ORIE 779: Functional Data Analysis

From last meeting

Finished ICA

Began Statistical Discrimination (i.e. Classification)

"automatic diagnosis"

- Naïve methods (Mean Difference, PCA)
- Fisher Linear Discrimination

Fisher Linear Discrimination (cont.)

Relationship to "Mahalanobis distance"

Idea: For $X_1, X_2 \sim N(\mu, \Sigma)$, a "natural distance measure" is:

$$d_M(X_1, X_2) = (X_1 - X_2)^t \Sigma^{-1} (X_1 - X_2)$$

- "unit free", i.e. "standardized"
- essentially "mod out" covariance structure
- Euclidean distance applied to $\Sigma^{-1/2} X_1$ & $\Sigma^{-1/2} X_2$
- Same as key transformation for FLD
- I.e. FLD is "mean difference in Mahalanobis space"

FLD Likelihood View

Assume: Class distributions are multivariate $N\left(\underset{\to}{\mu}^{(j)}, \Sigma^w\right)$

(strong distributional assumption + *common covariance*)

At a location $\underset{\to}{x}^0$, the likelihood ratio,

for choosing between Class 1 and Class 2, is:

$$LR\left(\underset{\to}{x}^0, \underset{\to}{\mu}^{(1)}, \underset{\to}{\mu}^{(2)}, \Sigma^w\right) = \varphi_{\Sigma^w}\left(\underset{\to}{x}^0 - \underset{\to}{\mu}^{(1)}\right) / \varphi_{\Sigma^w}\left(\underset{\to}{x}^0 - \underset{\to}{\mu}^{(2)}\right)$$

where φ_{Σ^w} is the Gaussian density with covariance Σ^w

FLD Likelihood View (cont.)

Simplifying, using the form of the Gaussian density:

$$\varphi_{\Sigma^w}(\underrightarrow{x}) = \frac{1}{(2\pi)^{d/2}\left|\Sigma^w\right|} e^{-\left(\underrightarrow{x}^t \Sigma^{w-1} \underrightarrow{x}\right)/2}$$

Gives (critically using the common covariance):

$$LR\left(\underrightarrow{x}^0, \underrightarrow{\mu}^{(1)}, \underrightarrow{\mu}^{(2)}, \Sigma^w\right) = e^{-\left[\left(\underrightarrow{x}^0 - \underrightarrow{\mu}^{(1)}\right)^t \Sigma^{w-1}\left(\underrightarrow{x}^0 - \underrightarrow{\mu}^{(1)}\right) - \left(\underrightarrow{x}^0 - \underrightarrow{\mu}^{(2)}\right)^t \Sigma^{w-1}\left(\underrightarrow{x}^0 - \underrightarrow{\mu}^{(2)}\right)\right]/2}$$

$$-2\log LR\left(\underrightarrow{x}^0, \underrightarrow{\mu}^{(1)}, \underrightarrow{\mu}^{(2)}, \Sigma^w\right) =$$

$$= \left(\underrightarrow{x}^0 - \underrightarrow{\mu}^{(1)}\right)^t \Sigma^{w-1}\left(\underrightarrow{x}^0 - \underrightarrow{\mu}^{(1)}\right) - \left(\underrightarrow{x}^0 - \underrightarrow{\mu}^{(2)}\right)^t \Sigma^{w-1}\left(\underrightarrow{x}^0 - \underrightarrow{\mu}^{(2)}\right)$$

FLD Likelihood View (cont.)

But:

$$\left(\underline{x}^0 - \underline{\mu}^{(j)}\right)^t \Sigma^{w^{-1}} \left(\underline{x}^0 - \underline{\mu}^{(j)}\right) = \underline{x}^{0^t} \Sigma^{w^{-1}} \underline{x}^0 - 2\underline{x}^{0^t} \Sigma^{w^{-1}} \underline{\mu}^{(j)} + \underline{\mu}^{(j)} \Sigma^{w^{-1}} \underline{\mu}^{(j)}$$

so:

$$-2\log LR\left(\underline{x}^0, \underline{\mu}^{(1)}, \underline{\mu}^{(2)}, \Sigma^w\right) =$$

$$= -2\underline{x}^{0^t} \Sigma^{w^{-1}} \left(\underline{\mu}^{(1)} - \underline{\mu}^{(2)}\right) + \left(\underline{\mu}^{(1)} + \underline{\mu}^{(2)}\right) \Sigma^{w^{-1}} \left(\underline{\mu}^{(1)} - \underline{\mu}^{(2)}\right)$$

