Imprecise Gaussian Discriminant Classification

heudiasyc

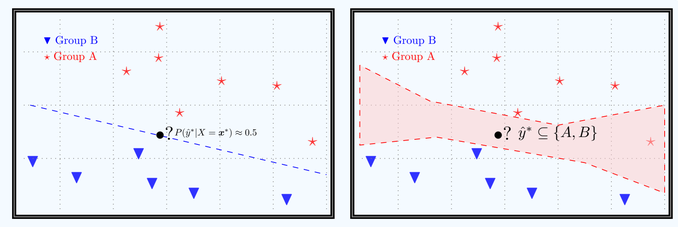
YC. Carranza-Alarcon, Sebastien Destercke

{yonatan-carlos.carranza-alarcon, sebastien.destercke}@hds.utc.fr

ALLIANCE SORBONNE UNIVERSITÉ | utc Recherche

Problem statement

- **Setting:** Training data $D = \{x_i, y_i\}_{i=0}^{N} \subseteq \mathscr{X} \times \mathscr{K}$ where $\mathscr{X} = \mathbb{R}^p$ and $\mathcal{K} = \{ m_1, ..., m_K\}$
- **Motivation:** avoids mistakes performed by the precise models in hard-to-predict unlabeled instances by making cautious decisions (Figure 1(a) and 1(b))
- **Our proposal:**
  - **Cautious decision:** assigns to a new instance $\boldsymbol{x}$ a set-valued predictions $\widehat{Y} \subseteq \mathcal{K}$ in cases of high uncertainty.
  - **New classifier:** a extension of Gaussian Discriminant analysis aiming to quantify the lack of evidence of the component $\mathbb{P}_{X|Y=m_k}$.



(a) **Precise decision-making** (b) **Cautious decision-making**

(1) Classification problem

Objective

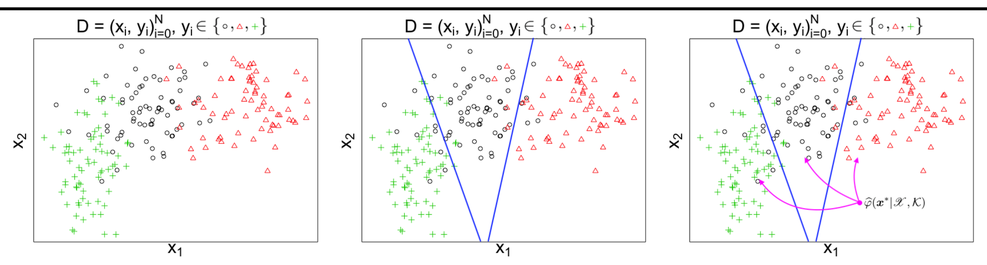
Given a training data

$$D = \{x_i, y_i\}_{i=0}^{N}$$

Learning a classification rule:

$$\varphi : \mathcal{X} \to \mathcal{Y}$$

for predicting new observations $\widehat{\varphi}(\boldsymbol{x})$



(c) Getting Training Data (d) Learning model φ (e) Prediction unlabeled instance $\boldsymbol{x}$

(2) Decision Making

Precise Decision

Definition 2 (Precise ordering [2, pp. 47]) *Given a general loss function $\mathcal{L}(\cdot,\cdot)$, a conditional probability distribution $\mathbb{P}_{Y|\boldsymbol{x}}$ and a new unlabeled instance $\boldsymbol{x}$, m_a is preferred to m_b, denoted by*

$$m_a \succ m_b \iff \mathbb{E}_{\mathbb{P}_{Y|\boldsymbol{x}}}[\mathcal{L}(\cdot, m_a)] < \mathbb{E}_{\mathbb{P}_{Y|\boldsymbol{x}}}[\mathcal{L}(\cdot, m_b)]$$

