Song segmentation via ordinal linear discriminant analysis

Brian McFee

August 19, 2013

1 Introduction

2 Background: Fisher's linear discriminant

Let $\mathcal{X} := \{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times [m]$ denote a labeled dataset of n points belonging to m distinct categorical labels. Fisher's linear discriminant analysis (FDA) finds the direction(s) that maximizes the between-class scatter, while minimizing the scatter within each class.

Formally, let $C_\ell = \{x_i | \; y_i = \ell\}$ denote the subset of points belonging to the ℓth class, and define the following empirical quantities:

$$\begin{aligned}\mu_\ell &:= \frac{1}{|C_\ell|} \sum_{x \in C_\ell} x\\ \Sigma_\ell &:= \frac{1}{|C_\ell|} \sum_{x \in C_\ell} (x - \mu_\ell)(x - \mu_\ell)^\mathsf{T}.\end{aligned}$$

When $m = 2$, FDA solves the following problem:

$$\max_w \quad \frac{w^\mathsf{T} S_B w}{w^\mathsf{T} S_W w}$$

where

$$\begin{aligned}\mu &:= \frac{1}{n} \sum_{\mathcal{X}} x\\ S_B &:= \sum_\ell |C_\ell| (\mu_\ell - \mu)(\mu_\ell - \mu)^\mathsf{T}\\ S_W &:= \sum_\ell |C_\ell| \Sigma_\ell.\end{aligned}$$

More generally, for $m > 2$ classes, FDA solves

$$\max_W \operatorname{tr} \left[\left(W^\mathsf{T} S_W W \right)^{-1} W^\mathsf{T} S_B W \right],$$

which reduces to a generalized eigenvalue problem.

3 Ordinal linear discriminant analysis

General FDA assumes no structure on the label space, or on linkage constraints in the data. As such, it is not directly applicable to optimizing features for sequentially constrained clustering, where each "class" in the data corresponds to a segment of audio, and the temporal structure between class label is highly informative.

While the notion of minimizing within-class scatter should carry over, maximizing between-class scatter is more problematic, for two reasons. First, a segment will only ever be directly compared with temporally neighboring segments, so it is inappropriate to compute scatter relative to the entire dataset. Second, in music especially, it is likely that a segment C_i may repeat later in the song as another segment C_{i+k}, and attempting to separate these classes can actually introduce error.

To correct these issues, we exploit the ordinal structure of the labels by defining the *ordinal scatter*:

$$S_O := \sum_{\ell < m} |C_\ell| (\mu_\ell - \mu_{\ell+})(\mu_\ell - \mu_{\ell+})^\mathsf{T},$$

where

$$\mu_{\ell+} := \frac{|C_\ell|\mu_\ell + |C_{\ell+1}|\mu_{\ell+1}}{|C_\ell| + |C_{\ell+1}|}$$

is the mean of the joined segment $C_\ell \cup C_{\ell+1}$.

We then define the ordinal linear discriminant analysis (OLDA) optimization accordingly:

$$\max_W \operatorname{tr}\left[\left(W^\mathsf{T} S_W W\right)^{-1} W^\mathsf{T} S_O W \right],$$

which again reduces to a generalized eigenvalue problem. The only difference between OFDA and FDA is that the between-class scatter matrix is replaced with a temporally-constrained and localized variant, so the resulting optimization only tries to separate adjacent segments. While the formulation above assumes a single ordering structure over labels (*i.e.*, the segmentation of one song), it is trivial to incorporate multiple segmentations from one or more songs by accumulating song-level matrices S_W and S_O.

In practice, for large d, the within-class scatter S_W may become rank-deficient, resulting in an ill-conditioned problem. In that case, we may replace S_W by a ridge-regularized variant

$$S_W \mapsto S_W + \sigma I,$$

for $\sigma > 0$.

After solving for W, a data matrix $X \in \mathbb{R}^{d \times n}$ can then be transformed according to

$$X \mapsto W^\mathsf{T} X,$$

and subsequently clustered.

4 Clustering algorithm

4.1 Graph-constrained agglomerative clustering

4.2 Graph-constrained Hartigan clustering (cleanup)

5 Features

5.1 Latent repetition features