University of Amsterdam
Informatics Institute

AMLAB
Amsterdam
Machine Learning Lab

Machine Learning 1

Lecture 6 - Classification - Probabilistic Generative Models - Maximum Likelihood - Discrete Data - Discriminant Functions

Erik Bekkers

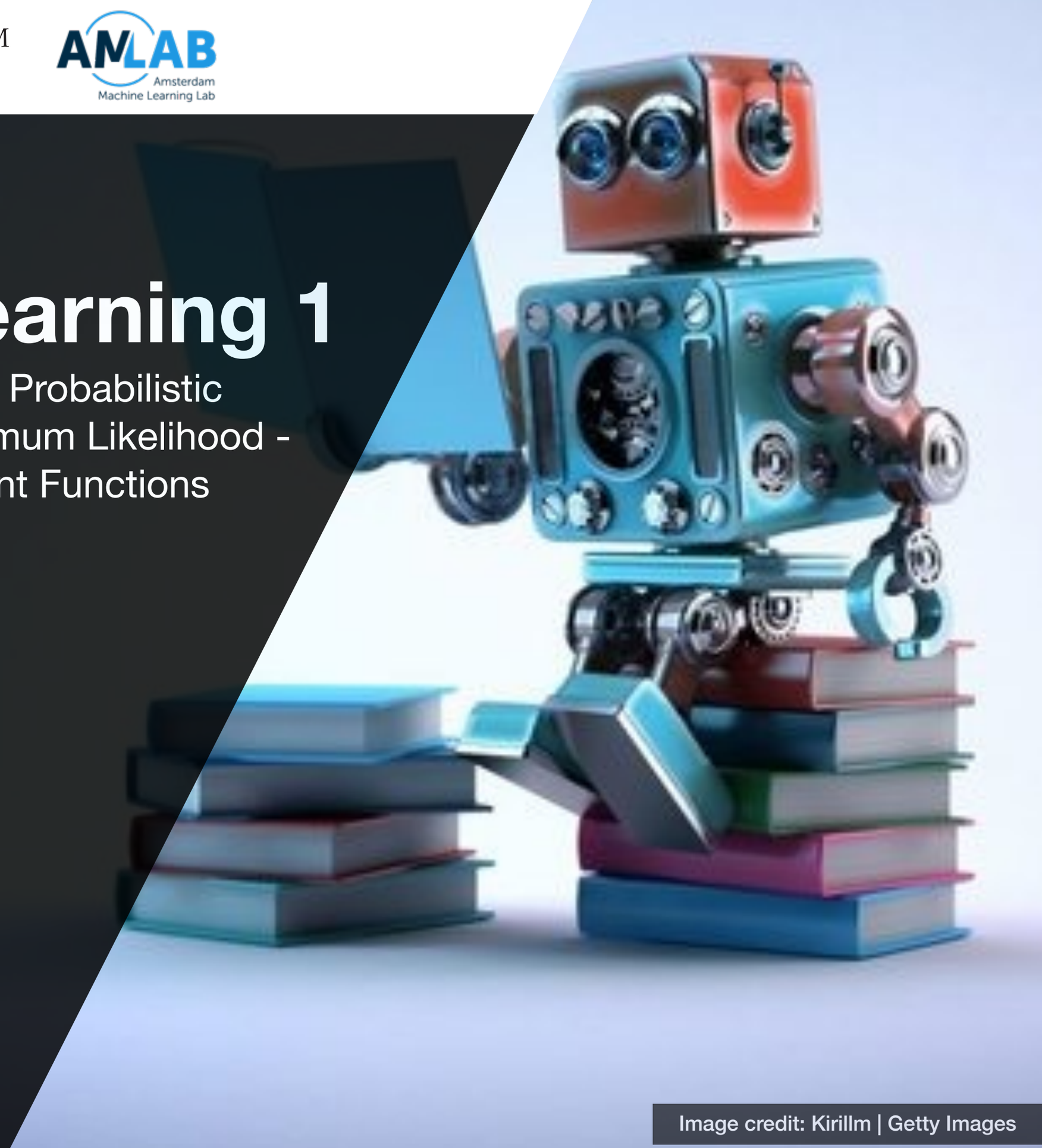
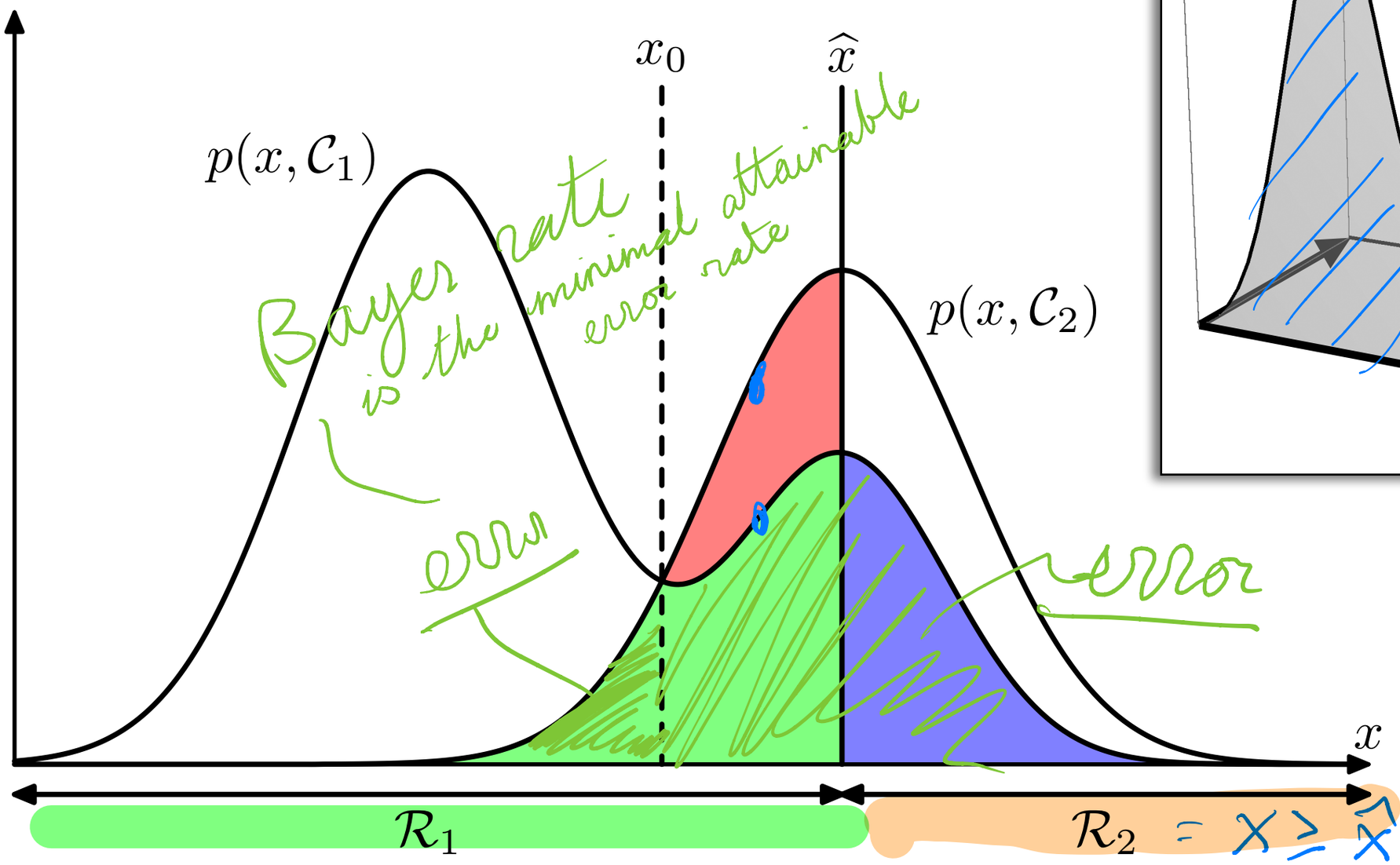


Image credit: Kirillm | Getty Images

Decision theory: Misclassification Rate

$\hat{x}$: Decision boundary

x_0 : Optimal Decision boundary



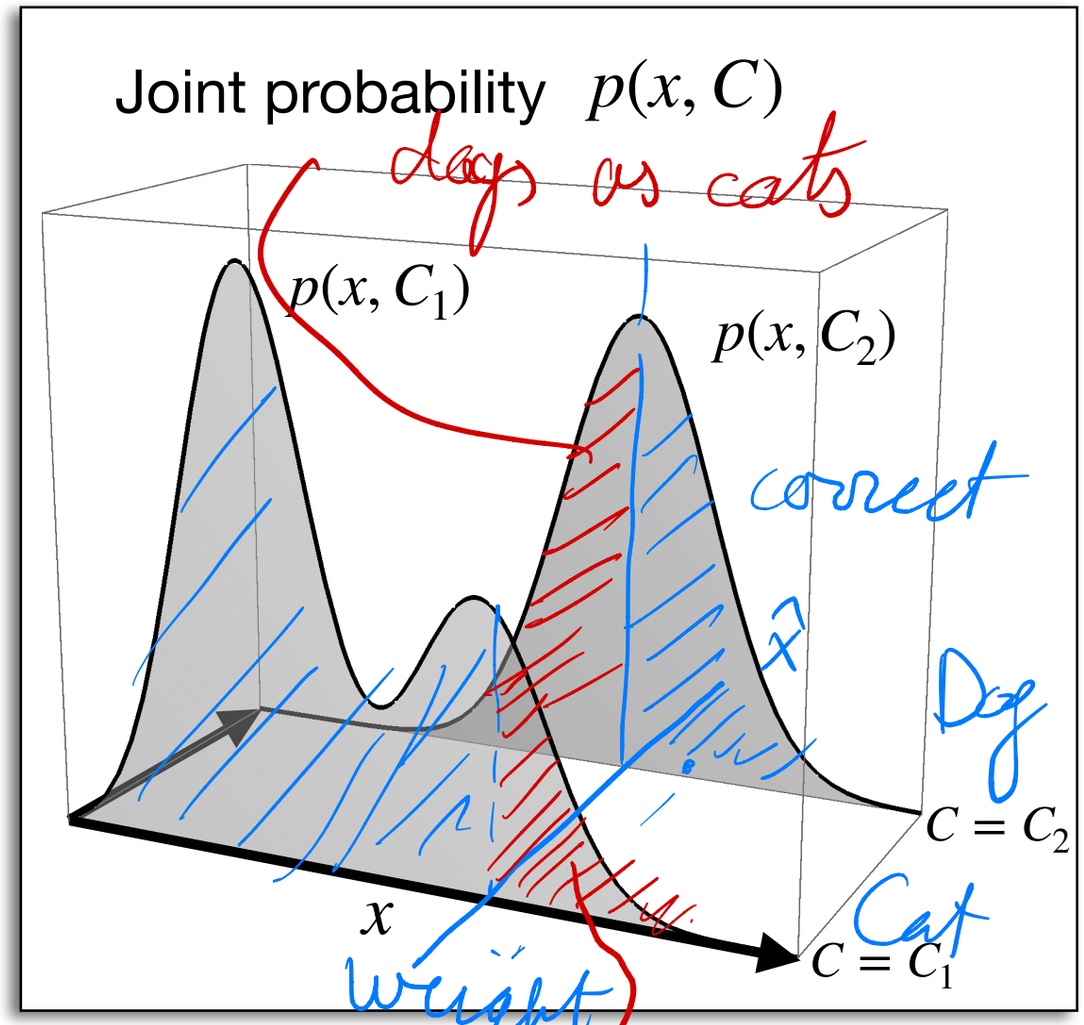


Figure: joint probability distributions and decision boundary (Bishop 1.24)

