Simplified Quadratic Discriminant Analysis

Ursula Garczarek

March 5, 2005

Overview

- where to apply
- how to apply
- what it means
- how it looks

Where to apply

sQDA is meant for

- two-class classification based on
- high dimensional data
- with variables measuring *intensity*,

as e.g. when diagnosing with gene array data, mass spectrometry data,...

How to apply

Given some training data

1. take the logarithm of the pure intensities

2. determine univariate discriminating features,

3. subdivide them in under-expressed and over-expressed features

4. standardize them appropriately

5. average the over-expressed, average the under-expressed

6. do classical QDA on the thus generated two features

Validate the rule on extra data...

Appropriately standardizing single features

Given the following notation for the standardization task:

$$s(i_{vn}) := w_v \left(\log(i_{vn}) - m_v\right)$$

with i_{vn}: intensity value of individuum n in variable v
we search for parameters that are easy to calculate and some

- m_v that provides some meaningful "mean"
- w_v that weights distances to that mean according to some notion of dispersion

Meaningful mean

On the training data

1. calculate groupwise arithmetic mean:

$$m_{vg} := \frac{1}{N_g} \sum_{n \in \mathcal{G}} \log(i_{vn}) = \log\left(\left(\prod_{n \in \mathcal{G}} i_{vn}\right)^{1/N_g}\right) =: \log\left(i_{vg}\right)$$

2. Calculate balance point of groupwise means

$$m_v := \frac{1}{2}\left(m_{v0} + m_{v1}\right)$$

3. Difference with balance point gives log of relative intensity:

$$\log(i_{vn}) - m_v = \log\left(\frac{i}{\sqrt{i_{v0} * i_{v1}}}\right)$$

Meaningful mean

Justifying

$$m_{vg} \quad =: \quad \log\left(i_{vg}\right)$$

- geometric mean

$$i_{vg} \quad := \quad \left(\prod_{n \in \mathcal{G}} i_{vn}\right)^{1/N_g}$$

- if $\log(I)$ symmetrically distributed

$$\begin{aligned} \frac{1}{N_g} \sum_{n \in \mathcal{G}} \log(i_{vn}) \quad &\approx \quad \textbf{median}\left\{\log(i_{vn}), n \in \mathcal{G}\right\} \\ &= \quad \log\left(\textbf{median}\left\{i_{vn}, n \in \mathcal{G}\right\}\right) \end{aligned}$$

One way to do weighting...

1. Calculate groupwise standard deviation:

$$d_{vg} := \sqrt{\frac{1}{N_g} \sum_{n \in \mathcal{G}} \left(\log(i_{vn}) - \log\left(i_{vg}\right)\right)}$$

2. Calculate inverse of balance point of groupwise standard deviations and adjust for sizes of $\mathcal{U}$ and $\mathcal{O}$

$$w_v \quad := \quad \begin{cases} \frac{2}{|\mathcal{U}| * (d_{v0} + d_{v1})}, v \in \mathcal{U} \\ \frac{2}{|\mathcal{O}| * (d_{v0} + d_{v1})}, v \in \mathcal{O} \end{cases}$$

Level of under- and over-expression of some sample

$$\begin{aligned} u(\vec{i}) &= \sum_{v \in \mathcal{U}} w_v \left(\log(i_{vn}) - m_v\right) \\ &= \sum_{v \in \mathcal{U}} w_v \log\left(\frac{i_v}{\sqrt{i_{v0} * i_{v1}}}\right) \\ o(\vec{i}) &= \sum_{v \in \mathcal{O}} w_v \log\left(\frac{i_v}{\sqrt{i_{v0} * i_{v1}}}\right) \end{aligned}$$

The final discriminant function and its components

The discrimination is based on the Mahalanobis distance to the mean level of under- and over-expression in the diseased group compared to respective levels in the control group.

$$\mathbf{df}\left(\begin{bmatrix} u \\ o \end{bmatrix}\right) = -a\left(\mathbf{MD}\left(\begin{bmatrix} u \\ o \end{bmatrix}, \begin{bmatrix} \bar{u}_1 \\ \bar{o}_1 \end{bmatrix}\right) - \mathbf{MD}\left(\begin{bmatrix} u \\ o \end{bmatrix}, \begin{bmatrix} \bar{u}_0 \\ \bar{o}_0 \end{bmatrix}\right)\right) + b$$

with

$a \in \mathbb{R}^+, b \in \mathbb{R}$

MD: Mahalanobis distance

u, o: levels of under- and over-expression in some individual

$\bar{u}_g, \bar{o}_g$: mean levels of under- and over-expression in group g

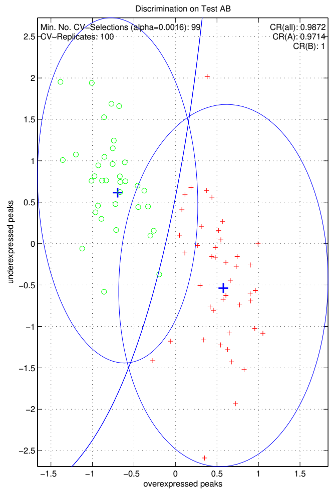
The centers of sQDA

$$\begin{aligned} \bar{u}_0 &= \frac{1}{N_0} \sum_{n \in \mathcal{C}} \sum_{v \in \mathcal{U}} w_v \left(\log(i_{vn}) - m_v\right) \\ &= \ldots \\ &= \ldots \\ &= \frac{1}{2} \sum_{v \in \mathcal{U}} w_v \left(\log\left(\frac{i_{v0}}{i_{v1}}\right)\right) >= 0 \end{aligned}$$

This results in

$$\begin{array}{c|c} \begin{array}{c} \bar{u}_0 > 0 \\ \bar{o}_0 < 0 \end{array} & \\ \hline & \begin{array}{c} \bar{u}_1 < 0 \\ \bar{o}_1 > 0 \end{array} \end{array}$$

What it looks like



Side notes

- With the proposed weighting, a subset of highly correlated features gets too much weight. Either reduce set of features then, or down-weight according to mean groupwise partial correlation with all other variables in either $\mathcal{U}$ or $\mathcal{O}$.
- QDA and not LDA is proposed for diagnostic questions as typically the population of "non-diseased" and the population of diseased are quite different.
- I assume, sQDA can be generalized as some QDA with certain restrictions on parameters. May be a way to find "optimal" weights.