Regularized Discriminant Analysis: A Large Dimensional Study

ISIT 2018

Xiaoke Yang
Khalil Elkhalil
Abla Kammoun
Tareq Y. Al-Naffouri
Mohamed-Slim Alouini

June 19, 2018

King Abdullah University of Science and Technology

KAUST
King Abdullah University of Science and Technology

Table of contents

1. Gaussian discriminant analysis

2. Asymptotic Performance of RDA

3. Discussion

4. Conclusions

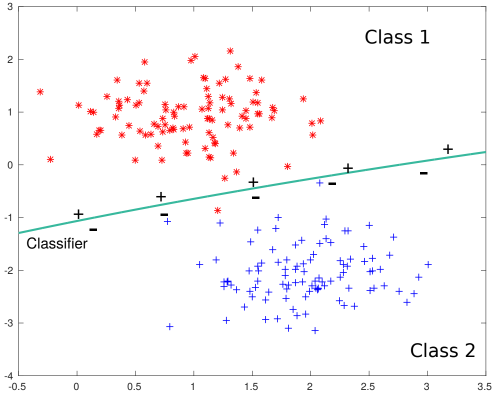
Gaussian discriminant analysis

Classification

Classification is the task of assigning a label to an input data among a set of possible categories.

Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-Based Learning Applied To Document Recognition. Proceedings of the IEEE, 86(11):22782324, 1998.

- Principle: Build a classification rule that allows to assign for an unseen observation its corresponding class.



Let $\mathbf{x}$ be the input data and f be the classification rule.

$$\text{Classifier} \triangleq \begin{cases} \text{Assign class 1} & \text{if} \quad f(\mathbf{x}) < 0 \\ \text{Assign class 2} & \text{if} \quad f(\mathbf{x}) > 0 \end{cases}$$

Model based classification

- Data is assumed to be sampled from a certain dist.
- The decision rule is constructed based on that.
- The MAP rule is considered in the design

$$\boxed{\widehat{k} = \arg \max_{k:classes} \mathbb{P}\left[\mathcal{C}_k | \mathbf{x}\right]}$$

The classifier is designed to satisfy this rule.

Gaussian mixture model for binary classification (2 classes)

- $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^p$
- **Class** k **is formed by** $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ **,** $k = 0, 1$

Linear discriminant analysis (LDA) $(\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1)$

$$\boxed{W^{LDA}(\mathbf{x}) = \left(\mathbf{x} - \frac{\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1}{2}\right)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) - \log \frac{\pi_1}{\pi_0}} \tag{1}$$

$$\begin{cases} \text{Assign } \mathbf{x} \text{ to class } 0 & \text{if } W^{LDA} > 0 \\ \text{Assign } \mathbf{x} \text{ to class } 1 & \text{otherwise} \end{cases}$$

$\rightarrow$ **Decision rule is linear in x.** .

Quadratic discriminant analysis $(\boldsymbol{\Sigma}_0 \neq \boldsymbol{\Sigma}_1)$

$$\boxed{W^{QDA}(\mathbf{x}) = -\frac{1}{2}\log\frac{|\boldsymbol{\Sigma}_0|}{|\boldsymbol{\Sigma}_1|} - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\mathbf{x}-\boldsymbol{\mu}_0) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)} \tag{2}$$

$$\begin{cases} \text{Assign} \quad \mathbf{x} \quad \text{to class} \quad 0 \quad \text{if} \quad W^{QDA}(\mathbf{x}) > 0 \\ \text{Assign} \quad \mathbf{x} \quad \text{to class} \quad 1 \quad \text{otherwise} \end{cases}$$

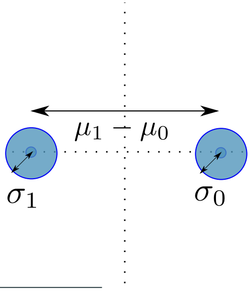
$\rightarrow$ **Decision rule is quadratic in x.**

LDA: finite regime (statistics are known)

- Assume $\boldsymbol{\Sigma}$, $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ known.
- Equal priors : $\pi_0 = \pi_1 = 0.5$
- No asymptotic regime, p is fixed.

The total misclassification rate is given by [1]

$$\epsilon = \Phi\left(-\frac{\Delta}{2}\right), \quad \Delta = \sqrt{(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)} = \|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|_{\boldsymbol{\Sigma}^{-1}}$$



[1] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. The Elements of Statistical Learning. Springer, 2009.