Classification Strategies

- Discriminant functions

Direct mapping of input to target: $t = y(\underline{x}; \underline{w})$

- Probabilistic discriminative models

Posterior class probabilities: $p(C_k | \underline{x})$

- Probabilistic generative models

Class-conditional densities: $p(x | C_k)$

Prior class probabilities: $p(C_k)$

$\Big\} \cdot p(\underline{x}, C_k)$

$\cdot \underline{p(C_k | \underline{x})}$

University of Amsterdam
Informatics Institute

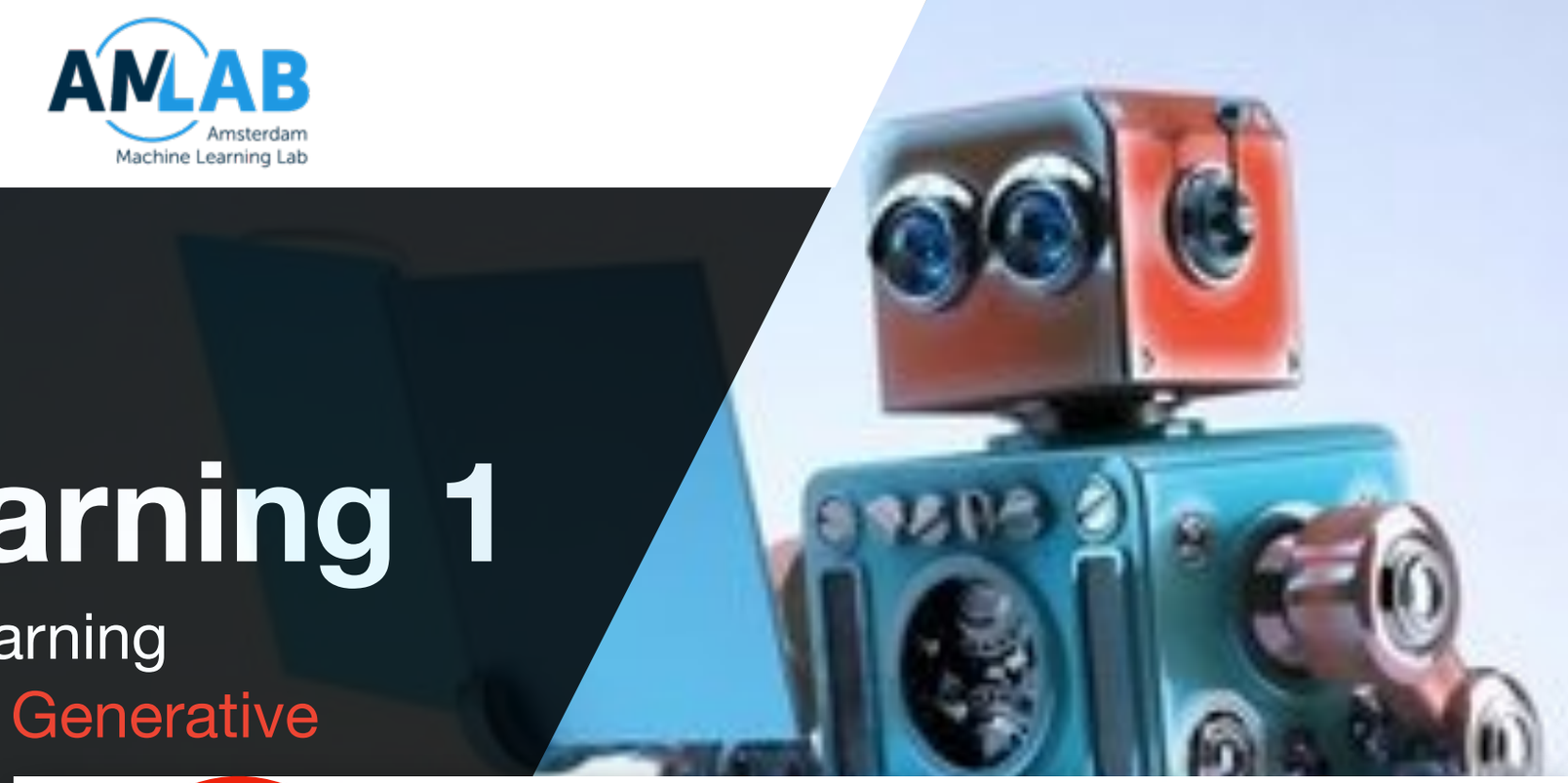
AMLAB
Amsterdam
Machine Learning Lab

Machine Learning 1

Lecture 5.6 - Supervised Learning
Classification - Probabilistic Generative Models

Erik Bekkers

(Bishop 1.5)

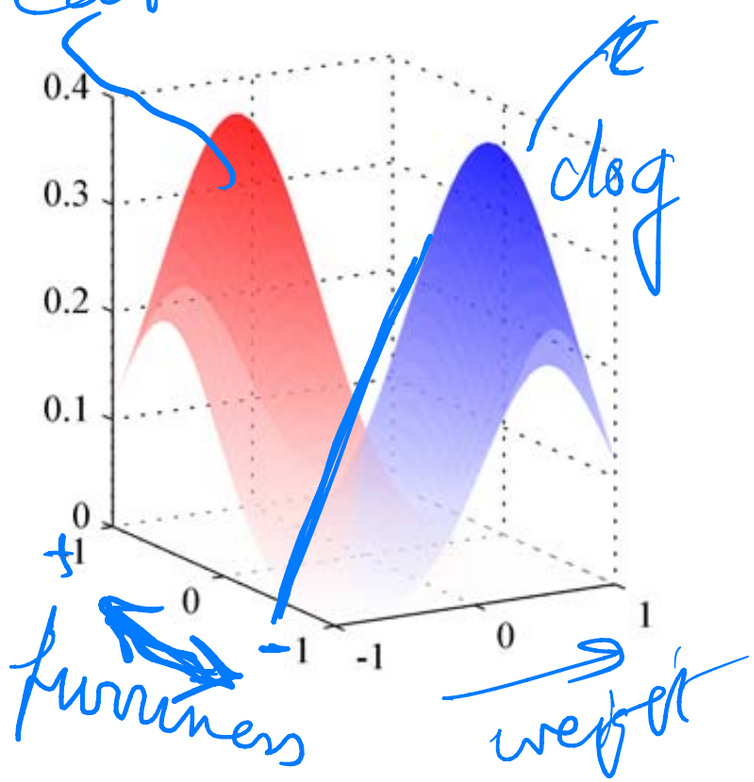


Slide credits: Patrick Forré and Rianne van den Berg

Image credit: Kirillm | Getty Images

Probabilistic Generative Models: general K

- Model:
  - Class-conditional densities: $p(\mathbf{x} \mid C_k)$
  - Prior class probabilities: $p(C_k)$



- This gives access to joint and posterior:
  - $p(\mathbf{x}, C_k)$ and $p(C_k \mid \mathbf{x})$

via product rule via Bayes

- Why are we interested in joint or posterior?
  - Decision theory tells us that the best prediction for input $\mathbf{x}$, is to choose the class with highest joint $p(\mathbf{x}, C_k)$ $= p(C_k|x)\, p(x)$
  - Or equivalently: choose class with highest posterior $p(C_k \mid \mathbf{x})$
  - Decision boundary between C_k and C_j are at $p(C_k \mid \mathbf{x}) = p(C_j \mid \mathbf{x})$

Probabilistic Generative Models: $K = 2$

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

$\log a^{-1} = -\log a$

Bayes

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

$p(\underline{x}) = \sum_{C_k} p(\underline{x}, C_k)$

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

"odds for class C_1 vs C_2"

- Log odds $a = \ln \frac{\sigma}{1 - \sigma} = \log \frac{p(\underline{x} \mid C_1) p(C_1)}{p(\underline{x} \mid C_2) p(C_2)}$

Logistic Sigmoid Function

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

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

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

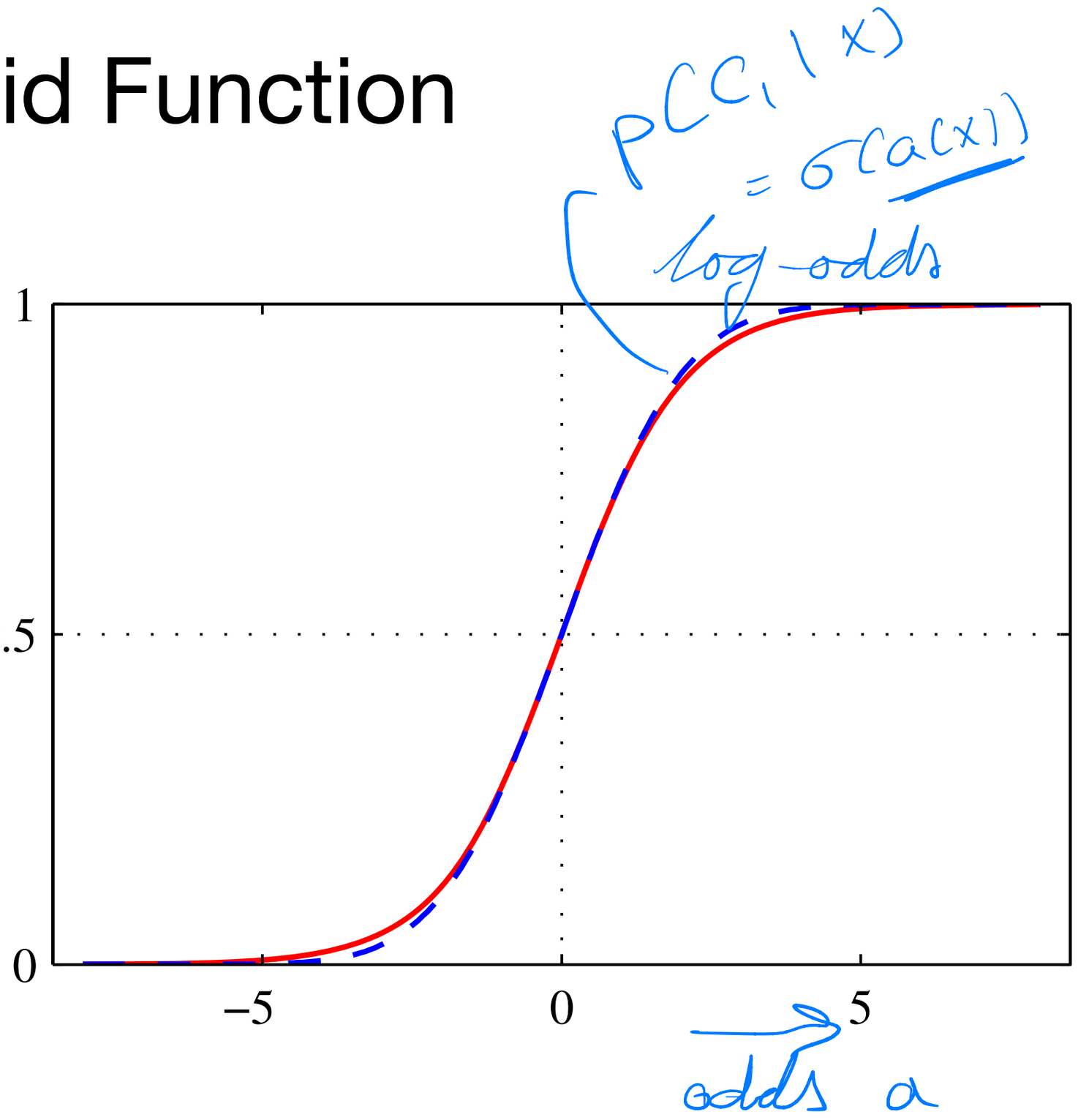


Figure: Logistic Sigmoid function (red) (Bishop 4.9)

Probabilistic Generative Models: general K

- For multiple classes (general K):

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

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

called logits

$p(C_k|x) \approx 1$

$p(C_j|x) \approx 0$

- Softmax: if $a_k \gg a_j$ for all $j \neq k$:

- Note: for $K = 2$:

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

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

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

Class Conditional Densities: Continuous Inputs

Modeling Choice

- Gaussian Class-conditional densities:

