University of Amsterdam
Informatics Institute

AMLAB Amsterdam Machine Learning Lab

Machine Learning 1

Lecture 6.1 - Supervised Learning Classification - Probabilistic Generative Models - Maximum Likelihood Solution

Erik Bekkers

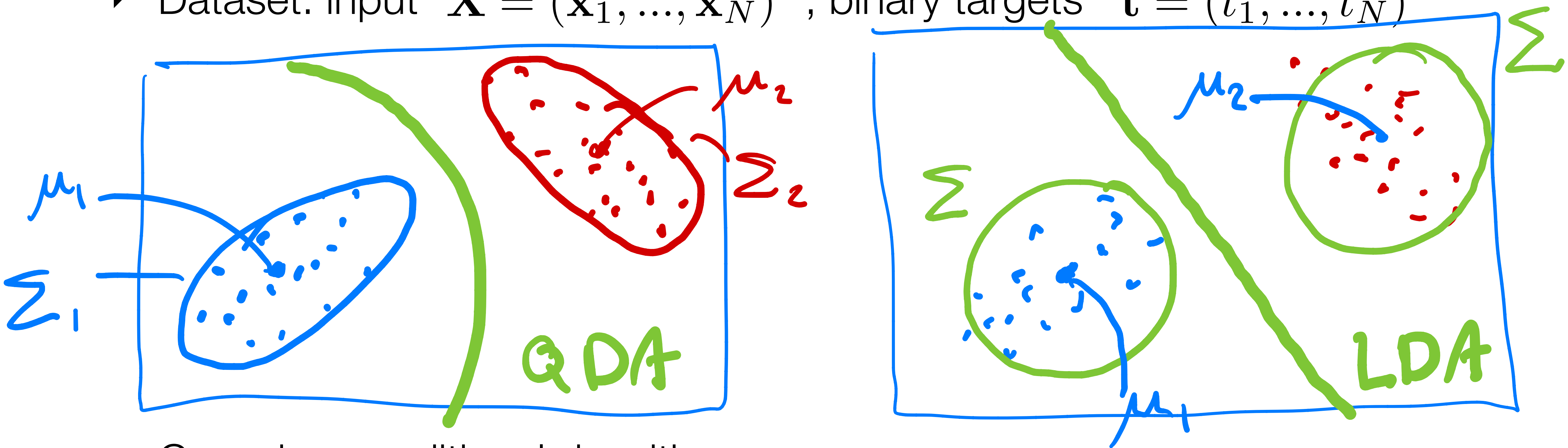
(Bishop 4.2.2)

Slide credits: Patrick Forré and Rianne van den Berg

Image credit: Kirillm | Getty Images

LDA: Maximum Likelihood for K=2

- Dataset: input $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_N)^T$, binary targets $\mathbf{t} = (t_1, ..., t_N)^T$



- Gaussian conditional densities

$$p(\mathbf{x}|C_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\mathbf{\Sigma}|^{1/2}} \exp\{\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_k)\}$$

- Use Maximum likelihood to estimate $\boldsymbol{\mu}_k$, $\mathbf{\Sigma}$ and priors $p(C_k)$
- Denote $p(C_1) = q$ and $p(C_2) = 1 - q$
- For $\mathbf{x}_n$ with $t_n = 1$: $p(\mathbf{x}_n, C_1) = p(\mathbf{x}_n|C_1)p(C_1) = q\,\mathcal{N}(\underline{x}_n \,|\, \underline{\mu}_1, \Sigma)$
- For $\mathbf{x}_n$ with $t_n = 0$: $p(\mathbf{x}_n, C_2) = p(\mathbf{x}_n|C_2)p(C_2) = (1-q)\,\mathcal{N}(\underline{x}_n \,|\, \underline{\mu}_2, \Sigma)$

LDA: Maximum Likelihood for K=2

- Gaussian conditional densities

$$p(\mathbf{x}|C_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\{\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_k)\}$$

- Use Maximum likelihood to estimate $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}$ and priors $p(C_k)$
- Denote $p(C_1) = q$ and $p(C_2) = 1 - q$

