Deltas and Delta-deltas

由 Tom Bäckström创建, 最后修改于四月 16, 2019

In recognition tasks, such as phoneme recognition or voice activity detection, a classic input feature are the mel-frequency cepstral coefficients (MFCCs). They describes the instantaneous, spectral envelope shape of the speech signal. However, speech signals are time-variant signals and in a constant flux. Though we describe speech in linguistics as concatenated sequences of phonemes, the acoustical signal is more accurately described as a sequence of transitions between phonemes.

The same observation applies to other features of speech like the fundamental frequency (F0), which describes an instantaneous value. However, it is often more informative to analyse the overall shape of the F0 track, than the absolute value. For example emphasis in a sentence is often encoded with a distinct high-low contrast in F0 and questions have in many languages a characteristic low-high F0 contour.

A common method for extracting information about such transitions is to determine the first difference of signal features, known as the *delta* of a feature. Specifically, for a feature f_k, at time-instant k, the corresponding delta is defined as

$$\Delta_k = f_k - f_{k-1}.$$

The second difference, known as the delta-delta, is correspondingly

$$\Delta\Delta_k = \Delta_k - \Delta_{k-1}.$$

Common short-hand notations for the deltas and delta-deltas are, respectively, Δ and $\Delta\Delta$ -features. Features in a recognition engine are then typically appended by their Δ and $\Delta\Delta$ -features to triple the number of features with a very small computational overhead.

A trivial observation/interpretation of the delta and delta-delta features is that they approximate first and second derivatives of the signal. As estimates of the derivatives, they are not particularly accurate, but their simplicity probably makes up for that. The issue with accuracy is that differentiators tend to amplify white noise, whereas the desired signal remains unchanged. Consequently, the output is more noisy than the original signal. Differentiation is applied twice in the delta-delta feature such that issues with noise are also accumulated.

Note that the delta-features are linear transforms of the input features, such that if they are combined with a linear layer in a subsequent neural network, then in principle, the two consecutive linear layers are redundant. However, using delta-features can still provide a benefit in convergence.

In any case, delta- and delta-delta features are a classic component of machine learning algorithms. They are successful because they are very simple to calculate and provide often a clear benefit over the instantaneous features.