j\frac{2\pi}{N}kn}$ is the complex exponential function

How does the DFT work in practice?

Some quiz questions.

- Consider an audio wave $y[n]$ with just a single frequency (e.g., a sine wave):
  - What will the Fourier Transform show?
- Audio waves are continuous signals. Why are we using the DFT?
- In practice, we may not know which frequency to look for. What do we do?

Now, back to spectral features beyond the spectral centroid.

- The spectral bandwidth describes how wide the frequency band is
  - A wider bandwidth implies a broad range of frequencies
  - Conversely, a narrower bandwidth indicates a more tonally pure sound
  - It can help, for example, to identify the complexity of sound
- The spectral flatness indicates how "noise-like" a sound is instead of being tonal
  - Values close to 1 indicate a noise-like sound
  - Values near 0 indicate a more tonal sound
  - It can be helpful, for instance, to disentangle tones and environmental noises