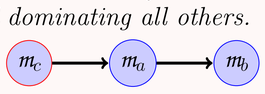
if $\mathcal{L}(\cdot,\cdot)$ is 0/1 loss function, then

$$m_a \succ m_b \Longleftrightarrow P(Y = m_a|X = \boldsymbol{x}) > P(Y = m_b|X = \boldsymbol{x})$$

Example 1 *Given a set of labels $\mathcal{K} = \{m_a, m_b, m_c\}$, a new unlabeled instance $\boldsymbol{x}$, and the probability estimates of the conditional distribution $\widehat{\mathbb{P}}_{Y|X}$:*

$$\widehat{P}(Y = m_a|X = \boldsymbol{x}) = 0.3,$$
$$\widehat{P}(Y = m_b|X = \boldsymbol{x}) = 0.1,$$
$$\widehat{P}(Y = m_c|X = \boldsymbol{x}) = 0.6,$$

the complete preorder over labels w.r.t. estimated probabilities is $m_c \succ m_a \succ m_b$ where m_c is the maximal predicted label dominating all others.



Cautious Decision

Definition 3 (Partial Ordering by Maximality Criterion [4, §3.2]) *Let $\mathcal{L}(\cdot,\cdot)$ be a general loss function, $\boldsymbol{x}$ an observed instance and $\mathscr{P}_{Y|\boldsymbol{x}}$ a set of conditional probability distributions. m_a is preferred to m_b according to the maximility criterion if the cost of exchanging m_a with m_b has a positive lower expectation:*

$$m_a \succ_M m_b \iff \inf_{\mathbb{P}_{Y|\boldsymbol{x}} \in \mathscr{P}_{Y|\boldsymbol{x}}} \mathbb{E}_{\mathbb{P}_{Y|\boldsymbol{x}}}\big[\mathcal{L}(\cdot, m_b) - \mathcal{L}(\cdot, m_a)\big] > 0 \tag{6}$$

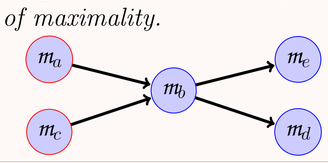
if $\mathcal{L}(\cdot,\cdot)$ is 0/1 loss function, $m_a \succ_M m_b$ if and only if:

$$\inf_{\mathbb{P}_{Y|\boldsymbol{x}} \in \mathscr{P}_{Y|\boldsymbol{x}}} P(Y = m_a|\boldsymbol{x}) - P(Y = m_b|\boldsymbol{x}) > 0 \tag{7}$$

Example 2 *Given a set of labels $\mathcal{K} = \{m_a, m_b, m_c, m_d, m_e\}$ and a possible partial ordering could be the following:*

$$\mathscr{B} = \{m_a \succ_M m_b, \quad m_c \succ_M m_b, \quad m_a \succ\prec_M m_c,$$
$$m_b \succ_M m_d, \quad m_b \succ_M m_d, \quad m_d \succ\prec_M m_e\}$$

where $\widehat{Y}_M = \{m_a, m_c\}$ is the predicted set obtained from the set $\mathscr{B}$ of comparisons derived by the criterion of maximality.



(3) Near-Ignorance on Gaussian Discriminant Analysis

Definition 1 (Prior near-ignorance for k-parameter exponential families [1, §4, eq. 16]) *Let $\mathbb{L}$ be a bounded closed convex subset of $\mathbb{R}^k$ strictly including the origin ([1, lem. 4.5]).*

$$\mathbb{L} = \Big\{\ell \in \mathbb{R}^k : \ell_i \in [-c_i, c_i], c_i > 0, i = \{1, \ldots, d\}\Big\} \tag{1}$$

The following set of priors

$$\mathscr{M}^w = \Big\{w \in \mathscr{W} \mid p(w) \propto \exp(\ell^T w), \ell = [\ell_1, \ldots, \ell_k]^T \in \mathbb{L}\Big\}, \tag{2}$$

satisfies following properties: (P1) the prior-invariance, (P2) the prior-ignorance, (P3) learning from data and (P4) convergence, as well as conjugacy between the likelihood and the set of posterior distributions.