For Gaussian data, LDA and QDA are respectively the *best* classifiers (with the smallest possible risk) in the case of equal covariances and distinct covariances!

How things will look like with noisy estimates of the statistics?

RDA

Consider the supervised setting

$$\bar{\mathbf{x}}_i = \frac{1}{n_i} \sum_{\mathbf{x}_j \in \mathcal{C}_i} \mathbf{x}_j$$

$$\widehat{\boldsymbol{\Sigma}}_i = \frac{1}{n_i - 1} \sum_{\mathbf{x}_j \in \mathcal{C}_i} (\mathbf{x}_j - \bar{\mathbf{x}}_i)(\mathbf{x}_j - \bar{\mathbf{x}}_i)^T$$

- Introduce regularization parameter [2] $\lambda \in [0, 1]$

$$\widehat{\boldsymbol{\Sigma}}_i(\lambda) = \frac{(1-\lambda)n_i \widehat{\boldsymbol{\Sigma}}_i + \lambda n \widehat{\boldsymbol{\Sigma}}}{(1-\lambda)n_i + \lambda n}, \quad i \in \{0, 1\}$$

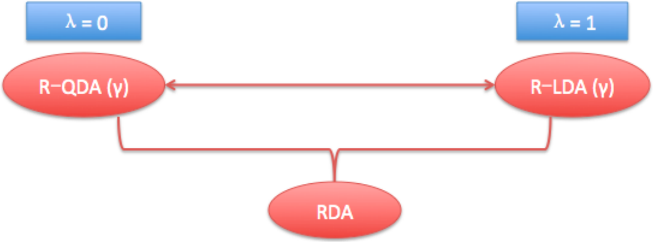
- Introduce regulariztion paramater $\gamma \in [0, 1]$ to avoid singularity

$$\boxed{\widehat{\boldsymbol{\Sigma}}_i(\lambda, \gamma) = \gamma \frac{(1-\lambda)n_i \widehat{\boldsymbol{\Sigma}}_i + \lambda n \widehat{\boldsymbol{\Sigma}}}{(1-\lambda)n_i + \lambda n} + (1-\gamma)\mathbf{I}_p.}$$

  Both R-LDA and R-QDA are special cases of RDA.
  - $\lambda = 0 \rightarrow$ R-QDA
  - $\lambda = 1 \rightarrow$ R-LDA
- Define $\mathbf{H}_i = \widehat{\boldsymbol{\Sigma}}_i^{-1}$

[2] J. Friedman, Regularized discriminant analysis, Journal of the American Sta- tistical Association, vol. 84, pp. 165{175, 1989}

RDA



The extreme cases (R-LDA and R-QDA) have been analyzed in [3] and [4]

[3] K. Elkhalil, A. Kammoun, R. Couillet, T. Y. Al-Naffouri, and M.-S. Alouini. A Large Dimensional Study of Regularized Discriminant Analysis Classifiers. ArXiv e-prints, Nov. 2017.
[4] A. Zollanvari and E. R. Dougherty, Generalized Consistent Error Estimator of Linear Discriminant Analysis, IEEE Transactions on Signal Processing, vol. 63, no. 11, pp. 28042814, June 2015

LDA: Asymptotic regime (equal covariances)

Asymptotic growth regime

Let $n = n_0 + n_1$.

- $n_0, n_1, p \to \infty$ such that $\frac{n_0}{n_1} \to 1$ and $\frac{p}{n} \to c < 1$
- $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$
- $\boldsymbol{\mu} \triangleq \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$ is such that $\|\boldsymbol{\mu}\| = O(1)$.

Under these assumptions, the misclassification rate converges to [5]:

$$\boxed{R_{LDA} - \Phi\left[-\frac{\Delta}{2}\sqrt{1-c}\right] \to_{a.s.} 0.} \tag{3}$$

→ When $c \to 1$, the misclassification rate tends to 0.5.

→ For the LDA to result in acceptable performance, we need c close to 0.

→ Because its use of the inverse of the pooled covariance matrix, the LDA applies only when $c < 1$.