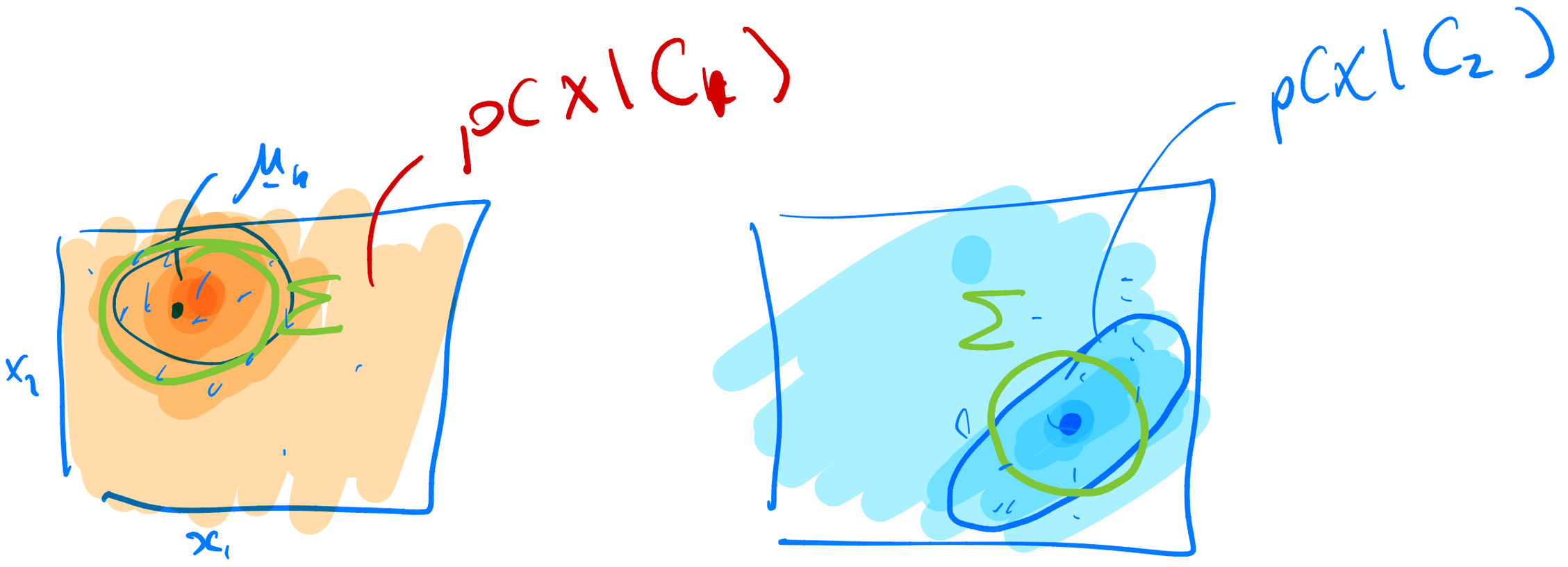
$$p(\mathbf{x}\,|\,C_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\,\Sigma_k\,|^{1/2}} \exp\{\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T\Sigma_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\}$$

- Class posteriors simplify with the assumption of shared covariance matrix: $\Sigma_k = \Sigma$

- For $K = 2$ classes they are given as $p(C_1\,|\,\mathbf{x}) = \dfrac{1}{1+\exp(-a)} = \sigma(a)$ with

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

putting bayes rule in logistic form



Class Conditional Densities: Continuous Inputs

- Gaussian Class-conditional densities:

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

- Class posteriors simplify with the assumption of shared covariance matrix: $\Sigma_k = \Sigma$

- For $K = 2$ classes they are given as $p(C_1\,|\,\mathbf{x}) = \dfrac{1}{1+\exp(-a)} = \sigma(a)$ with

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

Class Conditional Densities: Continuous Inputs

- Gaussian Class-conditional densities:

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

- Class posteriors simplify with the assumption of shared covariance matrix: $\Sigma_k = \Sigma$

- For $K = 2$ classes they are given as $p(C_1\,|\,\mathbf{x}) = \dfrac{1}{1+\exp(-a)} = \sigma(a)$ with

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

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

$$-\frac{1}{2}\mathbf{x}\Sigma^{-1}\mathbf{x} + \mathbf{x}\Sigma\boldsymbol{\mu}_1 - \frac{1}{2}\boldsymbol{\mu}_1\Sigma^{-1}\boldsymbol{\mu}_1 \qquad +\frac{1}{2}\mathbf{x}\Sigma^{-1}\mathbf{x} - \mathbf{x}\Sigma\boldsymbol{\mu}_2 + \boldsymbol{\mu}_2\Sigma\boldsymbol{\mu}_2$$

Class Conditional Densities: Continuous Inputs

- Gaussian Class-conditional densities:

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

- Class posteriors simplify with the assumption of shared covariance matrix: $\Sigma_k = \Sigma$

- For $K = 2$ classes they are given as $p(C_1\,|\,\mathbf{x}) = \dfrac{1}{1+\exp(-a)} = \sigma(a)$ with

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

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

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

$= \underline{w}^T\underline{x} + w_0$

$a(\underline{x})$ is a linear model !

Class Conditional Densities: Continuous Inputs

- Gaussian Class-conditional densities:

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

- Class posteriors simplify with the assumption of shared covariance matrix: $\Sigma_k = \Sigma$

- For $K = 2$ classes they are given as $p(C_1\,|\,\mathbf{x}) = \dfrac{1}{1+\exp(-a)} = \sigma(a)$ with

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

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

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

σ is non-linear

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

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

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

Example: Linear Discriminant Analysis for K=2

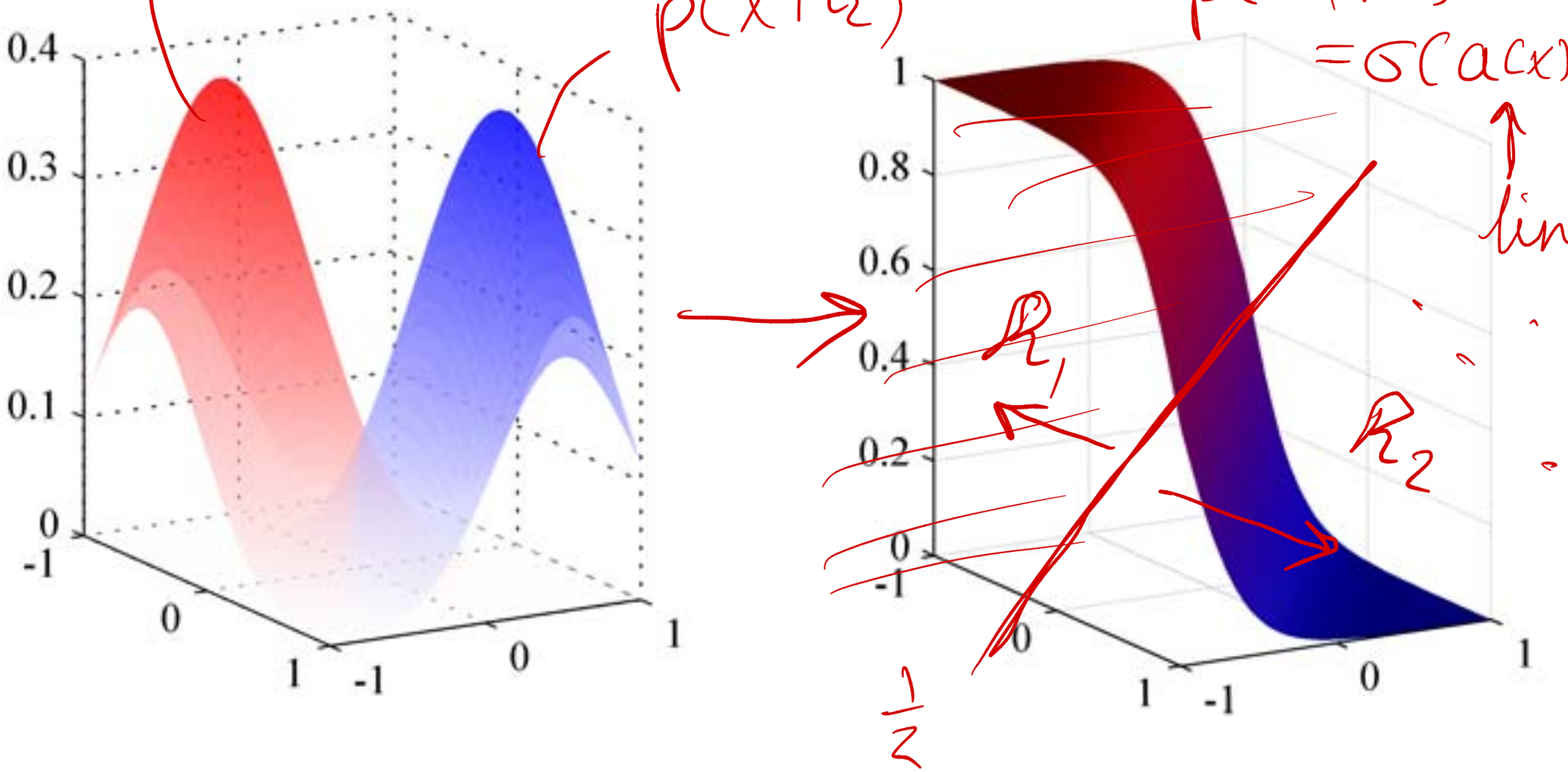


Figure: Left: class conditional densities $p(\mathbf{x} \mid C_k)$. Right: posterior $p(C_1 \mid \mathbf{x})$ as sigmoid of linear function of $\mathbf{x}$. (Bishop 4.9)

Linear Discriminant Analysis: General K

- Gaussian Class-conditional densities & fixed covariance: $\Sigma_k = \Sigma$

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

- Posterior distributions:

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

one linear model per class

has this form

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

$$\mathbf{w}_k = \Sigma^{-1} \boldsymbol{\mu}_k$$

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

- Decision boundary:

$$p(C_k \mid \mathbf{x}) = p(C_j \mid \mathbf{x}) \quad \rightarrow \quad a_k(\underline{x}) = a_j(\underline{x})$$

so linear decision boundaries!

- If all covariance matrices are different $\Sigma_k \neq \Sigma_j$ then $a_k(\mathbf{x})$ also contains quadratic terms in $\mathbf{x}$

→ non-linear d. b.

Example: LDA and QDA

Quadratic discriminant analysis

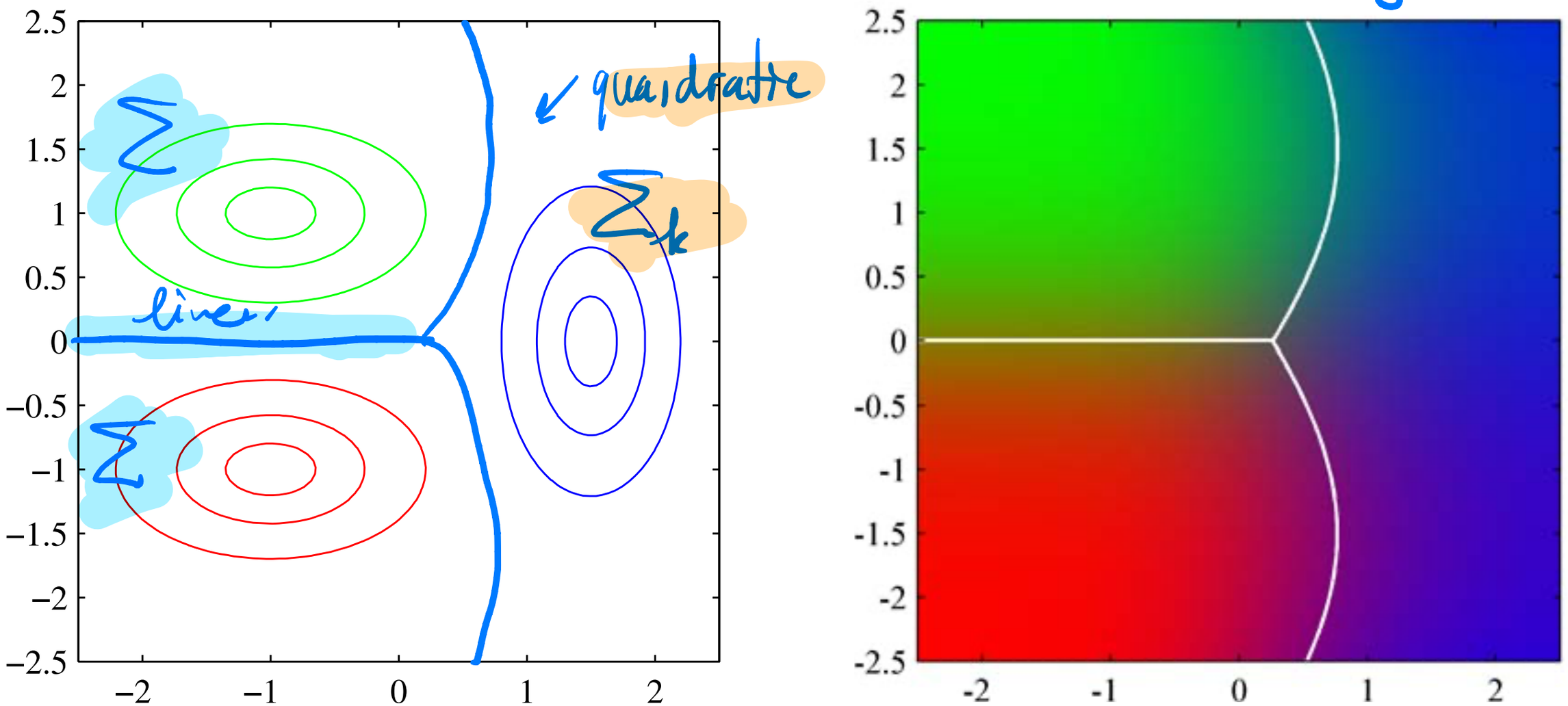


Figure: Left: Gaussian class conditional densities $p(\mathbf{x} \mid C_k)$, red and green have same covariance matrix. Right: posterior $p(C_k \mid \mathbf{x})$ distributions (RGB vectors) and decision boundaries. (Bishop 4.9)