Assumptions

- **(A1)** Normality of conditional probability distribution $\mathbb{P}_{X|Y=m_k} := \mathscr{G}_{m_k}$.
- **(A2)** A *precise* estimation of marginal distribution $\mathbb{P}_Y := \widehat{\pi}_y$
- **(A3)** A *precise* estimation of covariance matrix $\Sigma_{m_k} := \widehat{\Sigma}_{m_k} = \widehat{S}_{m_k}$ or $\widehat{S}$.

Assuming (A1) and (A3), and using the set of prior distribution of Equation (2), we can get a set of posterior distribution [3, §5.2]:

$$\mathscr{M}_n^{\mu_{m_k}} = \left\{\mu_{m_k} \;\middle|\; \overline{\boldsymbol{x}}_{n_{m_k}}, \ell \propto \mathcal{N}\left(\frac{\ell + n_{m_k}\overline{\boldsymbol{x}}_{n_{m_k}}}{n_{m_k}}, \frac{1}{n}\widehat{\Sigma}_{m_k}\right)\right\} \tag{3}$$

Using the lower and upper posterior expectations of μ_{m_k} on $\mathscr{M}_n^{\mu_{m_k}}$, we get a convex space of plausible values for the mean μ_{m_k}:

$$\mathbb{G}_{m_k} = \left\{\widehat{\mu}_{m_k} \in \mathbb{R}^d \;\middle|\; \widehat{\mu}_{i,m_k} \in \left[\frac{-c_i + n_{m_k}\overline{\boldsymbol{x}}_{i,n_{m_k}}}{n_{m_k}}, \frac{c_i + n_{m_k}\overline{\boldsymbol{x}}_{i,n_{m_k}}}{n_{m_k}}\right] \right.$$
$$\left. \forall i = \{1, ..., d\},\ c_i > 0 \right\} \tag{4}$$

On the basis of the set $\mathbb{G}_{m_k}$, we can simply consider the following set of conditional probability distributions $\mathscr{P}_{X|Y=m_k}$:

$$\mathscr{P}_{X|Y=m_k} = \left\{\mathbb{P}_{X|Y=m_k} \;\middle|\; \mathbb{P}_{X|Y=m_k} \sim \mathcal{N}(\mu_{m_k}, \widehat{\Sigma}_{m_k}), \mu_{m_k} \in \mathbb{G}_{m_k}\right\} \tag{5}$$

(4) Gaussian Discriminant Classification(GDC)

Precise GDC

Applying Baye's rules to $P(Y = m_a|X = \boldsymbol{x})$:

$$P(y = m_k|X = \boldsymbol{x}) = \frac{P(X = \boldsymbol{x}|y = m_k)P(y = m_k)}{\sum_{m_l \in \mathcal{K}} P(X = \boldsymbol{x}|y = m_l)P(y = m_l)}$$

Assuming normality on $P_{X|Y=m_k}$:

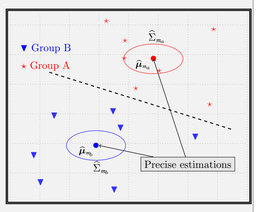
$$\mathscr{G}_{m_k} := \mathbb{P}_{X|Y=m_k} \sim \mathcal{N}(\mu_{m_k}, \Sigma_{m_k}) \tag{8}$$

Defining the marginal distribution as $\pi_{m_k} := P(Y = m_k)$, so under 0/1 loss function, the optimal prediction becomes:

$$\arg\max_{m_k \in \mathcal{K}} \ \log \pi_{m_k} - \log |\Sigma_{m_k}|^{\frac{1}{2}} - \frac{1}{2}(\boldsymbol{x}^T - \mu_{m_k})^T \Sigma_{m_k}^{-1}(\boldsymbol{x}^T - \mu_{m_k}) \tag{9}$$