[5] Cheng Wang and Binyan Jiang. On the dimension effect of regularized linear discriminant analysis, arXiv:1710.03136v1

Asymptotic Performance of RDA

RDA ($\lambda \in [0, 1)$)

- Make some simplification

$$\mathbf{H}_0 = \left[(1-\gamma)\mathbf{I}_p + \alpha_0 \widehat{\mathbf{\Sigma}}_0 + \beta_0 \widehat{\mathbf{\Sigma}}_1\right]^{-1}$$

$$\alpha_0 = \frac{n_0\gamma}{n_0 + \lambda n_1}, \beta_0 = \frac{n_1\gamma\lambda}{n_0 + \lambda n_1}$$

$$\mathbf{H}_1 = \left[(1-\gamma)\mathbf{I}_p + \alpha_1 \widehat{\mathbf{\Sigma}}_0 + \beta_1 \widehat{\mathbf{\Sigma}}_1\right]^{-1}$$

$$\alpha_1 = \frac{n_0\gamma\lambda}{n_1 + \lambda n_0}, \beta_1 = \frac{n_1\gamma}{n_1 + \lambda n_0}$$

- The Discriminant score for the RDA is:

$$\widehat{\delta}_i^{RDA}(\mathbf{x}) = \frac{1}{2}\log|\mathbf{H}_i| - \frac{1}{2}(\mathbf{x} - \overline{\mathbf{x}}_i)^T \mathbf{H}_i (\mathbf{x} - \overline{\mathbf{x}}_i) + \log \pi_i.$$

- The conditional misclassification error rate associated to class $\mathcal{C}_i$

$$\boxed{\epsilon_i^{RDA} = \mathbb{P}\left[(-1)^i\, \widehat{\delta}_0^{RDA}(\mathbf{x}) < (-1)^i\, \widehat{\delta}_1^{RDA}(\mathbf{x})\, |\mathbf{x} \in \mathcal{C}_i\right].} \qquad (4)$$

The conditional misclassification error can easily be shown to write as

$$\boxed{\epsilon_i = \mathbb{P}\left[\boldsymbol{\omega}^T \mathbf{B}_i \boldsymbol{\omega} + 2\boldsymbol{\omega}^T \mathbf{y}_i < \xi_i\right], \textit{where } \boldsymbol{\omega} \sim \mathcal{N}\left(\mathbf{0}_p, \mathbf{I}_p\right),} \tag{5}$$

$$\begin{aligned}
\mathbf{B}_i &= \boldsymbol{\Sigma}_i^{1/2}\left(\mathbf{H}_1 - \mathbf{H}_0\right)\boldsymbol{\Sigma}_i^{1/2}, \\
\mathbf{y}_i &= \boldsymbol{\Sigma}_i^{1/2}\left[\mathbf{H}_1\left(\boldsymbol{\mu}_i - \bar{\mathbf{x}}_1\right) - \mathbf{H}_0\left(\boldsymbol{\mu}_i - \bar{\mathbf{x}}_0\right)\right], \\
\xi_i &= -\log\left(\frac{|\mathbf{H}_0|}{|\mathbf{H}_1|}\right) + \left(\boldsymbol{\mu}_i - \bar{\mathbf{x}}_0\right)^T \mathbf{H}_0\left(\boldsymbol{\mu}_i - \bar{\mathbf{x}}_0\right) - \left(\boldsymbol{\mu}_i - \bar{\mathbf{x}}_1\right)^T \mathbf{H}_1\left(\boldsymbol{\mu}_i - \bar{\mathbf{x}}_1\right) + 2\log\frac{\pi_1}{\pi_0}.
\end{aligned}$$

Assumptions

1. $p, n_1, n_2 \to \infty$ with $\frac{n_i}{p} \to c\,(0, \infty)$, $\frac{n_0}{n} = \frac{n_1}{n} \to 0.5$
2. The difference in means $\boldsymbol{\mu} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$ satisfies $\|\boldsymbol{\mu}\|^2 = O(\sqrt{p})$.
3. $\|\boldsymbol{\Sigma}_i\| = O\left(1\right)$.
4. Matrix $\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1$ has at most $O(\sqrt{p})$ eigenvalues of order 1 while the remaining ones decay at an order of $\mathcal{O}\left(1/\sqrt{p}\right)$.