University of Amsterdam
Informatics Institute

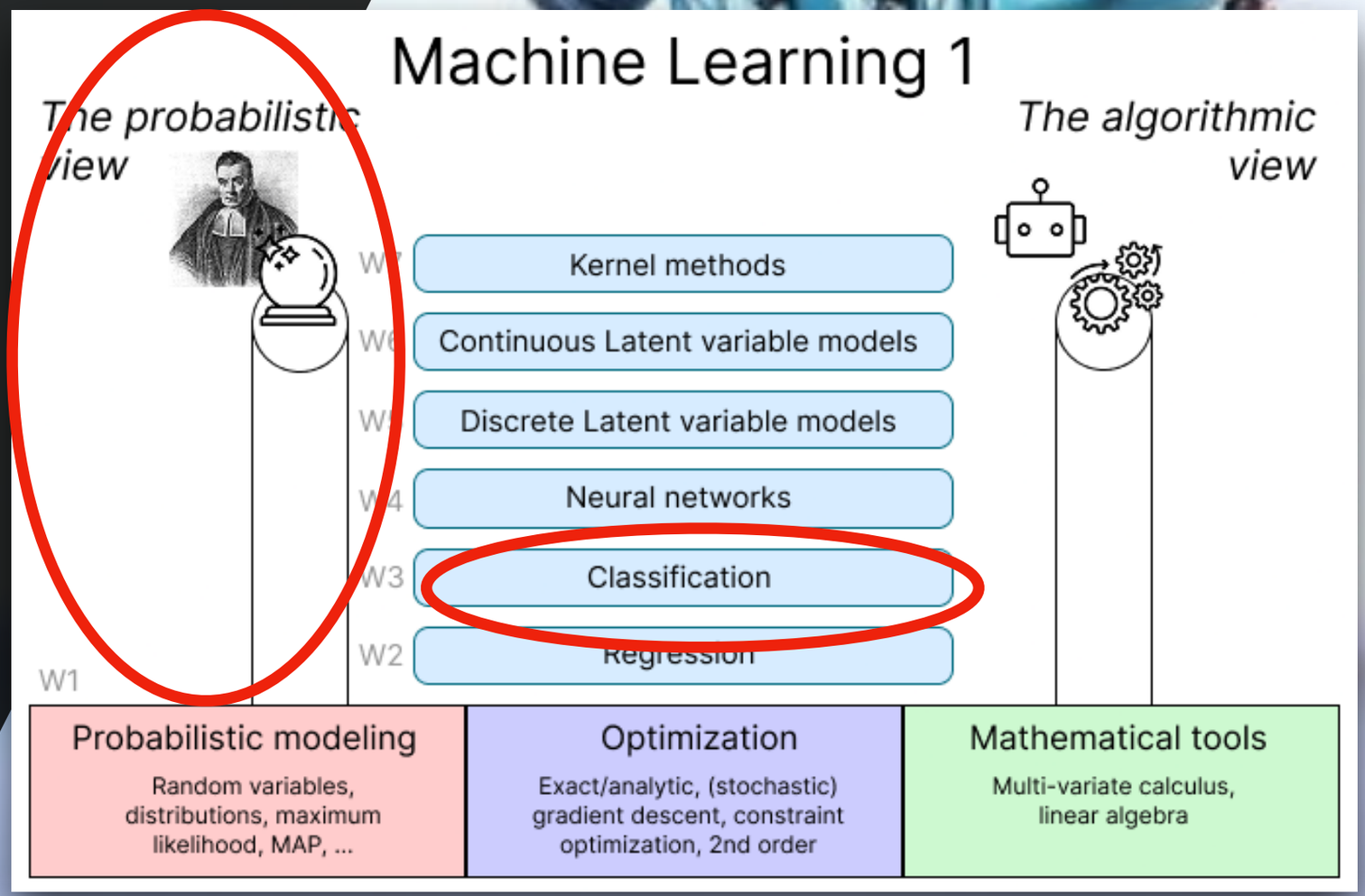
AMLAB
Amsterdam
Machine Learning Lab

Machine Learning 1

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

Erik Bekkers

(Bishop 4.2.2)

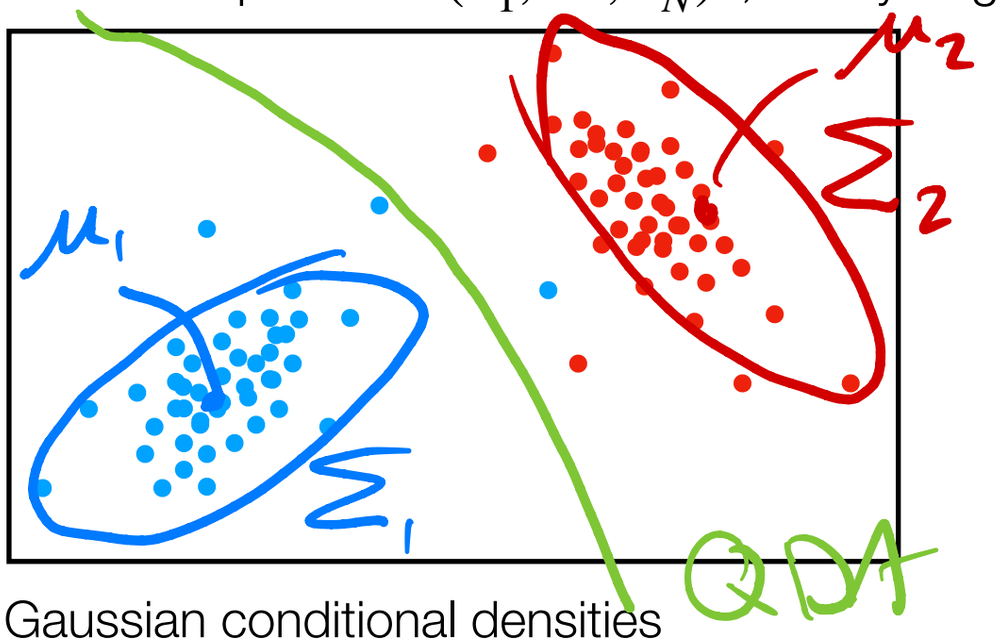


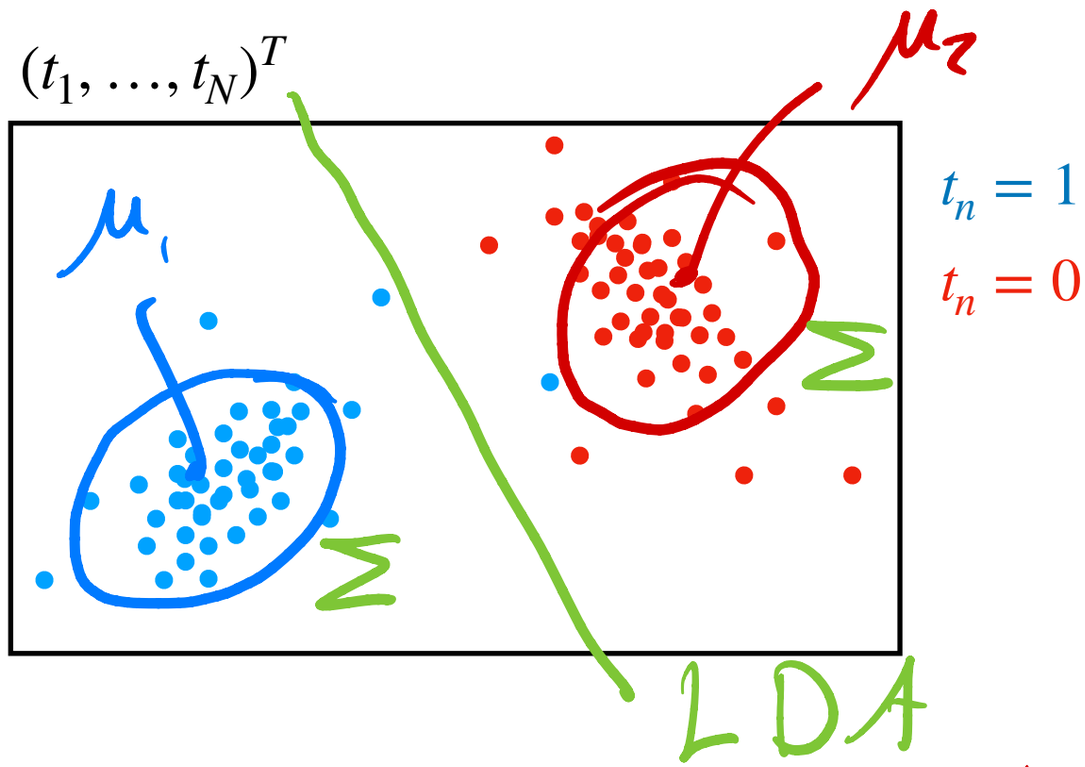
Slide credits: Patrick Forré and Rianne van den Berg

Image credit: Kirillm | Getty Images

LDA: Maximum Likelihood for K=2

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





- Gaussian conditional densities

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

how to obtain? ML!!

- Strategy: (the parameters ...)
  - use maximum likelihood to estimate $\boldsymbol{\mu}_k$, Σ and priors $p(C_k)$
  - let q parametrize the prior probabilities with q: $p(C_1) = q$ and $p(C_2) = 1 - q$

$q \in [0,1]$ Bernoulli

For $\mathbf{x}_n$ with $t_n = 1$: $p(\mathbf{x}_n, C_1) = p(\mathbf{x}_n \mid C_1) p(C_1) =$ $q\, N(\underline{x}_n \mid \mu_1, \Sigma)$

For $\mathbf{x}_n$ with $t_n = 0$: $p(\mathbf{x}_n, C_2) = p(\mathbf{x}_n \mid C_2) p(C_2) =$ $(1-q)\, N(\underline{x}_n \mid \mu_2, \Sigma)$

LDA: Maximum Likelihood for K=2

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

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

- Strategy:
  - use maximum likelihood to estimate $\boldsymbol{\mu}_k$, Σ and priors $p(C_k)$
  - let q parametrize the prior probabilities with q: $p(C_1) = q$ and $p(C_2) = 1 - q$
- Likelihood:

$$p(\mathbf{t}, \mathbf{X} \mid q, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) = \prod_{n=1}^{N} p(\mathbf{x}_n, t_n) = \prod_{n=1}^{N} p(\mathbf{x}_n \mid t_n) p(t_n)$$

$$= \prod_{n=1}^{N} \left[p(\mathbf{x}_n \mid C_1) p(C_1)\right]^{t_n} \left[p(\mathbf{x}_n \mid C_2) p(C_2)\right]^{1-t_n}$$

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

LDA: Maximum Likelihood for K=2

- Likelihood

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

$\ln a^t = t \ln a$

$\ln a \cdot b = \ln a + \ln b$

tricky!