Estimating parameters by MLE on a subset $\mathscr{D}_{m_k} = \{(x_{i,k}, y_{i,k=m_k}) | i = 1, \ldots, n_{m_k}\} \subseteq \mathscr{D}$:

- ☞ $\hat{\pi}_{m_k} = n_{m_k}/N$ (frequency of m_k)
- ☞ $\hat{\mu}_{m_k} = \overline{\boldsymbol{x}}_{m_k}$ (sample mean of $\mathscr{D}_{m_k}$)
- ☞ If we assume:
  - ① (heteroscedasticity) $\rightarrow$ $\hat{\Sigma}_{m_k} = \widehat{S}_{m_k}$ (sample covariance matrix of $\mathscr{D}_{m_k}$)
  - ② (homoscedasticity) $\rightarrow$ $\hat{\Sigma}_{m_k} = \widehat{S}$ (within-class covariance matrix $\mathscr{D}$)



Imprecise Gaussian Discriminant Classification(IGDC)

- Using the maximality criterion and applying Baye's rule, to know whether $m_a \succ_M m_b$, we need to solve

$$\inf_{\mathbb{P}_Y \in \mathscr{P}_Y} \inf_{\mathbb{P}_{X|m_a} \in \mathscr{P}_{X|m_a}} \inf_{\mathbb{P}_{X|m_b} \in \mathscr{P}_{X|m_b}} [P(X = \boldsymbol{x}|Y = m_a)P(Y = m_a) - P(X = \boldsymbol{x}|Y = m_b)P(y = m_b)] > 0 \tag{10}$$

- Assuming (A2), an precise estimation for marginal: $P(Y = m_k) := \widehat{\pi}_{m_k} > 0$:

$$\inf_{\mathbb{P}_{X|m_a} \in \mathscr{P}_{X|m_a}} \inf_{\mathbb{P}_{X|m_b} \in \mathscr{P}_{X|m_b}} [P(X = \boldsymbol{x}|Y = m_a)\widehat{\pi}_{m_a} - P(X = \boldsymbol{x}|Y = m_b)\widehat{\pi}_{m_b}] > 0 \tag{11}$$

- As conditional distributions sets $\mathscr{P}_{X|y=m_k}$ are independent of each others, then

$$\widehat{\pi}_{m_a} \inf_{\mathbb{P}_{X|m_a} \in \mathscr{P}_{X|m_a}} P(X = \boldsymbol{x}|Y = m_a) \ - \ \widehat{\pi}_{m_b} \sup_{\mathbb{P}_{X|m_b} \in \mathscr{P}_{X|m_b}} P(X = \boldsymbol{x}|Y = m_b) > 0 \tag{12}$$

- We then have two optimization problems with constraint convex space:

$$\sup_{\mathbb{P}_{X|m_b} \in \mathscr{P}_{X|m_b}} P(\boldsymbol{x}|Y = m_b) \iff \overline{\mu}_{m_b} = \arg\max_{\mu_{m_b} \in \mathbb{G}_{m_b}} -\frac{1}{2}(\boldsymbol{x} - \mu_{m_b})^T \widehat{\Sigma}_{m_b}^{-1}(\boldsymbol{x} - \mu_{m_b}) \tag{BQP}$$

$$\inf_{\mathbb{P}_{X|m_a} \in \mathscr{P}_{X|m_a}} P(\boldsymbol{x}|Y = m_a) \iff \underline{\mu}_{m_a} = \arg\min_{\mu_{m_a} \in \mathbb{G}_{m_a}} -\frac{1}{2}(\boldsymbol{x} - \mu_{m_a})^T \widehat{\Sigma}_{m_a}^{-1}(\boldsymbol{x} - \mu_{m_a}) \tag{NBQP}$$

☞ First problem box-constrained quadratic problem (BQP).
☞ Second problem non-convex BQP $\rightarrow$ solved through Branch and Bound method.