Thus $LR\left(\underline{x}^0, \underline{\mu}^{(1)}, \underline{\mu}^{(2)}, \Sigma^w\right) \geq 1$ when

$$-2\log LR\left(\underline{x}^0, \underline{\mu}^{(1)}, \underline{\mu}^{(2)}, \Sigma^w\right) \leq 0$$

i.e.

$$\underline{x}^{0^t} \Sigma^{w^{-1}} \left(\underline{\mu}^{(1)} - \underline{\mu}^{(2)}\right) \geq \frac{1}{2} \left(\underline{\mu}^{(1)} + \underline{\mu}^{(2)}\right) \Sigma^{w^{-1}} \left(\underline{\mu}^{(1)} - \underline{\mu}^{(2)}\right)$$

FLD Likelihood View (cont.)

Replacing $\underset{\to}{\mu}^{(1)}$, $\underset{\to}{\mu}^{(2)}$ and Σ^{w} by maximum likelihood estimates:

$$\underset{\to}{\overline{X}}^{(1)}, \ \underset{\to}{\overline{X}}^{(2)} \text{ and } \hat{\Sigma}^{w}$$

gives the likelihood ratio discrimination rule:

Choose Class 1, when

$$\underset{\to}{x}^{0^t} \hat{\Sigma}^{w^{-1}} \left(\underset{\to}{\overline{X}}^{(1)} - \underset{\to}{\overline{X}}^{(2)} \right) \leq \frac{1}{2} \left(\underset{\to}{\overline{X}}^{(1)} + \underset{\to}{\overline{X}}^{(2)} \right) \hat{\Sigma}^{w^{-1}} \left(\underset{\to}{\overline{X}}^{(1)} - \underset{\to}{\overline{X}}^{(2)} \right)$$

same as above, so FLD can be viewed as "Likelihood ratio rule"

FLD Generalization I

Gaussian Likelihood Ratio Discrimination

(a. k. a. "nonlinear discriminant analysis")

Idea: Assume class distributions are $N\left(\underset{\longrightarrow}{\mu}^{(j)}, \Sigma^{(j)}\right)$

Different covariances!

Likelihood Ratio rule is straightforward numerical calculation

(thus can easily implement, and do discrimination)

FLD Generalization I (cont.)

But no longer have "separating hyperplane" representation

(instead "regions determined by quadratics")

(fairly complicated case-wise calculations)

Graphical display: for each point, color as:

Yellow if assigned to Class 1

Cyan if assigned to Class 2

("intensity" is "strength of assignment")

Illustrate with FLD for canonical [Toy Example]

FLD Generalization I (cont.)

Toy Examples:

1. Standard Tilted Point clouds [graphic]:

   - Both FLD and LR work well.

2. Donut:

   - [FLD] poor (no separating plane can work)

   - [LR] much better

FLD Generalization I (cont.)

3. Split X: [FLD] [LR]

   - neither works well

   - although ∃ good separating quadratic surfaces

   - they are not “from Gaussian likelihoods”

   - so this is not “general quadratic discrimination”

FLD Generalization II

Different prior probabilities

Main idea: Give different weights to 2 classes

I.e. assume *not* a priori equally likely

Development is “straightforward”

- modified likelihood
- change intercept in FLD

Won’t explore further here

FLD Generalization III

Principal Discriminant Analysis

Idea: FLD-like approach to more than two classes

Assumption: Class covariance matrices are the *same* (similar)

(but not Gaussian, same situation as for FLD)

Main idea: quantify "location of classes" by their means

$$\underset{\rightarrow}{\mu}^{(1)}, \ \underset{\rightarrow}{\mu}^{(2)}, \ \ldots, \ \underset{\rightarrow}{\mu}^{(k)}$$

FLD Generalization III (cont.)

Simple way to find "interesting directions" among the means:

PCA on set of means

i.e. Eigen-analysis of "between class covariance matrix"

$$\Sigma^B = MM^t$$

where

$$M = \frac{1}{\sqrt{k}}\left(\underset{\rightarrow}{\mu}^{(1)} - \underset{\rightarrow}{\overline{\mu}} \quad \cdots \quad \underset{\rightarrow}{\mu}^{(k)} - \underset{\rightarrow}{\overline{\mu}}\right)$$

Aside: can show: overall $\sqrt{n}\Sigma = \sqrt{n}\Sigma^B + \sqrt{k}\Sigma^w$

FLD Generalization III (cont.)

But PCA only works like “mean difference”,

Expect can improve by “taking covariance into account”.

(recall FLD illustration)

Blind application of above ideas suggests eigen-analysis of:

$$\Sigma^{w^{-1}}\Sigma^{B}$$

FLD Generalization III (cont.)

