University of Amsterdam
Informatics Institute

AMLAB
Amsterdam
Machine Learning Lab

Machine Learning 1

Lecture 5.6 - Supervised Learning Classification - Probabilistic Generative Models

Erik Bekkers

(Bishop 1.5)

Slide credits: Patrick Forré and Rianne van den Berg

Image credit: Kirillm | Getty Images

Probabilistic Generative Models: K=2

- Class-conditional densities: $p(x \mid C_k)$
- Prior class probabilities: $p(C_k)$
- Joint distribution: $p(x, C_k) = p(x \mid C_k)\, p(C_k)$
- Posterior distribution: K=2

$$p(C_1|\mathbf{x}) = \frac{p(\mathbf{x}|C_1)p(C_1)}{p(\mathbf{x}|C_1)p(C_1) + p(\mathbf{x}|C_2)p(C_2)} = p(x)$$

$$= \frac{1}{1 + \frac{p(x|C_2)p(C_2)}{p(x|C_1)p(C_1)}} = \frac{1}{1 + e^{-a}}$$

- $a = \ln \frac{\sigma}{1-\sigma} = \ln \frac{p(x|C_1)\,p(C_1)}{p(x|C_2)\,p(C_2)}$

log odds

Logistic Sigmoid Function

"$p(C_1|x) = \sigma(a)$"