- Log likelihood

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

prior prob

- Let's estimate q by solving $\dfrac{\partial}{\partial q} \ln p(\mathbf{t}, \mathbf{X} \,|\, q, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) = 0$

$$\frac{\partial}{\partial q} \ln p(\mathbf{t}, \mathbf{X} \,|\, q, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) =$$

LDA: Maximum Likelihood for K=2

- Likelihood

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

- Log likelihood

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

- Let's estimate q by solving $\frac{\partial}{\partial q} \ln p(\mathbf{t}, \mathbf{X} \mid q, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) = 0$

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

LDA: Maximum Likelihood for K=2

- Likelihood

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

- Log likelihood

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

- Let's estimate q by solving $\dfrac{\partial}{\partial q} \ln p(\mathbf{t}, \mathbf{X} \,|\, q, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) = 0$

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

- Solve for q:

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

LDA: Maximum Likelihood for K=2

- Likelihood

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

- Log likelihood

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

- Let's estimate q by solving $\dfrac{\partial}{\partial q} \ln p(\mathbf{t}, \mathbf{X} \,|\, q, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) = 0$

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

- Solve for q:

$$\sum_{n=1}^{N} \frac{t_n - q}{q(1-q)} = 0 \quad \overset{(q \neq 0, q \neq 1)}{\Rightarrow} \quad \sum_{n=1}^{N} q = \sum_{n=1}^{N} t_n$$

$N \cdot q$ ← N_1 is # times $t_n = 1$

LDA: Maximum Likelihood for K=2

- Likelihood

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

- Log likelihood

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

- Let's estimate q by solving $\dfrac{\partial}{\partial q} \ln p(\mathbf{t}, \mathbf{X} \,|\, q, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) = 0$

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

- Solve for q:

$$\sum_{n=1}^{N} \frac{t_n - q}{q(1-q)} = 0 \quad \overset{(q \neq 0, q \neq 1)}{\Rightarrow} \quad \sum_{n=1}^{N} q = \sum_{n=1}^{N} t_n \quad \Leftrightarrow \quad \boxed{q = \frac{1}{N} \sum_{n=1}^{N} t_n}$$

Nq

ML solution to $p(C_1) = q$ is the fraction of points with label $t_n = 1$

LDA: Maximum Likelihood for K=2

- Log likelihood

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

- Let's estimate $\boldsymbol{\mu}_1$ by solving $\frac{\partial}{\partial \boldsymbol{\mu}_1} \ln p(\mathbf{t}, \mathbf{X} \mid q, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) = \mathbf{0}^T$

$$\frac{\partial}{\partial \boldsymbol{\mu}_1} \ln p(\mathbf{t}, \mathbf{X} \mid q, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) = \frac{\partial}{\partial \boldsymbol{\mu}_1} \sum_{n=1}^{N} t_n \ln \mathcal{N}(\mathbf{x}_n \mid \mu_1, \Sigma)$$
$$= -\frac{1}{2} \frac{\partial}{\partial \boldsymbol{\mu}_1} \sum_{n=1}^{N} t_n (\mathbf{x}_n - \boldsymbol{\mu}_1)^T \Sigma^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_1)$$

LDA: Maximum Likelihood for K=2

- Log likelihood

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

- Let's estimate $\boldsymbol{\mu}_1$ by solving $\frac{\partial}{\partial \boldsymbol{\mu}_1} \ln p(\mathbf{t}, \mathbf{X} \mid q, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) = \mathbf{0}^T$

$$\frac{\partial}{\partial \boldsymbol{\mu}_1} \ln p(\mathbf{t}, \mathbf{X} \mid q, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) = \frac{\partial}{\partial \boldsymbol{\mu}_1} \sum_{n=1}^{N} t_n \ln \mathcal{N}(\mathbf{x}_n \mid \mu_1, \Sigma)$$
$$= -\frac{1}{2} \frac{\partial}{\partial \boldsymbol{\mu}_1} \sum_{n=1}^{N} t_n (\mathbf{x}_n - \boldsymbol{\mu}_1)^T \Sigma^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_1) \quad (\Sigma \text{ is symmetric}) = \sum_{n=1}^{N} t_n (\mathbf{x}_n - \boldsymbol{\mu}_1)^T \Sigma^{-1}$$

- Solve for $\boldsymbol{\mu}_1$:

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

$$(\Sigma \text{ is pos. def.}) \Rightarrow \sum_{n=1}^{N} t_n \boldsymbol{\mu}_1 = \sum_{n=1}^{N} t_n \mathbf{x}_n$$

LDA: Maximum Likelihood for K=2

- Log likelihood

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

- Let's estimate $\boldsymbol{\mu}_1$ by solving $\frac{\partial}{\partial \boldsymbol{\mu}_1} \ln p(\mathbf{t}, \mathbf{X} \mid q, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) = \mathbf{0}^T$

$$\frac{\partial}{\partial \boldsymbol{\mu}_1} \ln p(\mathbf{t}, \mathbf{X} \mid q, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) = \frac{\partial}{\partial \boldsymbol{\mu}_1} \sum_{n=1}^{N} t_n \ln \mathcal{N}(\mathbf{x}_n \mid \mu_1, \Sigma)$$
$$= -\frac{1}{2} \frac{\partial}{\partial \boldsymbol{\mu}_1} \sum_{n=1}^{N} t_n (\mathbf{x}_n - \boldsymbol{\mu}_1)^T \Sigma^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_1) \quad \overset{(\Sigma \text{ is symmetric})}{=} \quad \sum_{n=1}^{N} t_n (\mathbf{x}_n - \boldsymbol{\mu}_1)^T \Sigma^{-1}$$

- Solve for $\boldsymbol{\mu}_1$:

$$\sum_{n=1}^{N} t_n (\mathbf{x}_n - \boldsymbol{\mu}_1)^T \Sigma^{-1} = \mathbf{0}^T \quad \overset{(\Sigma \text{ is pos. def.})}{\Rightarrow} \quad \sum_{n=1}^{N} t_n \boldsymbol{\mu}_1 = \sum_{n=1}^{N} t_n \mathbf{x}_n \quad \Leftrightarrow \quad N_1 \, \boldsymbol{\mu}_1 = \sum_{n=1}^{N} t_n \mathbf{x}_n$$

$\underline{\mu}_{1,ML} = \frac{1}{N_1} \sum_{n=1}^{N} t_n \underline{x}_n$ ← t_n = 0 or 1

average value for x_n for the cases $t_n = 1$

$\mu_{2,ML} = \frac{1}{N_2} \sum (1 - t_n) \underline{x}_n$

LDA: Maximum Likelihood for K=2

- Log likelihood

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

- Let's estimate Σ by solving $\dfrac{\partial}{\partial \Sigma} \ln p(\mathbf{t}, \mathbf{X} \,|\, q, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) = \mathbf{O}$

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

- Solve $\dfrac{\partial}{\partial \Sigma} \ln p(\mathbf{t}, \mathbf{X} \,|\, q, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) = \mathbf{O}$ for $\boldsymbol{\mu}_1$ (Bishop 4.2.2): + Notes on Canvas

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

Σ_1 — sample cov of class 1 — Σ_2 — .. class 2

Σ_1 Σ_2

LDA: Maximum Likelihood for K=2

The ML solutions:

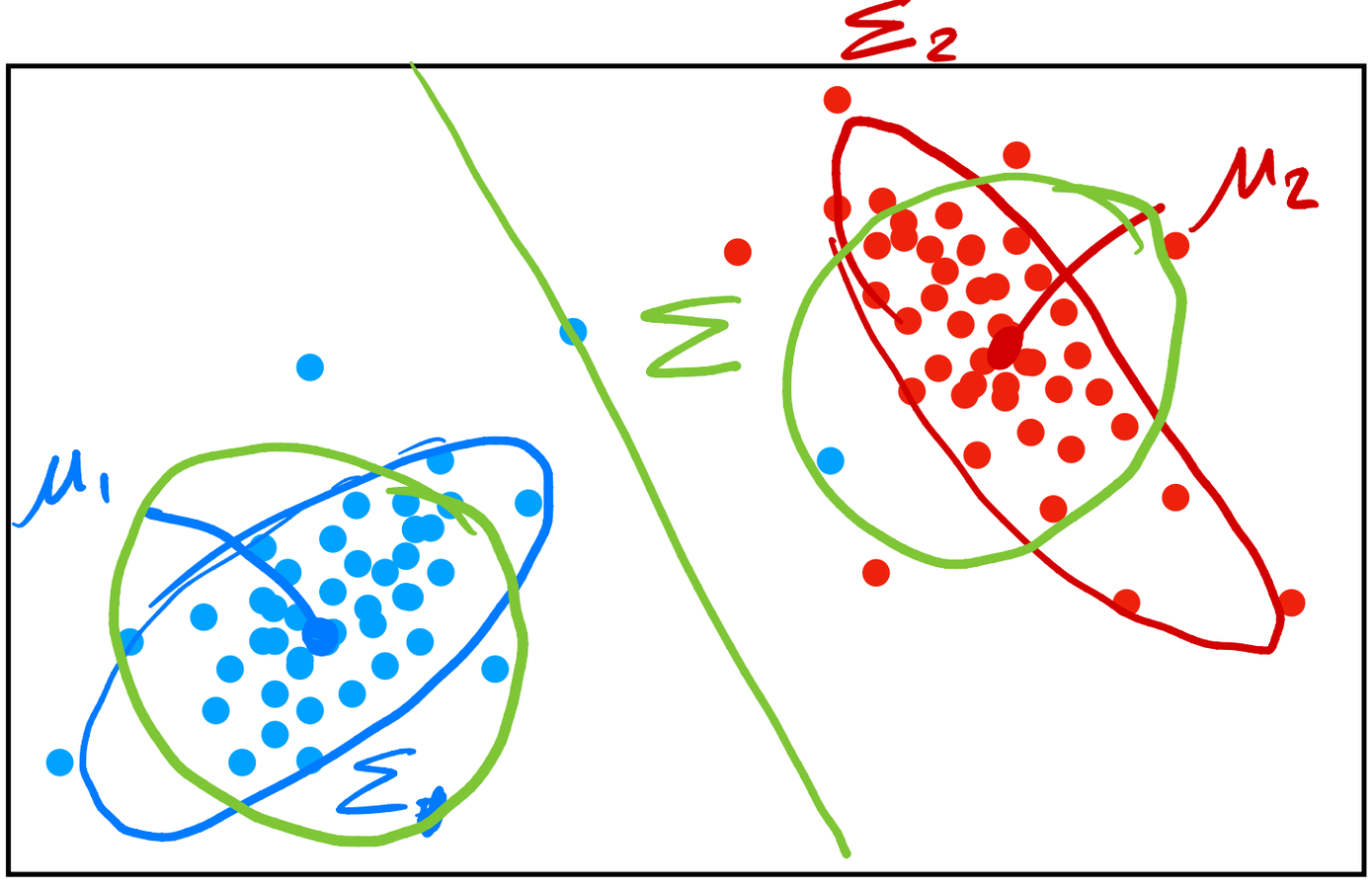
$$\boldsymbol{\mu}_{1,\mathrm{ML}} = \frac{1}{N_1}\sum_{n=1}^{N} t_n \mathbf{x}_n \qquad \boldsymbol{\mu}_{2,\mathrm{ML}} = \frac{1}{N_2}\sum_{n=1}^{N} (1-t_n)\mathbf{x}_n \qquad q_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} t_n = \frac{N_1}{N}$$

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

For the joint probabilities:

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

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



LDA: prediction for K=2

- For new datapoint $\mathbf{x}'$:

derived before

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

lin model given by ML

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

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

- Assign $\mathbf{x}'$ to C_1 if $p(C_1 \mid \mathbf{x}') > \frac{1}{2}$

- Disadvantage of LDA:
  - Gaussian distribution is sensitive to outliers (see averaging)
  - Linearity/handcrafted features restrict application X or use $\phi(k)$
  - Maximum likelihood is prone to overfitting no regularization

University of Amsterdam
Informatics Institute

AMLAB
Amsterdam
Machine Learning Lab

Machine Learning 1

Lecture 6.2 - Supervised Learning
Classification - Probabilistic Generative Models - Discrete Variables (Naive Bayes)

Erik Bekkers

(Bishop 4.2.3)

Slide credits: Patrick Forré and Rianne van den Berg

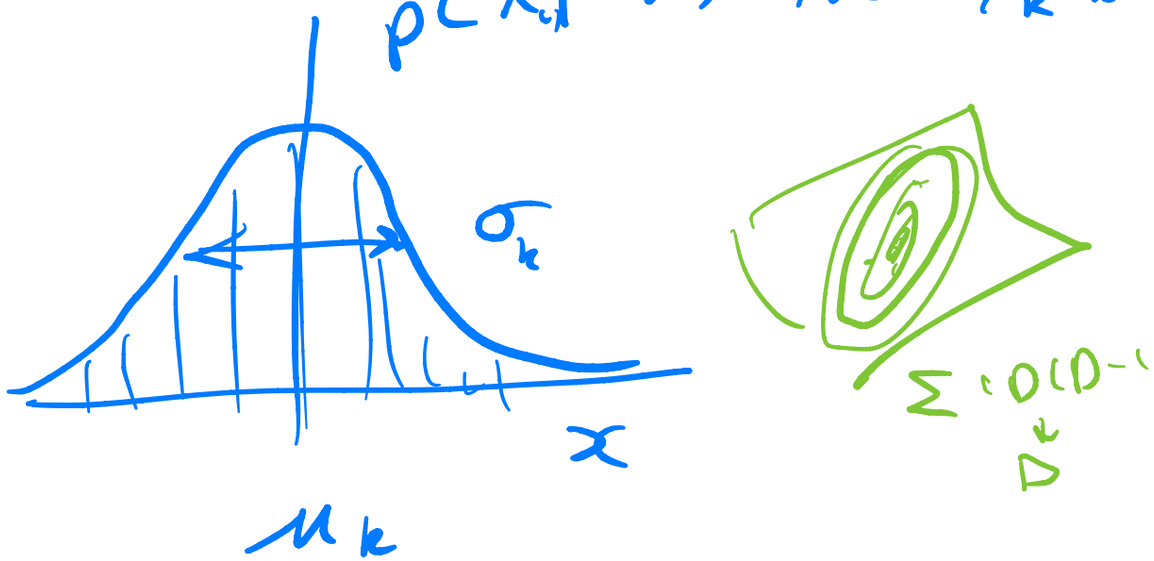
Image credit: Kirillm | Getty Images

Probabilistic Generative Models: Discrete ($K = 2$)