Central Limit Theorem (CLT)[6]

Assume $\lambda \neq 1$. Under assumptions 1-4, $\boldsymbol{\omega}^T \mathbf{B}_i \boldsymbol{\omega} + 2\boldsymbol{\omega}^T \mathbf{y}_i$ can be treated as a Gaussian random variable $\boldsymbol{z} \sim \mathcal{N}(\operatorname{tr} \mathbf{B}_i, 2 \operatorname{tr} \mathbf{B}_i^2 + 4\mathbf{y}_i^T \mathbf{y}_i)$, so the conditional classification error of RDA satisfies

$$\boxed{\epsilon_i^{RDA} - \Phi\left((-1)^i \frac{\xi_i - \operatorname{tr} \mathbf{B}_i}{\sqrt{2 \operatorname{tr} \mathbf{B}_i^2 + 4\mathbf{y}_i^T \mathbf{y}_i}}\right) \xrightarrow{\text{a.s.}} 0.} \tag{6}$$

[6] K. Elkhalil, A. Kammoun, R. Couillet, T. Y. Al-Naffouri, and M.-S. Alouini. A Large Dimensional Study of Regularized Discriminant Analysis Classifiers. ArXiv e-prints, Nov. 2017.

Recent results

Benaych and Couillet, 2016

$\mathbf{H}_i = \left(\frac{\mathbf{Y}_0 \mathbf{Y}_0^T}{p} + \frac{\mathbf{Y}_1 \mathbf{Y}_1^T}{p} + (1-\gamma)\,\mathbf{I} \right)^{-1}$: Gram matrix of mixture models [7]

Define

$$\delta_i = \frac{1}{n_i} \operatorname{tr} \mathbf{\Sigma}_i \mathbf{Q}_i, \; \tilde{\delta}_i = \frac{\alpha_i}{1+\alpha_i \delta_i}$$

where

$$\mathbf{Q}_i = \left[(1-\gamma)\mathbf{I}_p + \frac{\alpha_i}{1+\alpha_i\delta_i}\mathbf{\Sigma}_0 + \frac{\beta_i}{1+\beta_i\delta_i}\mathbf{\Sigma}_1 \right]^{-1},$$

$$\boxed{\frac{1}{p} \operatorname{tr} \mathbf{A} \left(\mathbf{H}_i - \mathbf{Q}_i \right) \to_p 0, \quad \boldsymbol{u}^T \left(\mathbf{H}_i - \mathbf{Q}_i \right) \boldsymbol{v} \to_p 0.} \tag{7}$$

[7] F. Benaych-Georges and R. Couillet, Spectral Analysis of the Gram Matrix of Mixture Models, ESAIM: Probability and Statistics, vol. 20, pp. 217237, 2016.

RDA (Asymptotic result)

With assumptions 1-4 satisfied, we have

$$\boxed{\epsilon_i - \Phi\left((-1)^i \frac{\overline{\xi}_i - \overline{b}_i}{\sqrt{2\overline{B}_i}}\right) \xrightarrow{p} 0.} \tag{8}$$

$\overline{\xi}_i$, $\overline{b}_i$ and $\overline{B}_i$ depend on the classes' statistics.

$$\overline{\xi}_i \triangleq \frac{1}{\sqrt{p}} \log \left(\frac{1+\alpha_0\delta_0}{1+\alpha_1\delta_1}\right)^{n_0} \left(\frac{1+\beta_0\delta_0}{1+\beta_1\delta_1}\right)^{n_1} + \frac{1}{\sqrt{p}}\left[\frac{\alpha_1\delta_1 n_0 - \alpha_0\delta_0 n_0}{(1+\alpha_1\delta_1)(1+\alpha_0\delta_0)} + \frac{\beta_1\delta_1 n_0 - \beta_0\delta_0 n_0}{(1+\beta_1\delta_1)(1+\beta_0\delta_0)}\right]$$
$$+ \frac{1}{\sqrt{p}} \log \frac{|\mathbf{Q}_1|}{|\mathbf{Q}_0|} + \frac{1}{\sqrt{p}} (-1)^{i+1} \boldsymbol{\mu}^T \mathbf{Q}_{1-\mathbf{i}} \boldsymbol{\mu}.$$

$$\overline{b}_i = \frac{1}{\sqrt{p}} \operatorname{tr} \boldsymbol{\Sigma}_i (\mathbf{Q}_1 - \mathbf{Q}_0).$$

$$\overline{B}_i \triangleq \frac{1}{p} \frac{2n_1\phi}{1-\left(\tilde{\delta}_1^2 + \tilde{\delta}_0^2\right)\phi} - \frac{1}{p}\frac{2n_1\phi}{1-2\tilde{\delta}_0\tilde{\delta}_1\phi}.$$

$$\phi = \frac{1}{n_1} \operatorname{tr} \boldsymbol{\Sigma}_1 \mathbf{Q}_1 \boldsymbol{\Sigma}_1 \mathbf{Q}_1.$$

Discussion

What happens if

- $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| = O(1)$?