$$\sigma(a) = \frac{1}{1 + \exp(-a)}$$

$$\sigma(-a) = 1 - \sigma(a)$$

$$\sigma'(a) = \sigma(a)(1 - \sigma(a))$$

verify

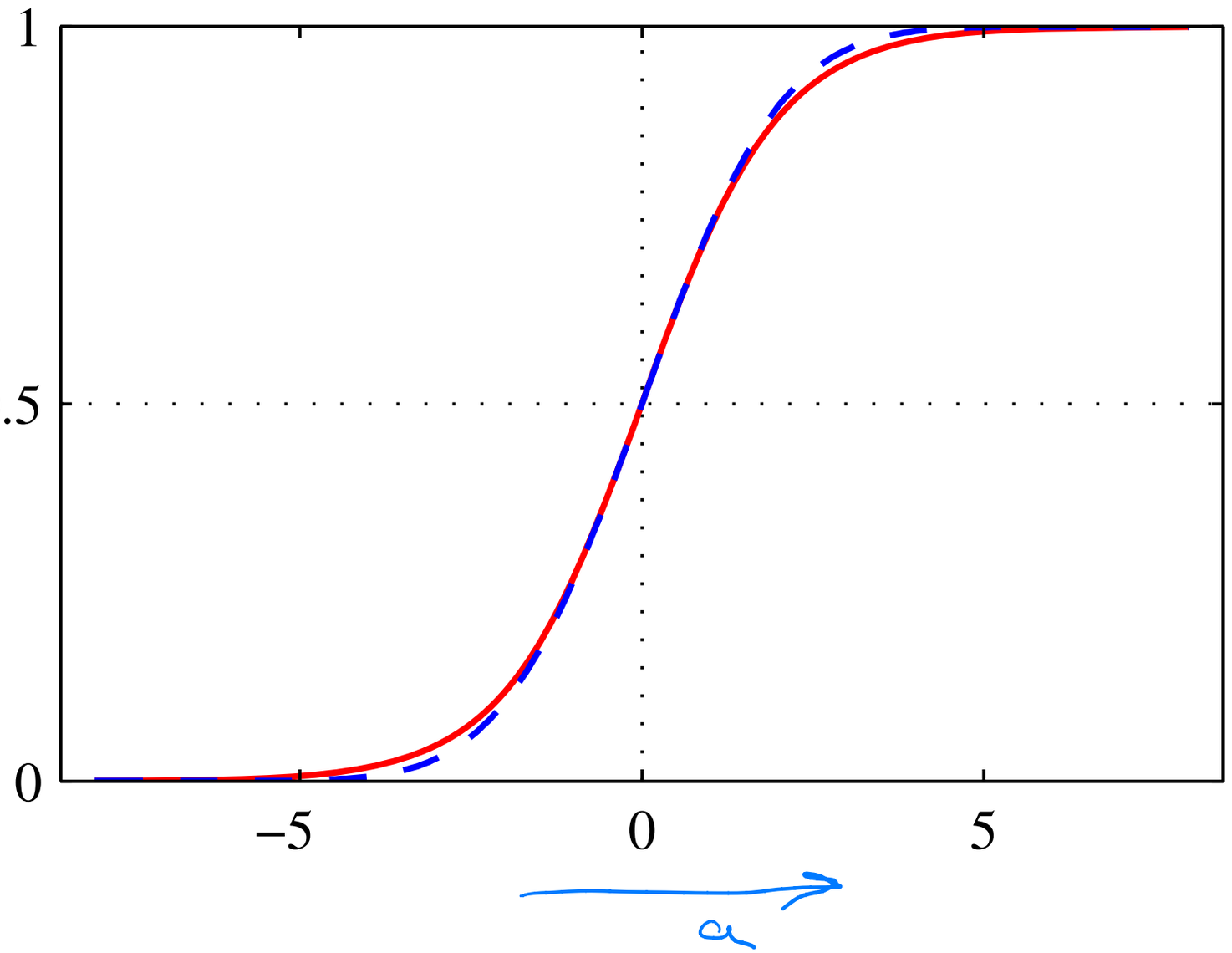

Figure: Logistic Sigmoid function (red) (Bishop 4.9)

Probabilistic Generative Models: general K

- For multiple classes (general K):

$$p(C_k|\mathbf{x}) = \frac{p(\mathbf{x}|C_k)p(C_k)}{\sum_{j=1}^{K} p(\mathbf{x}|C_j)p(C_j)} = \frac{\exp(a_k)}{\sum_{j=1}^{K} \exp(a_j)}$$

$p(C_k) = \pi_k$

- $a_k = \ln(p(\mathbf{x}|C_k)p(C_k))$

- Softmax: if $a_k >> a_j$ for all $j \neq k$: $p(C_k|x) \approx 1$, $p(C_j|x) \approx 0$

- Note: for K=2:

$$p(C_1|\mathbf{x}) = \frac{p(\mathbf{x}|C_1)p(C_1)}{p(\mathbf{x}|C_1)p(C_1) + p(\mathbf{x}|C_2)p(C_2)} = \frac{1}{1 + \frac{p(\mathbf{x}|C_2)p(C_2)}{p(\mathbf{x}|C_1)p(C_1)}}$$

$$= \sigma(a), \quad a = a_1 - a_2$$

$$a = \ln \frac{p(\mathbf{x}|C_1)p(C_1)}{p(\mathbf{x}|C_2)p(C_2)}$$

Class Conditional Densities: Continuous Inputs

- Gaussian Class-conditional densities:

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

For each k

- Assume shared covariance matrix: $\mathbf{\Sigma}_k = \Sigma$

↳ linear discriminant analysis (LDA)

- K=2 classes: $p(C_1|\mathbf{x}) = \frac{1}{1+\exp(-a)} = \sigma(a)$

$$a = \ln \frac{p(\mathbf{x}|C_1)p(C_1)}{p(\mathbf{x}|C_2)p(C_2)} = \ln \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_1, \mathbf{\Sigma}) - \ln \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_2, \mathbf{\Sigma}) + \ln \frac{p(C_1)}{p(C_2)}$$

$$= -\frac{1}{2}\ln|\mathbf{\Sigma}| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \frac{1}{2}\ln|\mathbf{\Sigma}| + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) + \ln\frac{p(C_1)}{p(C_2)}$$

$$= (\mu_1-\mu_2)^T\Sigma^{-1}x - \tfrac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \tfrac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 + \ln\frac{p(C_1)}{p(C_2)}$$

- Generalized Linear Model: $p(C_1|\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x} + w_0)$

$$\mathbf{w} = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

$$w_0 = -\frac{1}{2}\boldsymbol{\mu}_1^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln\frac{p(C_1)}{p(C_2)}$$