- Likelihood

$$\begin{aligned} p(\mathbf{t}, \mathbf{X} \,|\, q, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) &= \prod_{n=1}^{N} p(\mathbf{x}_n, t_n) = \prod_{n=1}^{N} p(\mathbf{x}_n \,|\, t_n) p(t_n) \\ &= \prod_{n=1}^{N} \left[p(\mathbf{x}_n \,|\, C_1) p(C_1) \right]^{t_n} \left[p(\mathbf{x}_n \,|\, C_2) p(C_2) \right]^{1-t_n} \\ &= \prod_{n=1}^{N} \left[q \mathcal{N}(\mathbf{x}_n \,|\, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) \right]^{t_n} \left[(1-q) \mathcal{N}(\mathbf{x}_n \,|\, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \right]^{1-t_n} \end{aligned}$$

LDA: Maximum Likelihood for K=2

- Likelihood

$$p(\mathbf{t}, \mathbf{X} | q, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} \left[q\mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_1, \boldsymbol{\Sigma})\right]^{t_n} \left[(1-q)\mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_2, \boldsymbol{\Sigma})\right]^{1-t_n}$$

- Log likelihood

$$\ln p(\mathbf{t}, \mathbf{X} | q, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} t_n \ln q + t_n \ln \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) +$$

$$+ (1-t_n)\ln(1-q) + (1-t_n)\ln \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_2, \boldsymbol{\Sigma})$$

- Estimate for q:

$$\frac{\partial}{\partial q} \ln p(\mathbf{t}, \mathbf{X} | q, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \frac{t_n}{q} + \frac{1-t_n}{1-q} \cdot (-1) = \sum_{n=1}^{N} \frac{t_n(1-q) - (1-t_n)q}{q(1-q)}$$

$$t_n - \cancel{t_n q} - q + \cancel{t_n q}$$

$$= \sum_{n=1}^{N} \frac{t_n - q}{q(1-q)} = 0$$

$(q \neq 0)$ $(q \neq 1)$ $\Longrightarrow \underbrace{\sum_{n=1}^{N} q}_{Nq} = \sum_{n=1}^{N} t_n \Longrightarrow$

$$\boxed{q_{ML} = \frac{1}{N}\sum_{n=1}^{N} t_n = \frac{N_1}{N}}$$

LDA: Maximum Likelihood for K=2

- log likelihood:

$$\ln p(\mathbf{t}, \mathbf{X}|q, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} t_n \ln q + t_n \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) +$$

$$+ (1 - t_n) \ln(1 - q) + (1 - t_n) \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$$

- Estimate for $\boldsymbol{\mu}_1$

$$\frac{\partial}{\partial \boldsymbol{\mu}_1} \ln p(\mathbf{t}, \mathbf{X}|q, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \frac{\partial}{\partial \boldsymbol{\mu}_1} \sum_{n=1}^{N} t_n \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$$

$$= -\frac{1}{2} \frac{\partial}{\partial \boldsymbol{\mu}_1} \sum_{n=1}^{N} t_n (\mathbf{x}_n - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_1) = \sum_{n=1}^{N} t_n (x_n - \mu_1)^T \Sigma^{-1} = 0^T$$

(Σ is symmetric)

(Σ pos. def.)

$$\Rightarrow \sum_{n=1}^{N} t_n \mu_1 = \sum_{n=1}^{N} t_n x_n$$

$$\Rightarrow N_1 \mu_1 = \sum_{n=1}^{N} t_n x_n$$

$$\boxed{\mu_{1,ML} = \frac{1}{N_1} \sum_{n=1}^{N} t_n x_n , \quad \mu_{2,ML} = \frac{1}{N_2} \sum_{n=1}^{N} (1-t_n) x_n}$$

LDA: Maximum Likelihood for K=2

- log likelihood:

$$\ln p(\mathbf{t}, \mathbf{X} | q, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} t_n \ln q + t_n \ln \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) +$$
$$+ (1 - t_n) \ln(1 - q) + (1 - t_n) \ln \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_2, \boldsymbol{\Sigma})$$