What happens if

- $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| = O(1)$?

  The difference in means will not be asymptotically used by RDA.

What happens if

- $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| = O(1)$?

  The difference in means will not be asymptotically used by RDA.

- Matrix $\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1$ has more than $O(\sqrt{p})$ eigenvalues of order 1?

What happens if

- $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| = O(1)$?

  The difference in means will not be asymptotically used by RDA.

- Matrix $\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1$ has more than $O(\sqrt{p})$ eigenvalues of order 1?

  RDA will perform asymptotically perfect classification.

- Matrix $\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1$ has less than $O(\sqrt{p})$ eigenvalues of order 1 and $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| = O(1)$?

RDA Discussions

What happens if

- $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| = O(1)$?

  The difference in means will not be asymptotically used by RDA.

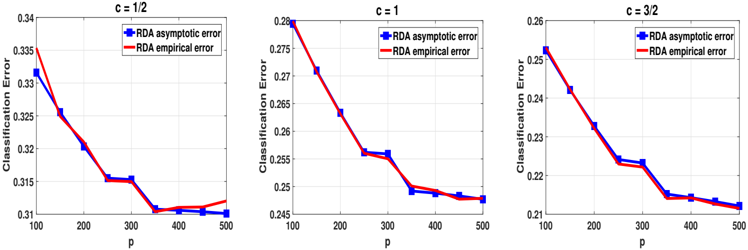
- Matrix $\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1$ has more than $O(\sqrt{p})$ eigenvalues of order 1?

  RDA will perform asymptotically perfect classification.

- Matrix $\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1$ has less than $O(\sqrt{p})$ eigenvalues of order 1 and $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| = O(1)$?

  The misclassification rate of RDA will converge to 0.5. Classification is asymptotically impossible.

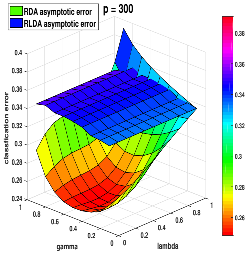
$$\{\boldsymbol{\Sigma}_0\}_{i,j} = 0.6^{|i-j|},\ \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_0 + 2\begin{bmatrix} \mathbf{I}_k & \mathbf{0}_{k\times(p-k)} \\ \mathbf{0}_{(p-k)\times k} & \mathbf{0}_{(p-k)\times(p-k)} \end{bmatrix},\ k = \lfloor\sqrt{p}\rfloor,\ \boldsymbol{\mu}_0 = \mathbf{1}_{p\times 1},$$
$$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_0 + 2p^{-\frac{1}{4}}\mathbf{1}_{p\times 1},\ \lambda = 0.5,\ \gamma = 0.5,\ c = \frac{n_i}{p}$$



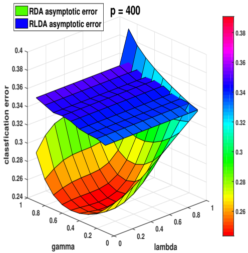
(a) (b) (c)

Figure 3.1: RDA classifier performance in terms of classification error with equal training, $n_0 = n_1$. The x axis is the number of the data dimension.

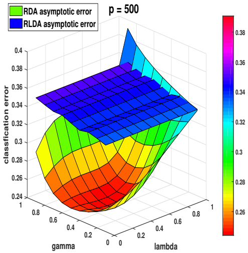
RDA: Synthetic data Experiment (2)



(a)



(b)



(c)

RDA: Synthetic data

- λ approaching the minimum classification error is neither 1 nor 0

The optimal classifier minimizing the classification error is neither neither R-LDA nor R-QDA.

- RDA offers better classification performance than R-LDA and R-QDA with proper regularizers selection

Conclusions

Conclusion

- Leveraging results from RMT, we identified the growth rate regime in which RDA results in a non-trivial classification risk.
- We derived a closed-form expression for the asymptotic classification risk reflecting the impact of the data statistics on the performance.
- Practical insights are drawn to help design the optimal classifier.

Thank you!