Decision Boundary

$a_1 = a_2$ $(a = 0)$

$(\sigma(a) = \frac{1}{2})$

Example: Linear Discriminant Analysis for K=2

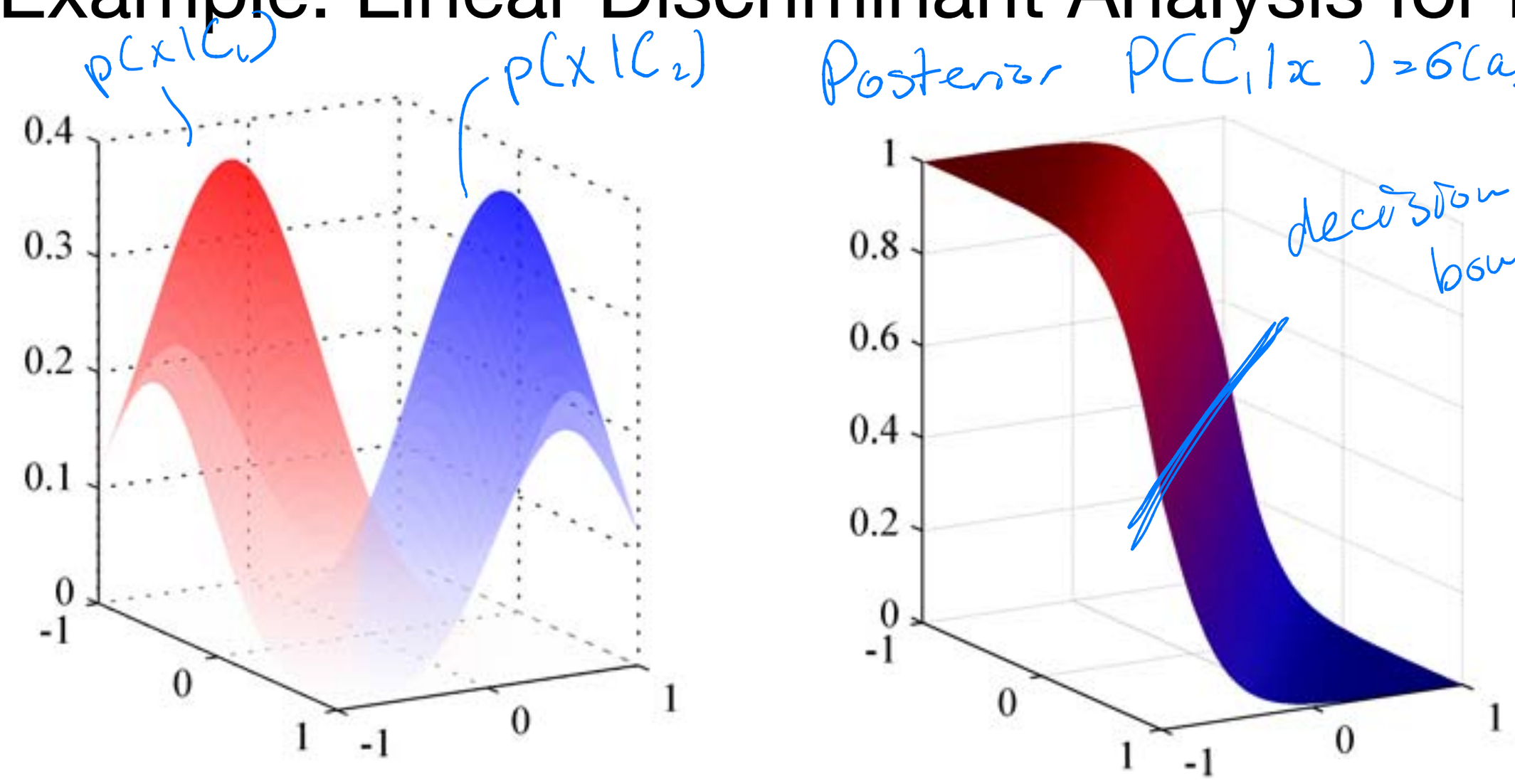

Figure: Left: class conditional densities $p(x \mid C_k)$. Right: posterior $P(C_1|x)$ as sigmoid of linear function of x. (Bishop 4.9)

Linear Discriminant Analysis: General K

- Gaussian Class-conditional densities & fixed covariance:

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

- Posterior distributions:

$$p(C_k|\mathbf{x}) = \frac{\exp(a_k(\mathbf{x}))}{\sum_{j=1}^{K} \exp(a_j(\mathbf{x}))}$$

verify

- $a_k(\mathbf{x}) = \mathbf{w}_k^T \mathbf{x} + w_{k0}$

$\mathbf{w}_k = \Sigma \, \mu_k$

$w_{k0} = -\frac{1}{2} \mu_k \Sigma^{-1} \mu_k + \ln p(C_k)$

- Decision boundary:

$p(C_k|\mathbf{x}) = p(C_j|\mathbf{x})$ → $a_k(\underline{x}) = a_j(\underline{x})$