- Estimate for $\boldsymbol{\Sigma}$

$$\frac{\partial}{\partial \boldsymbol{\Sigma}} \ln p(\mathbf{t}, \mathbf{X} | q, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \frac{\partial}{\partial \boldsymbol{\Sigma}} \left[-\frac{N}{2} \ln |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{n=1}^{N} t_n (\mathbf{x}_n - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_1) \right.$$
$$\left. -\frac{1}{2} \sum_{n=1}^{N} (1 - t_n)(\mathbf{x}_n - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_2) \right] = 0$$

(Bishop 4.2.2 + notes on Canvas)

- ML solution:

$$\boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{N_1}{N} \left[\overbrace{\frac{1}{N_1} \sum_{n=1}^{N} t_n (\mathbf{x}_n - \boldsymbol{\mu}_{1,\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{1,\mathrm{ML}})^T}^{\Sigma_1} \right]$$
$$+ \frac{N_2}{N} \left[\overbrace{\frac{1}{N_2} \sum_{n=1}^{N} (1 - t_n)(\mathbf{x}_n - \boldsymbol{\mu}_{2,\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{2,\mathrm{ML}})^T}^{\Sigma_2} \right]$$

LDA: Maximum Likelihood for K=2

The ML solutions:

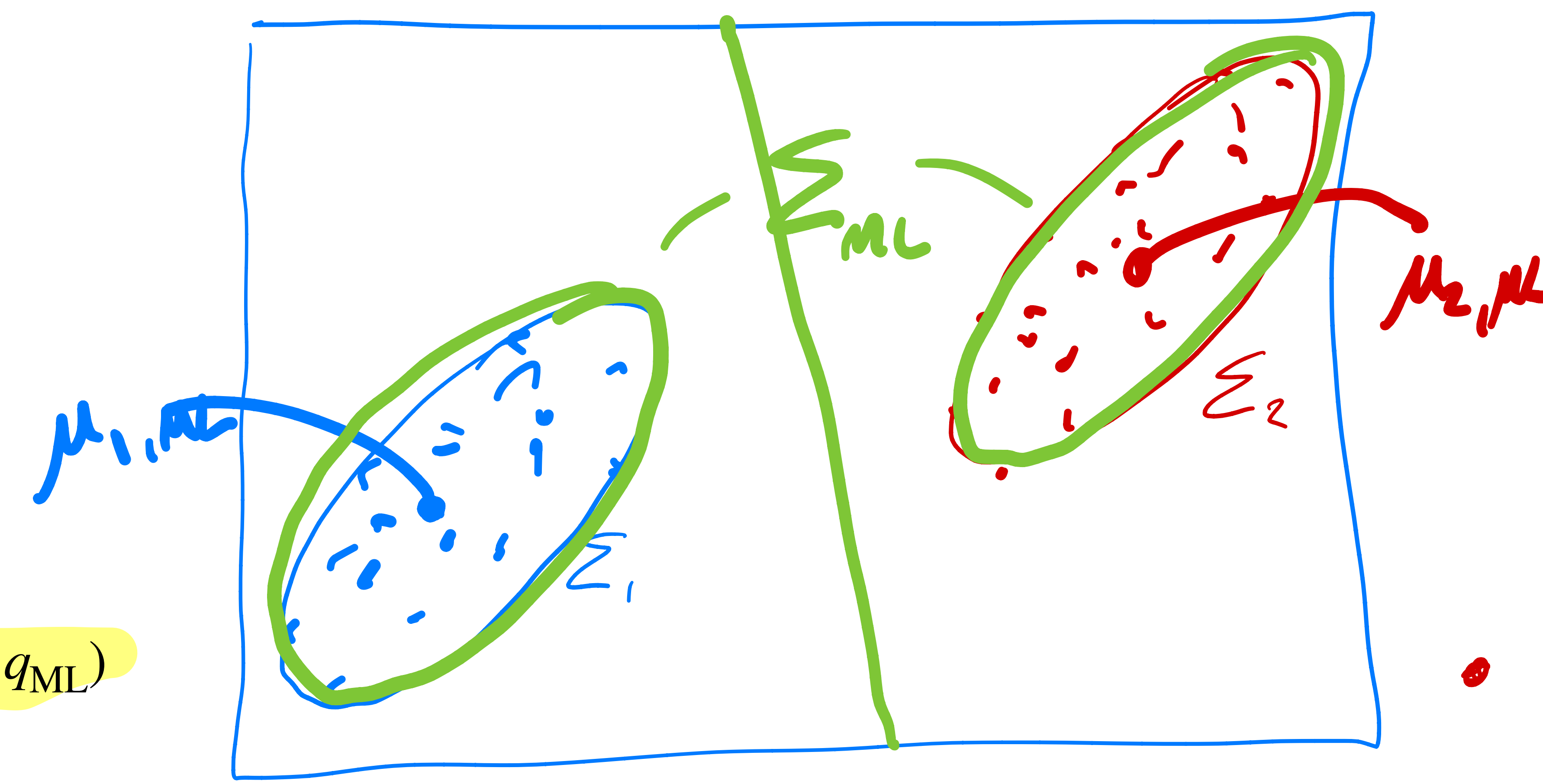
$$\boldsymbol{\mu}_{1,\mathrm{ML}} = \frac{1}{N_1}\sum_{n=1}^{N} t_n \mathbf{x}_n \qquad \boldsymbol{\mu}_{2,\mathrm{ML}} = \frac{1}{N_2}\sum_{n=1}^{N} (1-t_n)\mathbf{x}_n \qquad q_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} t_n = \frac{N_1}{N}$$

