High Dimensional Binary Classification
with Rare and Weak Signals

Arman Jahangiri

Department of Mathematical Sciences
University of Calgary

May 24, 2024
2024 Ottawa Mathematics and Statistics Conference

Table of Contents

Outline

Introduction

Basic setup

- **Classification:** predict a discrete label Y from features X
- $(X, Y) \in \mathbb{R}^p \times \{1, \ldots, K\}$
- Data: $\{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^p$
- Goal: learn a classifier $C : \mathbb{R}^p \to \{1, \ldots, K\}$

Performance metric

$$MR(C) = \mathbb{P}\big(C(X) \neq Y\big)$$

Why high-dimensional classification?

- **Genomics:** identify disease-related mutations from DNA measurements
- **Cancer subtyping:** classify tumors using gene expression profiles
- **Spam detection:** classify emails using large collections of word-based features

Common feature

In all three examples, the number of features p can be very large relative to the sample size n.

How can we build a classifier?

1. **Directly minimize classification error** over a class of candidate rules
2. **Model the posterior:**

$$C(x) = \arg\max_{1 \le i \le K} \mathbb{P}(Y = i \mid X = x)$$

3. **Model the class-conditional densities and use Bayes' rule:**

$$\mathbb{P}(Y = i \mid X = x) = \frac{\mathbb{P}(X = x \mid Y = i)\mathbb{P}(Y = i)}{\sum_{j=1}^{K} \mathbb{P}(X = x \mid Y = j)\mathbb{P}(Y = j)}$$

Outline

Logistic Regression

- A generalized linear model widely used for binary classification
- It models the posterior probability:

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta^\top x$$

where $p(x) = \mathbb{P}(Y = 1 \mid X = x)$

- A classification rule is obtained by thresholding the estimated probability
- **Advantage:** fewer parameters than QDA
- **Advantage:** does not require Gaussian class-conditional assumptions
- Typically estimated by maximum likelihood, often via Newton–Raphson / IRLS
- Related work: Abramovich and Grinshtein (2018), *Sparse Logistic Regression*

Discriminant analysis: model assumptions

- For binary classification, assume class-conditional Gaussian distributions:

$$X_i \mid \{Y_i = 0\} \sim \mathcal{N}_p(\mu_0, \Sigma_0), \qquad X_i \mid \{Y_i = 1\} \sim \mathcal{N}_p(\mu_1, \Sigma_1)$$

- Gaussian density:

$$f(x) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu) \right)$$

Discriminant analysis: simplification

- By recentering, we may write

$$\mu_0 = -\mu, \qquad \mu_1 = \mu$$

Note that

$$\mathbb{E}[X_i + \frac{\mu_0 - \mu_1}{2} | \{Y_i = 1\}] = \frac{\mu_0 + \mu_1}{2} = \mu$$
$$\mathbb{E}[-X_i + \frac{\mu_0 - \mu_1}{2} | \{Y_i = 0\}] = -\frac{\mu_0 + \mu_1}{2} = -\mu$$

- After whitening the first class covariance, we reduce to

$$X_i \mid \{Y_i = 0\} \sim \mathcal{N}_p(-\mu, I), \qquad X_i \mid \{Y_i = 1\} \sim \mathcal{N}_p(\mu, \Omega)$$

- This normalization simplifies the asymptotic analysis without changing the core classification problem

Quadratic Discriminant Analysis (QDA)

QDA's Decision rule

$$QDA(x) = \mathbb{I}\{\delta_1(x) > \delta_0(x)\}$$

where

$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^\top \Sigma_k^{-1}(x-\mu_k) + \log \mathbb{P}(Y=k)$$

- The decision boundary is quadratic in x
- QDA allows the two classes to have different covariance matrices

Linear Discriminant Analysis (LDA)

- **Additional assumption:** $\Sigma_0 = \Sigma_1 = \Sigma$
- Under homoscedasticity, the quadratic terms cancel

LDA's Decision rule

$$LDA(x) = \mathbb{I}\{x^\top w + s > 0\}$$

with

$$w = \Sigma^{-1}(\mu_1 - \mu_0), \qquad s = -\frac{1}{2}\left(\mu_1^\top \Sigma^{-1}\mu_1 - \mu_0^\top \Sigma^{-1}\mu_0\right) + \log\frac{\mathbb{P}(Y=1)}{\mathbb{P}(Y=0)}$$

Curse of Dimensionality

Zollanvari, James, Sameni (2020): With a fixed number of training samples, the average (expected) predictive power of a classifier or regressor first increases as the number of dimensions or features used is increased but beyond a certain dimensionality it starts deteriorating instead of improving steadily.