- If all covariance matrices are different $\mathbf{\Sigma}_k \neq \mathbf{\Sigma}_j$ then $a_k(\mathbf{x})$ will also contain quadratic terms in $\mathbf{x}$

Example: LDA and QDA

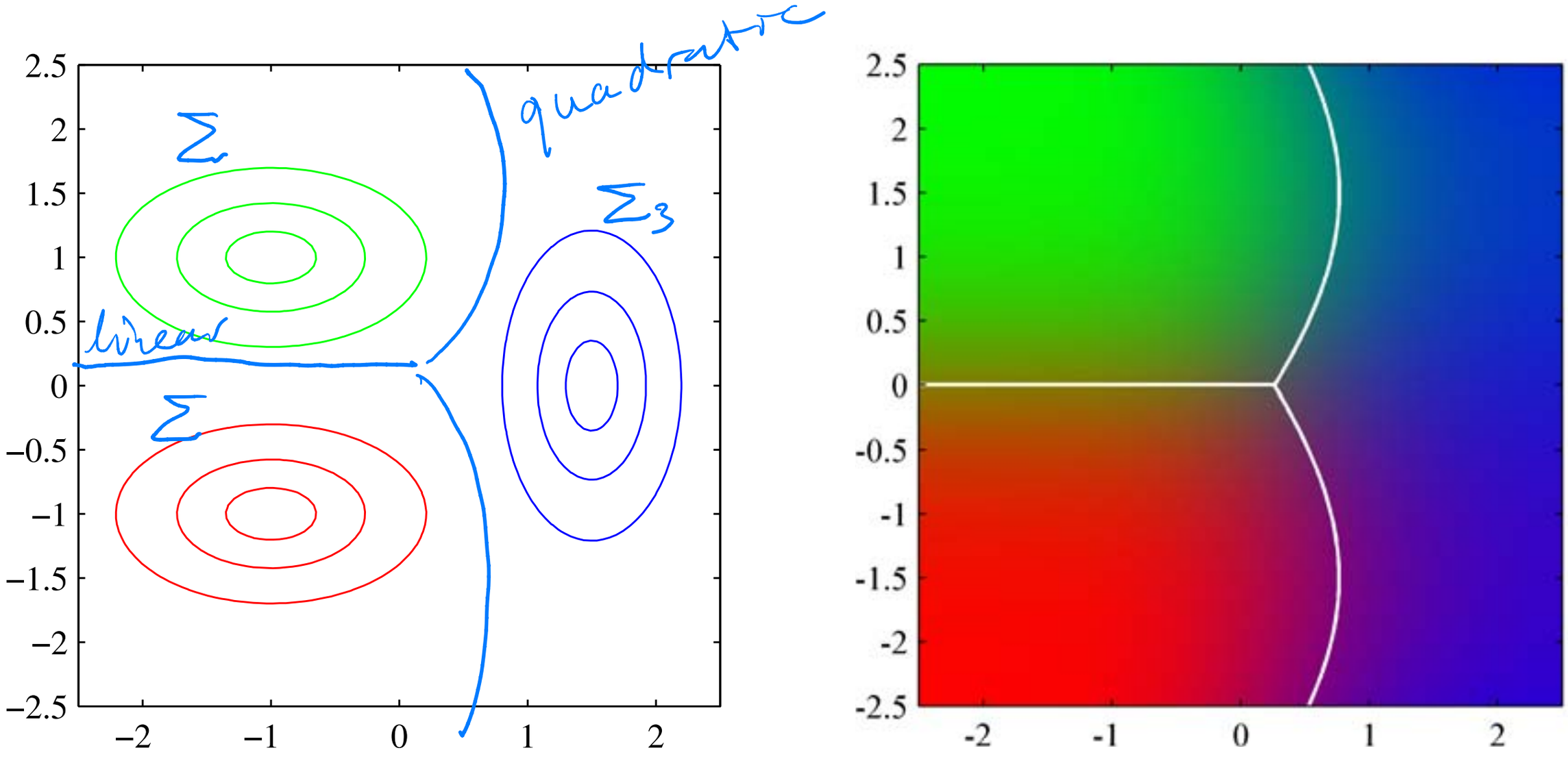

Figure: Left: Gaussian class conditional densities $p(x \mid C_k)$, red and green have same covariance matrix. Right: posterior $P(C_k \mid x)$ distributions (RGB vectors) and decision boundaries. (Bishop 4.9)