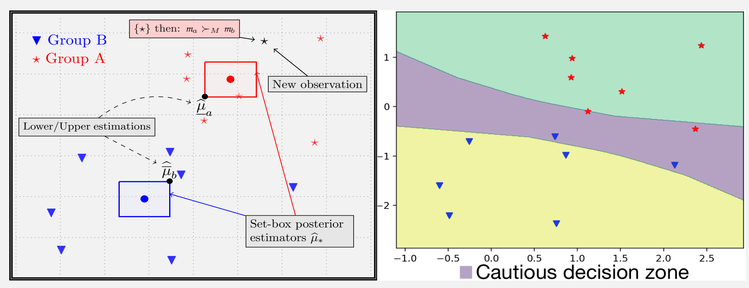


Figure 1: Cautious decision zone for a binary classification

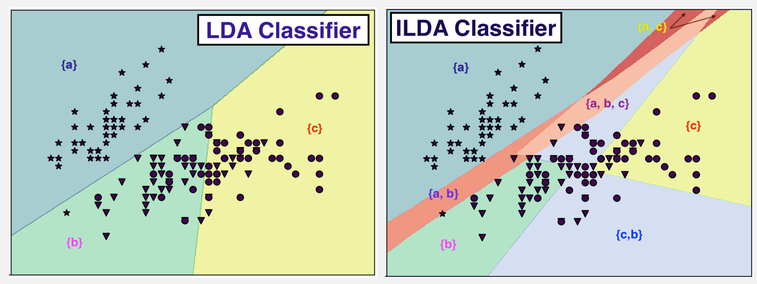


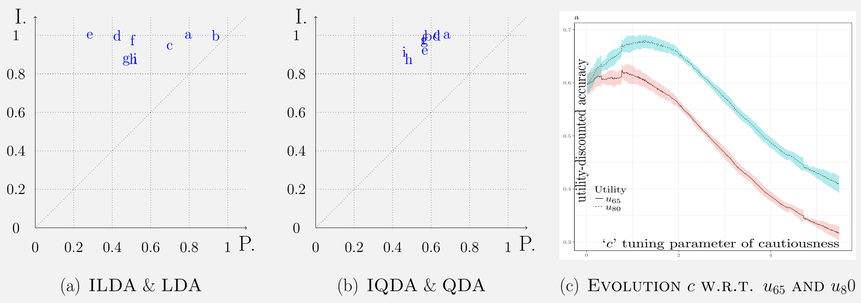
Figure 2: Cautious decision zone with three class $\{a, b, c\}$

(5) Experimental results

- ☞ 9 data sets issued from UCI repository.
- ☞ 10×10-fold cross-validation procedure.
- ☞ Utility-discounted accuracy measure proposed to *Zaffalon et al* on [5].

$$u(y, \widehat{Y}_M) = \begin{cases} 0 & \text{if } y \notin \widehat{Y}_M \\ \frac{\alpha}{|\widehat{Y}_M|} - \frac{1-\alpha}{|\widehat{Y}_M|^2} & \text{else} \end{cases} \Longrightarrow \begin{array}{l}\textbf{Goal:} \\ \text{Reward cautiousness} \\ \text{to some degree}\end{array} \begin{cases} \alpha = 1: & \text{cautiousness = randonness} \\ \alpha \to \infty: & \text{best classifier vacuous} \end{cases}$$

where the usual measures $u_{65}(\cdot,\cdot)$ with $\alpha = 1.6$ and $u_{80}(\cdot,\cdot)$ with $\alpha = 2.2$ have been used in this work.



(a) ILDA & LDA (b) IQDA & QDA (c) Evolution c w.r.t. u_{65} and u_{80}

Figure 3: (a) Correctness of the **I**mprecise LDA in the case of abstention versus accuracy of the **P**recise LDA. (b) Correctness of the **I**mpreciseQLDA in the case of abstention versus accuracy of the **P**recise QDA (graphs are given for the u_{80} accuracies), and the Figure (c) Prediction performance of ILDA model w.r.t. utility-discount accuracy and c tuning parameter on vowel dataset.