- Input: discrete feature vectors $\mathbf{x}_n = (x_1, \ldots, x_D)^T$ $= (0, 1, 1, 0, \ldots)^T$
  - $x_i \in \{0,1\}$ $\quad \underline{x_n} \in \{0,1\}^D$ ask after drawing
- For D-dimensional input # of parameters for $p(\mathbf{x} \mid C_k)$ per class: ? $2^D - 1$

(D = 1)

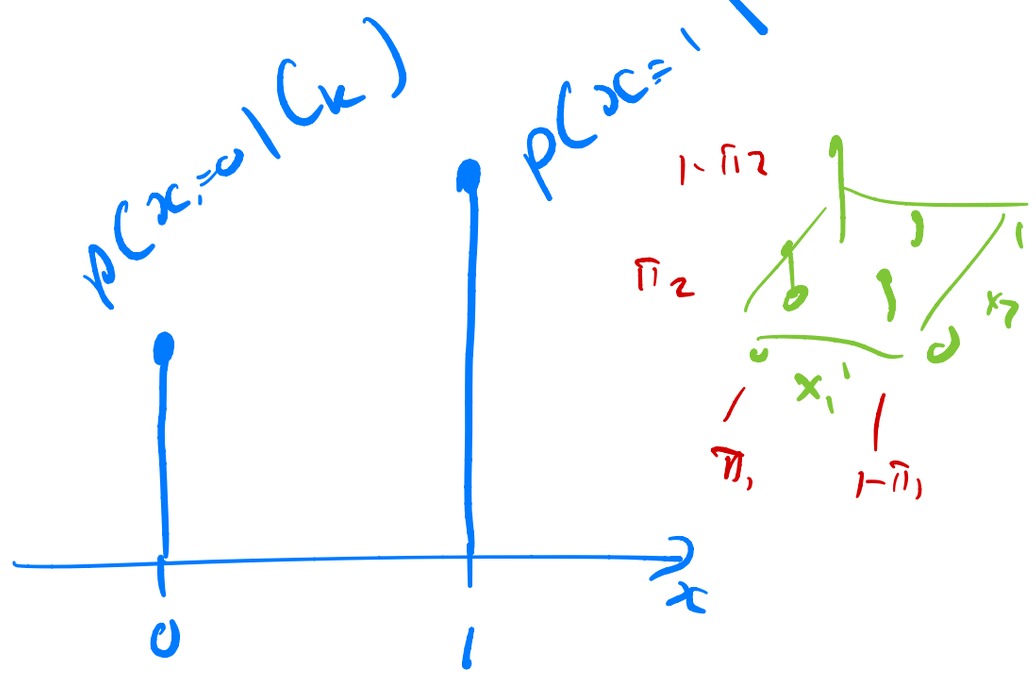
Before $x \in \mathbb{R}$



A Gaussian $N(x \mid \mu, \sigma)$ assigns prob to all possible $x \in \mathbb{R}$

(D = 1)

Now



assign probs to all possible cases separately

Probabilistic Generative Models: Discrete ($K = 2$)

- Input: discrete feature vectors $\mathbf{x}_n = (x_1, \ldots, x_D)^T$
  - $x_i \in \{0,1\}$
- For D-dimensional input # of parameters for $p(\mathbf{x} \mid C_k)$ per class: $2^D - 1$
- Naive Bayes assumption: feature values are treated as independent *when conditioned on class C_k!*

$$p(\mathbf{x} \mid C_k) = \prod_{i=1}^{D} p(x_i \mid C_k) = \prod_{i=1}^{D} (\pi_{ki})^{x_i} (1 - \pi_{ki})^{1 - x_i}$$

$$p(x_1, x_2, \ldots, x_D \mid C_k) = p(x_1 \mid C_1)\, p(x_2 \mid C_k) \cdots p(x_D \mid C_k)$$

with class conditional parameters $\boldsymbol{\pi}_{ki}$.

params = D

Bernoulli: $p(x_i = 1 \mid C_k) = \pi_{ki}$

Probabilistic Generative Models: Discrete ($K = 2$)

- Input: discrete feature vectors $\mathbf{x}_n = (x_1, \ldots, x_D)^T$
  - $x_i \in \{0,1\}$
- For D-dimensional input # of parameters for $p(\mathbf{x} \mid C_k)$ per class: $2^D - 1$
- Naive Bayes assumption: feature values are treated as independent *when conditioned on class C_k!*

$$p(\mathbf{x} \mid C_k) = \prod_{i=1}^{D} p(x_i \mid C_k) = \prod_{i=1}^{D} \pi_{ki}^{x_i}(1 - \pi_{ki})^{1-x_i}$$

with class conditional parameters π_{ki}.

- Class posterior: $p(C_k \mid \mathbf{x}) = \dfrac{\exp(a_k(\mathbf{x}))}{\sum_{j=1}^{K} \exp(a_j(\mathbf{x}))}$ and prior model parameters $p(C_k) = q_k$
  - $a_k = \ln p(\mathbf{x} \mid C_k)p(C_k) = \ln p(\mathbf{x} \mid C_k) + \ln p(C_k)$

then simplifies to:

$$= \sum_{i=1}^{D} x_i \ln \pi_{ki} + (1 - x_i) \ln (1 - \pi_{ki}) + \ln q_k$$

University of Amsterdam
Informatics Institute

AMLAB
Amsterdam
Machine Learning Lab

Machine Learning 1

Lecture 6.3 - Supervised Learning
Classification - Discriminative Models

Erik Bekkers

(Bishop 4.0, 4.1.1, 4.1.2)

Image credit: Kirillm | Getty Images

Slide credits: Patrick Forré and Rianne van den Berg

Classification Strategies

- Discriminant functions

Direct mapping of input to target: $t = y(\mathbf{x}, \mathbf{w})$

- Probabilistic discriminative models

Posterior class probabilities: $p(C_k \mid \mathbf{x})$

- Probabilistic generative models

Class-conditional densities: $p(\mathbf{x} \mid C_k)$

Prior class probabilities: $p(C_k)$

means we can sample from the learned data distribution

LDA

Discriminant Functions: Two Classes

- Input: $\mathbf{x} \in \mathbb{R}^D$
- Targets: $t \in \{C_1, C_2\}$ = $t=0$, $t=1$

Discriminant functions:

- Direct mapping of input to target

We now consider generalized Linear Models (GLM)

- $y(\mathbf{x}, \tilde{\mathbf{w}}) = f(\tilde{\mathbf{w}}^T \boldsymbol{\phi}(\mathbf{x}))$
- with $\boldsymbol{\phi}(\mathbf{x}) = (1, \phi_1(\mathbf{x}), \ldots, \phi_{M-1}(\mathbf{x}))^T$
- Decision boundary: $y(\mathbf{x}, \tilde{\mathbf{w}}) = \text{const}$

"so, why not directly model such a y?"

Discriminant Functions: Two Classes

- Simplest discriminant function ($f(\mathbf{x}) = x$, canonical basis $\phi_0(\mathbf{x}) = 1$, $\phi_j(\mathbf{x}) = x_j, \quad j = 1, \ldots, D$)

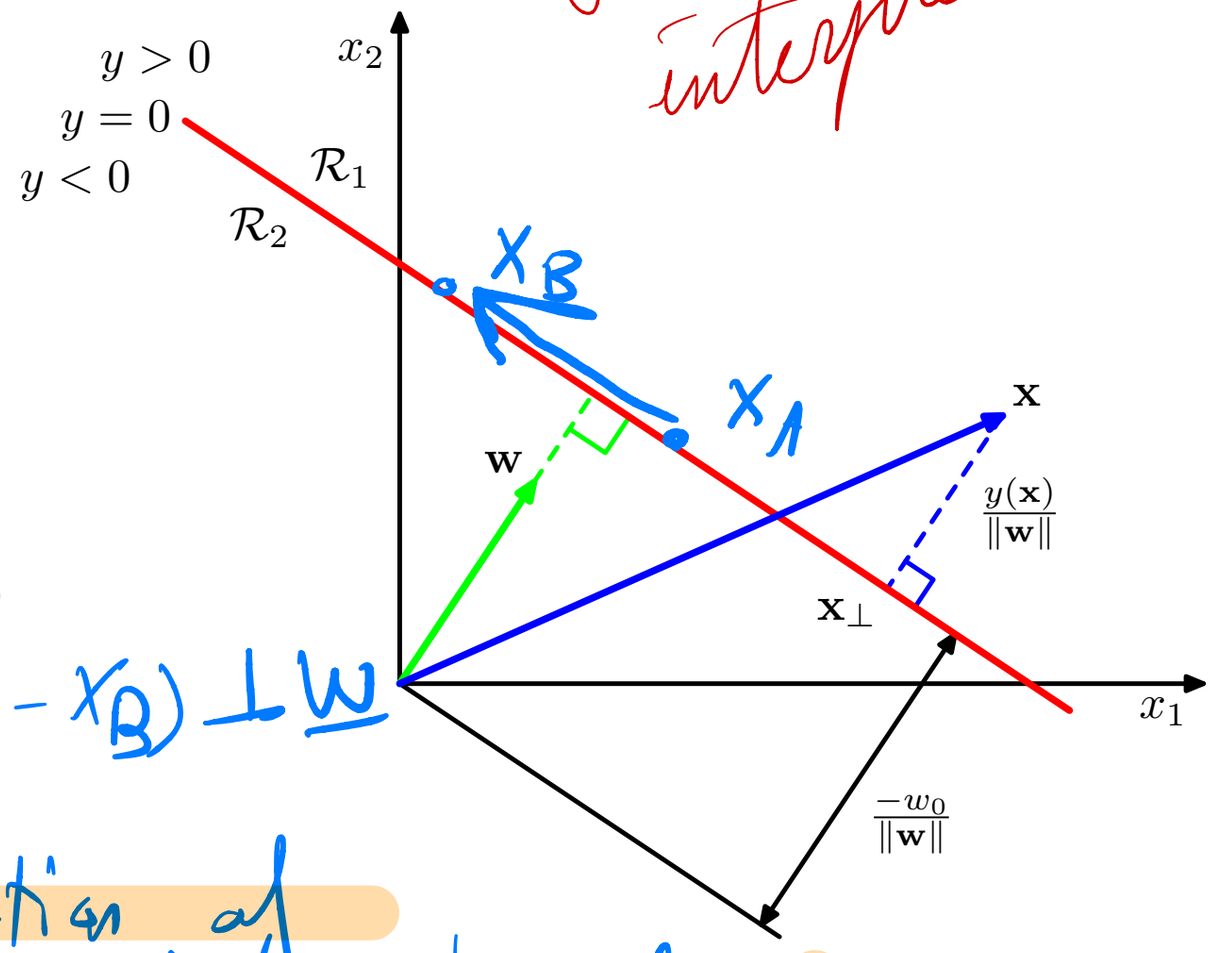
$$y(\mathbf{x}, \tilde{\mathbf{w}}) = \mathbf{w}^T\mathbf{x} + w_0$$

Still useful to establish some intuition.