There are:

- smarter ways to compute ("generalized eigenvalue")
- other representations (this solves optimization prob's)

Special case: 2 classes,   reduces to standard FLD

Good reference for more: Section 3.8 of:

Duda, R. O., Hart, P. E. and Stork, D. G. (2001) *Pattern Classification*, Wiley.

Fisher Linear Discrimination (cont.)

Corpora Callosa application:

Recall data: Schizophrenics Controls

Movie display of FLD direction vector and projections

- Great separation of subpopulations?!?
- Image doesn't change when marching along vector?!?

Corpora Callosa Fisher Linear Discrimination

Major problem: $n = 71 < 80 = d$:

- gives "directions of perfect separation" (~8 dim subspace!)
- $\exists$ a very small change in this direction (watch pixels)
- numerics: use pseudo-inverse of covariance matrix
- is FLD direction interesting or useful?

Corpora Callosa Fisher Linear Discrimination (cont.)

Zoom in on FLD direction:

- Only pixel sampling artifacts
- Expect big changes with new data
- Direction neither useful nor insightful

Big Picture View

This motivate new area of statistical analysis:

High Dimension - Low Sample Size (HDLSS)

Idea: face common Problem: $n << d$

Standard Approach to HDLSS

"Dimensionality Reduction"

- Find *some way* to "reduce dimension"
- Problem: how to do this?
- PCA is viewed as leading method
- Has obvious limitations
- Expect some "loss"
- Major question: when is that fatal????

A dimensionality approach to CCF data

Completely different data representation:

E.g. recall from Lecture 2-20-02:

Medial Representation of Corpora Callosa data

Yushkevich, P., Pizer, S. M., Joshi, S., and Marron, J. S. (2001) "Intuitive, Localized Analysis of Shape Variability", *Information Processing in Medical Imaging (IPMI)*, eds: Insana, M. F. and Leahy. R. M., 402-408. [http://www.cs.unc.edu/~pauly/ipmi2001/]

Idea: discrete "skeleton" of shape

Summarization: features are "location and angle parameters"

A dimensionality approach to CCF data

Recall: Raw data

- from same data as above Fourier boundary rep'n
- but they look different
- since different type of fitting was done
- also, worst outlier was deleted

Modes of variation? Recall PCA

PC1 PC2 PC3 PC4

A dimensionality approach to CCF data (cont.)

Some benefits from medial representation:

- "more efficient use of parameters"
- far fewer "parameters with little information"
- such as "high frequency" Fourier coefficients
- results in good representation with fewer parameters
- Here: Fourier $d = 80$ reduced to M-rep $d = 20$
- No longer have HDLSS, now have $d = 20 < n = 31, 40$
- Practical benefits? (maybe not "far from HDLSS"?)

A dimensionality approach to CCF data (cont.)

Toy Examples: simulated Corpora Collosa data sets ($n = 25$)

Simulated data set 1: (from Gaussian pop'n "like controls")

Simulated data set 2: (like 1, but "less overall bending")

Simulated data set 3: (like 1, but "bump on top center")

Reasons:

- Want to study *known* differences
- Unsure about differences in Schizophrenics vs. Controls
- Are there any? (in sense of statistical significance)

A dimensionality approach to CCF data (cont.)

Simulated Data 1 vs. Simulated Data 2:

FLD direction:

- Doesn't find "overall bending" direction
- Small change suggests "spurious direction"?
- Because of "near HDLSS setting"?

Mean Difference Direction:

- Found "overall bending" (as constructed)
- Seems more stable in "near HDLSS setting"?

A dimensionality approach to CCF data (cont.)

Simulated Data 1 vs. Simulated Data 2 (cont.):

Projection views: FLD Mean Difference

- Similar amounts of “separation of subpopulations”
- But FLD is slightly more separated?
- But FLD is “smaller scale effect” (see x-axes)
- So Mean Difference found “better separation”?
- Much less likely to be spurious

A dimensionality approach to CCF data (cont.)

Simulated Data 1 vs. Simulated Data 3:

FLD Direction:

- Keys on “width of 4th medial atom”
- Not on “bump in center” (the constructed difference)
- Again missed due to “near HDLSS”?

Mean Difference Direction:

- Nicely finds “bump in center”
- Again seems more stable in “near HDLSS situations”

A dimensionality approach to CCF data (cont.)

Simulated Data 1   vs.   Simulated Data 3 (cont.):

Projection Views: FLD   Mean Difference

- FLD seems to give "better separation"
- But again note an order of magnitude smaller
- So FLD again found a "spurious sampling direction"
- Again seems unstable for this "near HDLSS setting"