#	name	# instances	# features	# labels
a	iris	150	4	3
b	wine	178	13	3
c	forest	198	27	4
d	seeds	210	7	3
e	dermatology	385	34	6
f	vehicle	846	18	4
g	vowel	990	10	11
h	wine-quality	1599	11	6
i	wall-following	5456	24	4

Table 1: Data sets used in the experiments

	LDA	**ILDA**		**QDA**	**IQDA**		**Avg.**
#	acc.	u_{80}	u_{65}	acc	u_{80}	u_{65}	**time (sec.)**
a	97.96	**98.38**	97.16	97.29	**98.08**	97.13	0.56
b	98.85	**98.99**	98.95	99.03	**99.39**	99.09	1.49
c	94.61	**94.56**	94.05	89.43	**91.77**	88.90	12.14
d	96.35	**96.59**	96.51	94.64	**95.20**	94.72	1.50
e	96.58	**97.06**	96.94	82.47	**84.24**	84.05	19.24
f	77.96	**81.98**	79.59	85.07	**87.96**	86.13	3.10
g	60.10	**67.45**	62.41	87.83	**89.96**	88.40	4.95
h	59.25	**65.83**	60.31	55.62	**65.85**	60.36	34.85
i	67.96	**71.34**	66.65	65.87	**71.79**	69.75	10.77
avg.	83.68	**86.05**	84.03	80.34	**87.16**	85.33	10.1

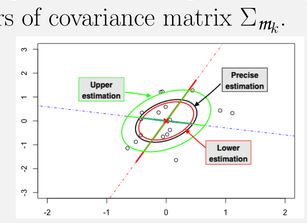
Table 2: Average utility-discounted accuracies (%)

(6)Conclusion and Perspectives

- ☞ Increasing imprecision on the estimators has allowed us to be more cautious and to improve the prediction of classification.
- ☞ Works done since submission of ISIPTA paper:
  - Considering a diagonal structure of the covariance matrix, i.e. $\Sigma_{m_k} = \boldsymbol{\sigma}_{m_k}^T \mathbb{I}$.
  - Considering a set of marginals distribution $\mathscr{P}_Y$ instead of $\mathbb{P}_Y$ (i.e. release Assumption (A2))
  - Considering the use of a generic loss function instead of *zero-one* loss function $\mathscr{L}_{0/1}$.
- ☞ What remains to do:
  - Make imprecise the covariance matrix Σ_{m_k} by using the following set of prior distributions

$$\mathscr{M} \propto \left\{|\Lambda|^{\frac{v_0}{2}} \exp\left\{-\frac{1}{2} tr\left(\Lambda \ell \ell^T\right)\right\}, \ell \in \mathbb{L}, \ell_i \in [-c_i, c_i], v_0 > p\right\}$$

  - Making imprecise the components eigenvalues and eigenvectors of covariance matrix Σ_{m_k}.



Acknowledgments

This work was carried out in the framework of the Labex MS2T, funded by the French Government, through the National Agency for Research (Reference ANR-11-IDEX-0004-02).

References

[1] Alessio Benavoli and Marco Zaffalon. Prior near ignorance for inferences in the k-parameter exponential family. *Statistics*, 49(5):1104–1140, 2014.

[2] James O Berger. *Statistical decision theory and Bayesian analysis; 2nd ed.* Springer Series in Statistics. Springer, New York, 1985.

[3] José M Bernardo and Adrian FM Smith. *Bayesian Theory*. John Wiley & Sons Ltd., 2000.

[4] Matthias CM Troffaes. Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, 45(1):17–29, 2007.

[5] Marco Zaffalon, Giorgio Corani, and Denis Mauá. Evaluating credal classifiers by utility-discounted predictive accuracy. *International Journal of Approximate Reasoning*, 53(8):1282–1301, 2012.