- Decision boundary: $y(\mathbf{x}, \tilde{\mathbf{w}}) = 0$

$\underline{w}^T \underline{x} + w_0 = 0$

Geometric interpretation



- Consider 2 datapoints $\mathbf{x}_A$ and $\mathbf{x}_B$ on decision boundary

$w^T x_A + w_0 = w^T x_B + w_0$

$$y(\mathbf{x}_A) = y(\mathbf{x}_B) = 0$$

i.e.,

$w^T(x_B - x_A) = 0$

$(x_A - x_B) \perp w$

$\Rightarrow$ w determines orientation of decision boundary

"normal vector of the hyperplane"

Figure: 2-class decision boundary (Bishop 4.1)

Discriminant Functions: Two Classes

- Take $\mathbf{x}'$ a point on a decision surface: $y(\mathbf{x}') = 0$

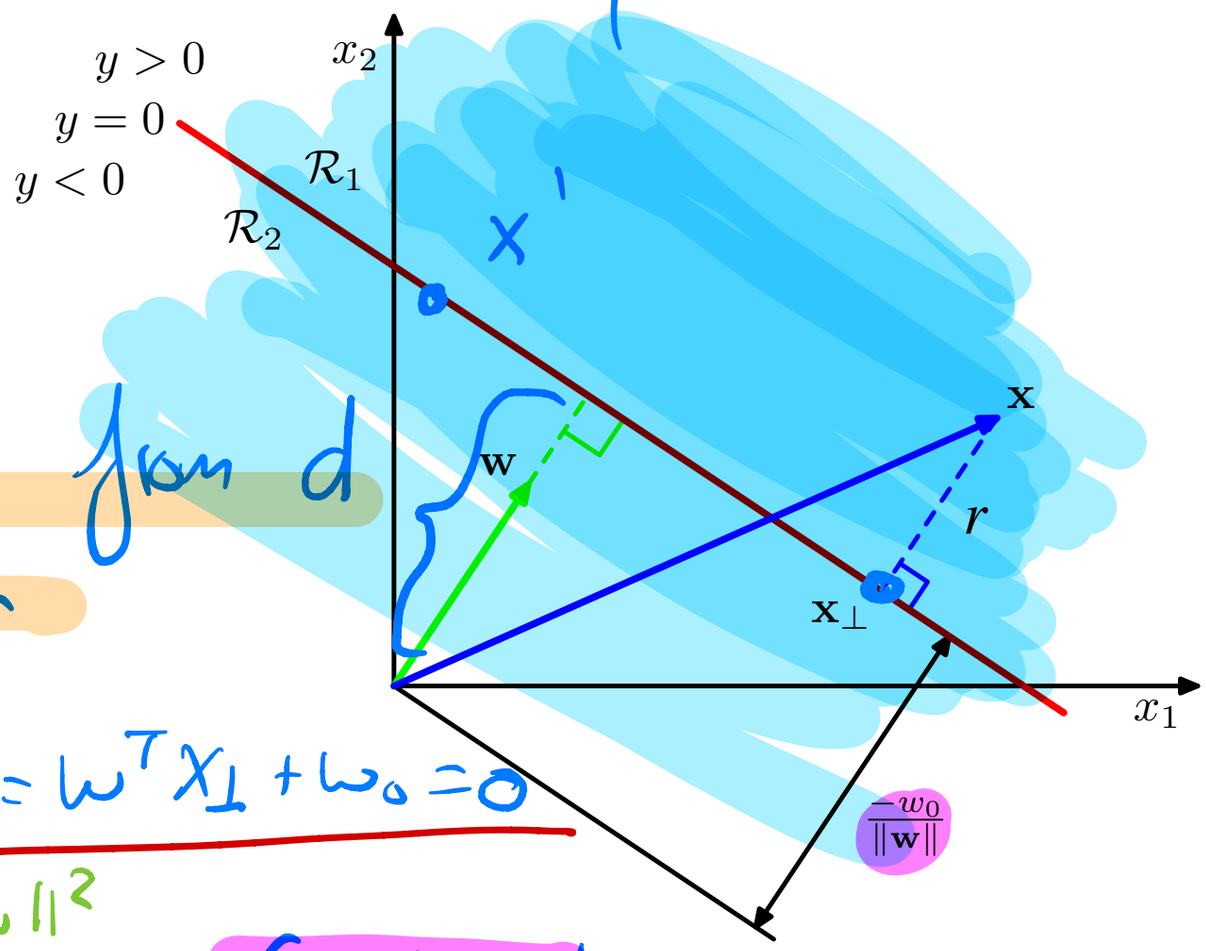
$$\mathbf{w}^T\mathbf{x}' + w_0 = 0 \quad \Leftrightarrow \quad \mathbf{w}^T\mathbf{x}' = -w_0$$

normal vector

- Normal (signed) distance d from origin to decision surface

$$d = \frac{\mathbf{w}^T\mathbf{x}'}{\|\mathbf{w}\|} = \frac{-w_0}{\|w\|}$$

$y(x)$ increases in the direction of w



⇒ w_0 shifts the d.b. away from d the origin

- Normal (signed) distances r from general $\mathbf{x}$ to decision surface:

Let $\mathbf{x} = \mathbf{x}_\perp + r\frac{\mathbf{w}}{\|\mathbf{w}\|}$

$y(x_\perp) = w^T x_\perp + w_0 = 0$

Then. $y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 = \mathbf{w}^T\mathbf{x}_\perp + r\frac{\mathbf{w}^T\mathbf{w}}{\|\mathbf{w}\|} + w_0 = r\cdot\|w\|$

$\mathbf{w}^T\mathbf{w} = \|w\|^2$

$y(x_\perp) = 0$

⇒ Distance r of a point to the d.b. is $r = \frac{y(x)}{\|w\|}$

Figure: 2-class decision boundary (Bishop 4.1)

So the distance of a point $\underline{x}$ to the decision boundary is proportional to $y(\underline{x})$

$$r = \frac{y(x)}{\|w\|}$$

Discriminant Functions: Multiple Classes

- K-class discriminant

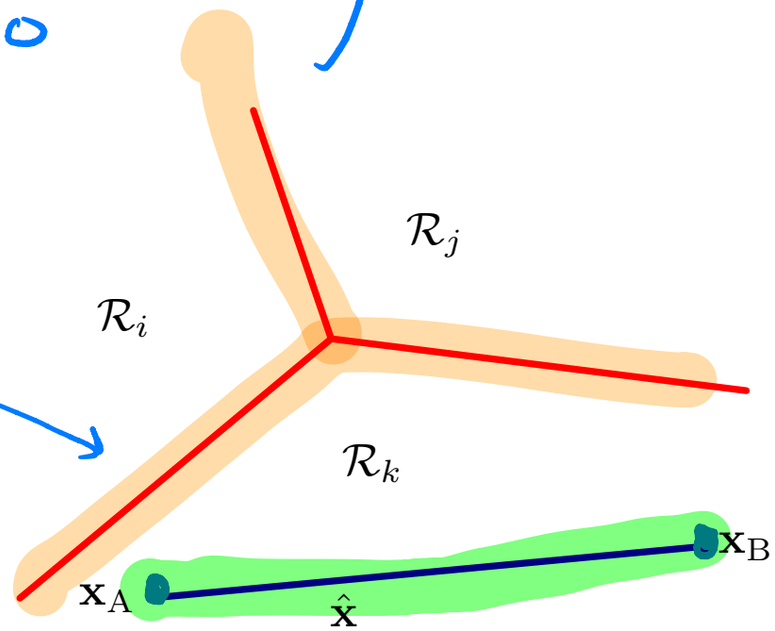
$$y_k(\mathbf{x}) = \mathbf{w}_k^T \mathbf{x} + w_{k0}$$

- Assign $\mathbf{x}$ to C_k if $\forall_{j \neq k}: \quad y_k(\mathbf{x}) > y_j(\mathbf{x})$

- Decision boundary between R_k and R_j:

$$y_k(\mathbf{x}) = y_j(\mathbf{x})$$

- Decision regions (for GLM) are convex



University of Amsterdam
Informatics Institute

AMLAB
Amsterdam
Machine Learning Lab

Machine Learning 1

Lecture 6.4 - Supervised Learning
Classification? - Discriminative Models - Least Squares Regression

Erik Bekkers

(Bishop 4.1.3)

Slide credits: Patrick Forré and Rianne van den Berg

Image credit: Kirillm | Getty Images

Least Squares for Classification

- Each class C_k has its own linear model:

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

- Shorter notation: $y(\mathbf{x}) = \tilde{\mathbf{W}}^T \tilde{\mathbf{x}}$

  - Matrix $\tilde{\mathbf{W}}$, column k contains: $\tilde{w}_k = \text{Concat}(w_{k0}, \mathbf{w}_k) \in \mathbb{R}^M$

  - Input vector: $\tilde{\mathbf{x}} = \text{Concat}(1, \mathbf{x}) \in \mathbb{R}^M$

  - Output/prediction vector: $\mathbf{y}(\mathbf{x}) = \begin{pmatrix} y_1(\mathbf{x}) \\ y_2(\mathbf{x}) \\ \vdots \\ y_K(\mathbf{x}) \end{pmatrix} = \tilde{\mathbf{W}}^T \tilde{\mathbf{x}} \in \mathbb{R}^K$

- **Classification**: Assign $\mathbf{x}$ to class C_K if $k = \underset{j}{\text{argmax}} \; y_j(\mathbf{x})$

Least Squares for Classification (II)

- Data set: $N \times (D+1)$ data matrix $N \times K$ target matrix

$$\tilde{X} = \begin{pmatrix} \tilde{\mathbf{x}}_1^T \\ \vdots \\ \tilde{\mathbf{x}}_N^T \end{pmatrix} \qquad T = \begin{pmatrix} \mathbf{t}_1^T \\ \vdots \\ \mathbf{t}_N^T \end{pmatrix}$$

- Use sum-of-squares regression error function

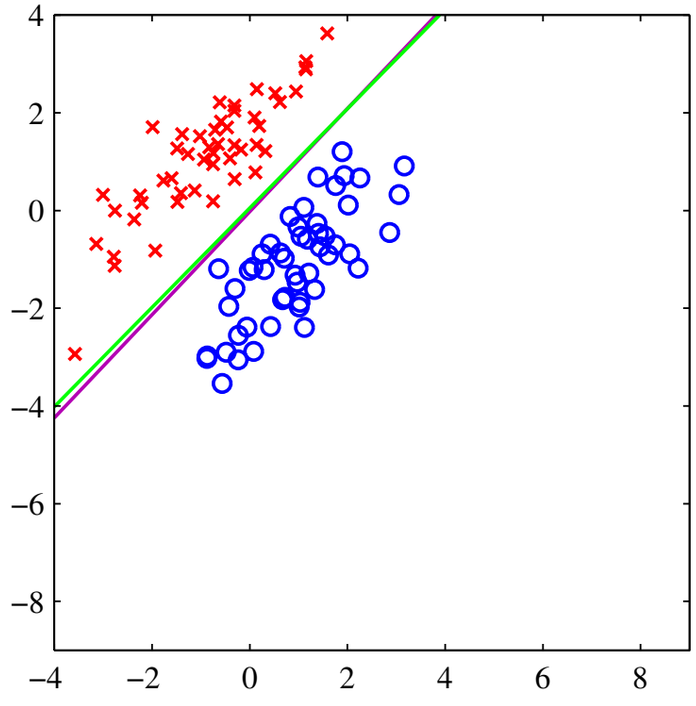
$$E_D(\tilde{\mathbf{W}}) = \frac{1}{2} \mathrm{Tr}\left[(\tilde{\mathbf{X}}\tilde{\mathbf{W}} - \mathbf{T})^T(\tilde{\mathbf{X}}\tilde{\mathbf{W}} - \mathbf{T})\right] =$$

- Minimize $E_D(\tilde{\mathbf{W}})$ as a function of $\tilde{\mathbf{W}}$:

- Solution: $\tilde{\mathbf{W}}_{\mathrm{LS}} = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\mathbf{T} = \tilde{\mathbf{X}}^\dagger\mathbf{T}$

- Discriminant function: $\mathbf{y}_{\mathrm{LS}}(\mathbf{x}) =$

Least Squares for Classification: Problems



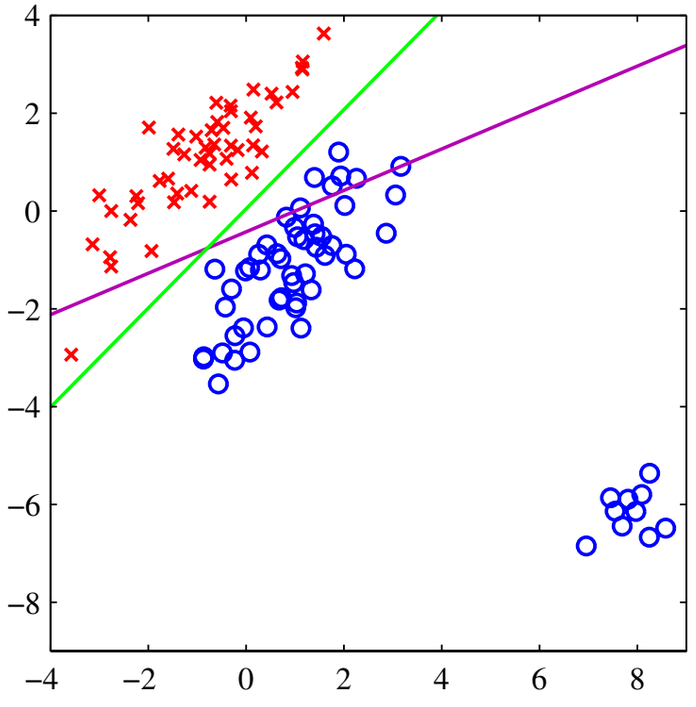


Figure: least squares is very sensitive to outliers (Bishop 4.4)