Assumption: $p >> n$

1. $\mu \in \mathbb{R}^p$ $(p >> n)$
2. μ is a noisy vector
3. High sparsity: non-zero entries << zero entries
4. Asymptotically Rare and Weak (ARW) signals
5. sparsity/weakness $\uparrow$ $\Rightarrow$ Classification accuracy $\downarrow$

Outline

Modeling sparsity and weakness in the mean

Mean vector model

For $\mu = (\mu_1, \ldots, \mu_p)$, assume

$$\mu_1, \ldots, \mu_p \overset{\text{iid}}{\sim} (1-\epsilon)M_0 + \epsilon M_\tau, \qquad \tau > 0$$

with

$$\epsilon = p^{-\zeta}, \qquad \tau = g_p\, p^{-\theta}, \qquad \zeta, \theta \in (0,1)$$

- ϵ controls **sparsity**
- τ controls **signal strength**
- M_a denotes a point mass at a
- g_p is either a positive constant sequence or a logarithmic factor

Modeling the covariance structure

Covariance model

$$\Omega = I + V + \xi W, \qquad \xi = p^{-\alpha}, \qquad \alpha \in (0, 1)$$

- V controls sparse diagonal perturbations
- W controls sparse off-diagonal dependence
- ξ scales the off-diagonal contribution

Sparse diagonal and off-diagonal effects

Diagonal entries:

$$v_1, \ldots, v_p \overset{\text{iid}}{\sim} (1-\omega)M_0 + \frac{\omega}{2}M_\gamma + \frac{\omega}{2}M_{-\gamma}$$

with

$$\omega = p^{-\delta}, \qquad \gamma = h_p\, p^{-u}, \qquad \delta > 0,\ u \in (0,1)$$

Off-diagonal entries:

$$w_{ij} = w_{ji} \overset{\text{iid}}{\sim} (1-\nu)M_0 + \frac{\nu}{2}M_1 + \frac{\nu}{2}M_{-1}, \qquad w_{ii} = 0$$

with

$$\nu = p^{-\beta}, \qquad \beta \in (0,2)$$

Outline

Previous research

Chen (2019)

Under assumptions (9) and (10), when $\mu = 0$ and $\Omega = I + V$,

$$\lim_{p \to \infty} MR(QDA) = 0 \quad \text{in the region } \{2u + \delta \leq 1\}.$$

Wang, Wu, and Yao

Under assumptions (7), (8), (10), and (11), when $\mu \neq 0$ and $\Omega = cI_p + \eta W$, the impossibility region is

$$\{\beta + 2\nu > 2\} \cap \{\zeta + 2\theta > 1\},$$

and QDA is optimal.

Main results

Jahangiri (2024)

Under assumptions (10) and (11), when $\mu = 0$ and $\Omega = I + V$, QDA is optimal in the region

$$\{2u + \delta \leq 1\}.$$

Jahangiri (2024)

Under assumptions (7), (8), (10), and (11), when $\mu \neq 0$ and $\Omega = I + V$,

$$\{2u + \delta > 1\} \cap \{\zeta + 2\theta > 1\}$$

is a subset of the impossibility region.

Outline

Computational comparison

Number of parameters

$$\text{QDA: } \Sigma_0, \Sigma_1, \mu_0, \mu_1, \mathbb{P}(Y=0), \mathbb{P}(Y=1) \quad \Rightarrow \quad O(p^2)$$

$$\text{Logistic regression: } \beta_0, \beta_1, \ldots, \beta_p \quad \Rightarrow \quad O(p)$$

Takeaway

QDA is more flexible, but this flexibility comes with a much heavier estimation burden in high dimensions.

Do not use LDA/QDA if...

- Press, Wilson (1978)
- Halperin, Blackwelder, and Verter (1971)
- Hastie, Wittenc, Ersbolld (2013)
- ...

Common Result:
The "use of the maximum likelihood method would be preferable, whenever practical, in situations where the normality assumptions are violated, especially when many of the independent variables are qualitative."

Related Works

- Clemmensen et al. (2011), Sparse discriminant analysis based on the optimal scoring interpretation of linear discriminant analysis
- Mai et al. (2012), Discriminant analysis constructed via lasso penalized least squares
- Pang et al. (2010), Shrinkage-based and regularization diagonal discriminant methods
- Ghosh et al. (2021), Robust generalised quadratic discriminant analysis

References

[1] BICKEL, P. J. and LEVINA, E. (2004). Some theory of Fisher's linear discriminant function, "naive Bayes", and some alternatives when there are many more variables than observations. Bernoulli 10 989–1010. MR2108040

[2] CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. Ann. Statist. 35 2313–2351. MR2382644

[3] DONOHO, D. and JIN, J. (2009). Feature selection by higher criticism thresholding achieves the optimal phase diagram. Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. 367 4449–4470. MR2546396

[4] S. James Press, Sandra Wilson (1978) "Choosing Between Logistic Regression and Discriminant Analysis", Journal of the American Statistical Association, Vol. 73, No. 364 (Dec., 1978), pp. 699-705.

References

[5] CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. Ann. Statist. 35 2313–2351. MR2382644

[6] DONOHO, D. and JIN, J. (2008). Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. Proc. Natl. Acad. Sci. USA 105

[7] HALL, P., PITTELKOW, Y. and GHOSH, M. (2008). Theoretical measures of relative performance of classifiers for high dimensional data with small sample sizes. J. R. Stat. Soc. Ser. B Stat. Methodol. 70 159–173. MR2412636

[8] Zollanvari, A.; James, A. P.; Sameni, R. (2020). "A Theoretical Analysis of the Peaking Phenomenon in Classification". Journal of Classification. 37 (2): 421–434. doi:10.1007/s00357-019-09327-3. S2CID 253851666.

Thank you!