$$\Sigma_{\mathrm{ML}} = \frac{N_1}{N}\left[\frac{1}{N_1}\sum_{n=1}^{N} t^n \left(\mathbf{x}_n - \boldsymbol{\mu}_{1,\mathrm{ML}}\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_{1,\mathrm{ML}}\right)^T\right] + \frac{N_2}{N}\left[\frac{1}{N_2}\sum_{n=1}^{N} (1-t^n)\left(\mathbf{x}_n - \boldsymbol{\mu}_{2,\mathrm{ML}}\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_{2,\mathrm{ML}}\right)^T\right]$$

For the joint probabilities:

$$p(\mathbf{x}, C_1) = p(\mathbf{x}\,|\,C_1)p(C_1) = \mathcal{N}\left(\mathbf{x}\,|\,\boldsymbol{\mu}_{1,\mathrm{ML}}, \Sigma_{\mathrm{ML}}\right)\ q_{\mathrm{ML}}$$

$$p(\mathbf{x}, C_2) = p(\mathbf{x}\,|\,C_2)p(C_2) = \mathcal{N}\left(\mathbf{x}\,|\,\boldsymbol{\mu}_{2,\mathrm{ML}}, \Sigma_{\mathrm{ML}}\right)\ (1 - q_{\mathrm{ML}})$$



LDA: prediction for K=2

- For new datapoint **x**':

$$p(C_1|\mathbf{x}') = \sigma(\underline{\mathbf{w}_{\mathrm{ML}}^T \mathbf{x}' + w_{0,\mathrm{ML}}})$$

$$\mathbf{w}_{\mathrm{ML}} = \boldsymbol{\Sigma}_{\mathrm{ML}}^{-1}(\boldsymbol{\mu}_{1,\mathrm{ML}} - \boldsymbol{\mu}_{2,\mathrm{ML}})$$

$$w_{0,\mathrm{ML}} = -\frac{1}{2}\boldsymbol{\mu}_{1,\mathrm{ML}}^T \boldsymbol{\Sigma}_{\mathrm{ML}}^{-1} \boldsymbol{\mu}_{1,\mathrm{ML}} + \frac{1}{2}\boldsymbol{\mu}_{2,\mathrm{ML}}^T \boldsymbol{\Sigma}_{\mathrm{ML}}^{-1} \boldsymbol{\mu}_{2,\mathrm{ML}} + \ln \frac{q_{\mathrm{ML}}}{1 - q_{\mathrm{ML}}}$$

- Assign **x**' to C_1 if $p(C_1|\mathbf{x}') \geq \frac{1}{2}$ $\left(w_{ML}^T x' + w_{0,ML} \geq 0 \right)$

- Disadvantage of LDA:
  - Gaussian distribution is sensitive to outliers
  - Linearity/handcrafted features restrict application
  - Maximum likelihood is prone to overfitting