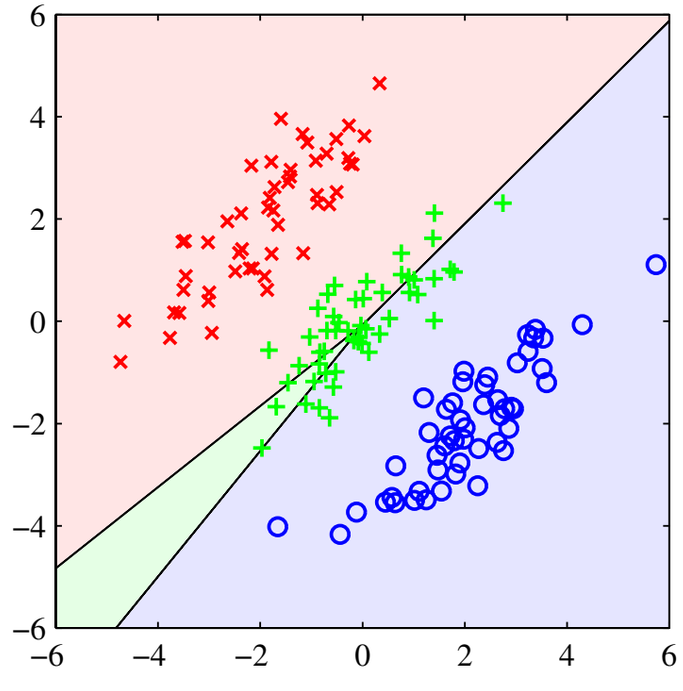


Figure: masking for least squares for K>2 (Bishop 4.5)

1. The decision boundaries are very sensitive to outliers
2. For $K > 2$ some decision regions can become very small or are even completely ignored
3. The components of $\mathbf{y}_{\mathrm{LS}}(\mathbf{x})$ are not probabilities!
   - $y_k(\mathbf{x})$
   - if $\sum_{k=1}^{K} t_k = 1$

University of Amsterdam
Informatics Institute

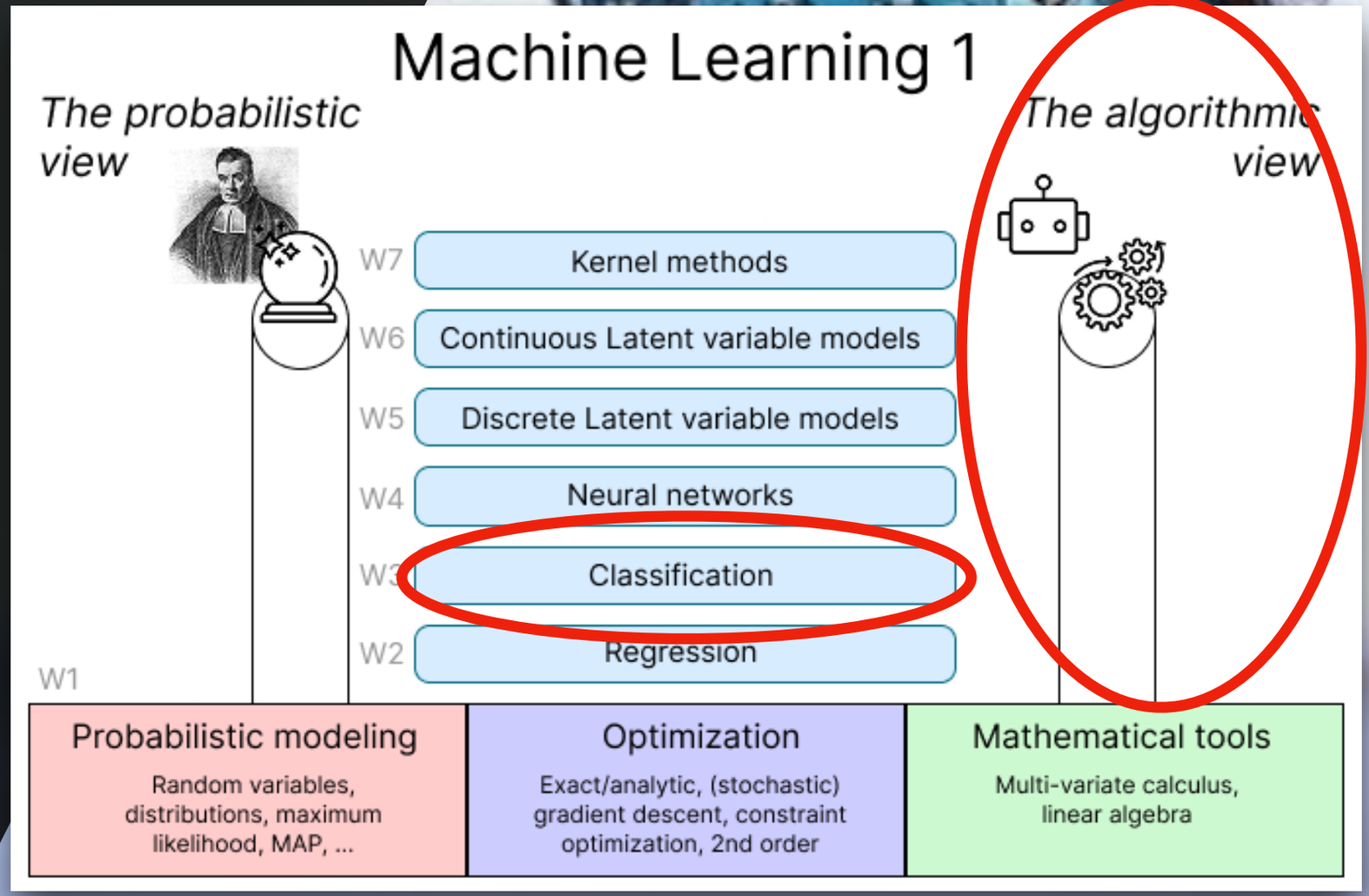
AMLAB
Amsterdam
Machine Learning Lab

Machine Learning 1

Lecture 6.5 - Supervised Learning
Classification - Discriminative Models -
The Perceptron

Erik Bekkers

(Bishop 4.1.7)



Slide credits: Patrick Forré and Rianne van den Berg

Image credit: Kirillm | Getty Images

The Perceptron Algorithm

- Input: $\mathbf{x} \in \mathbb{R}^D$
- Targets: $t \in \{C_1, C_2\}$
- Prediction: $y(\mathbf{x}) = f(\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}))$, with $f(a) = \begin{cases} 1 & , a \geq 0 \\ -1 & , a < 0 \end{cases}$
- Class decisions: assign $\mathbf{x}$ to class C_1 if

  (and to C_2 if)
- For correct classification: find $\mathbf{w}$ such that for all $(\mathbf{x}_n, t_n)$:
- Perceptron criterion: $E_P(\mathbf{w}) = -\sum_{n \in \mathcal{M}} \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_n) t_n$

  with $\mathcal{M} =$

Perceptron: Stochastic Gradient Descent

- $$E_P(\mathbf{w}) = -\sum_{n \in \mathcal{M}} \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_n) t_n$$

$$= -\sum_{n \in \mathcal{M}} E_n(\mathbf{w})$$

- Stochastic Gradient Descent (SGD). For each misclassified $\mathbf{x}_n$:

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \nabla E_n(\mathbf{w})^T$$

$$=$$

- If $\mathbf{X}$ is linearly separable, then perceptron SGD will converge

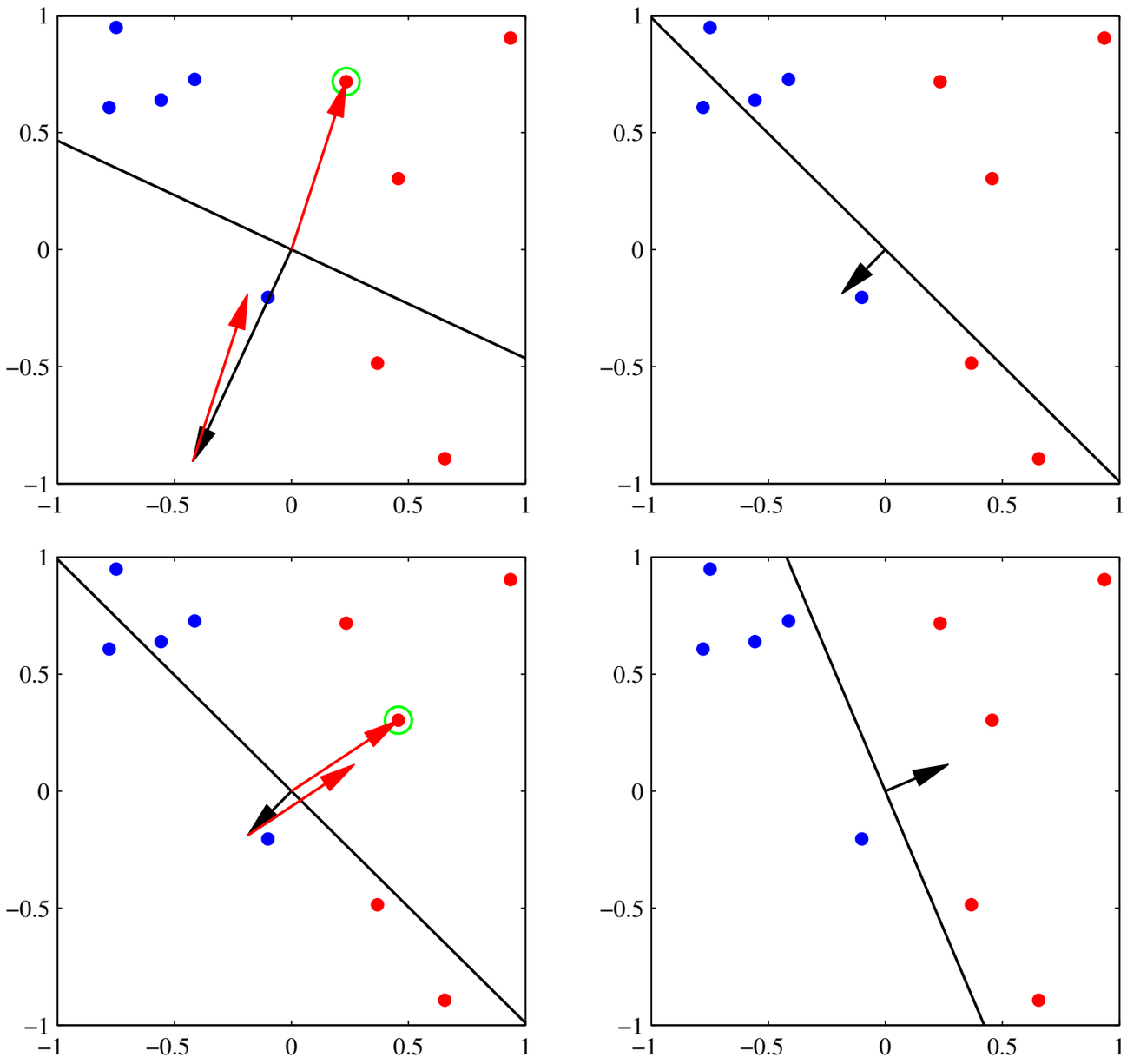


Figure: SGD for perceptron criterion (Bishop 4.7)
for $\mathbf{x}_n$ in C_1: add $\boldsymbol{\phi}(\mathbf{x}_n)$ to $\mathbf{w}$
for $\mathbf{x}_n$ in C_2: subtract $\boldsymbol{\phi}(\mathbf{x}_n)$ from $\mathbf{w}$.

Problems: Perceptron

- Perceptron only works for 2 classes
- There might be many solutions depending on the initialization of $\mathbf{w}$ and on the order in which data is presented in SGD
- If dataset is not linearly separable, the perceptron algorithm will not converge.
- Based on linear combination of fixed basis functions.