

Basics of Machine Learning

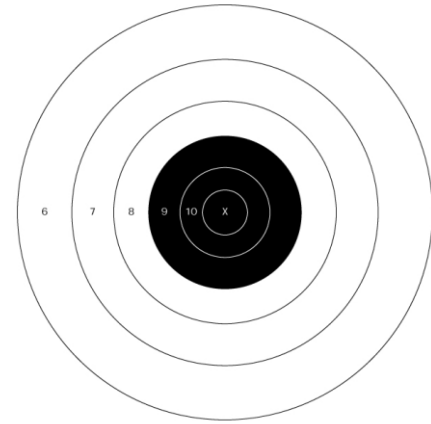
Dmitry Ryabokon, github.com/dryabokon





Lesson 10

Parametrical ML methods



Supervised Learning

Summary

- Naive Bayesian classifier
- Gaussian classifier

Differentiation	Parametric Model	Non-parametric Model
Features	A finite number of parameters to predict new data	Unbounded number of parameters to predict new data.
Algorithm	Logistic regression Linear discriminant analysis Perceptron Naive Bayes	K-nearest neighbours Decision trees (E.g. CART and C4.5) Support vector machines
Benefits	Easy to use Quick in functioning Less data	Flexibility Power Performance
Limitations	Constrained Limited complexity Poor fit	More data Slower Overfit

Naive Bayesian classifier



Naive Bayesian classifier

Advantages

- Very simple, easy to implement and fast.
- If the NB conditional independence assumption holds, then it will converge quicker than discriminative models like logistic regression.
- Even if the NB assumption doesn't hold, it works great in practice.
- Need less training data.
- Highly scalable. It scales linearly with the number of predictors and data points.
- Can be used for both binary and multi-class classification problems.
- Can make probabilistic predictions.
- Handles continuous and discrete data.
- Not sensitive to irrelevant features.

Naive Bayesian classifier

The problem

$\bar{x} = (x_1, x_2, \dots, x_N)$ feature

$$p(\bar{x}|k) = p(x_1|k) \cdot p(x_2|k) \cdot \dots \cdot p(x_N|k)$$

$k \in K = \{1, 2\}$ few states are possible

$$\frac{p(\bar{x}|k=1)}{p(\bar{x}|k=2)} \geq \theta$$

decision strategy

Naive Bayesian classifier

Example

$$p(x_2 | \text{red}) \quad p(x_2 | \text{blue})$$

$$1/6$$

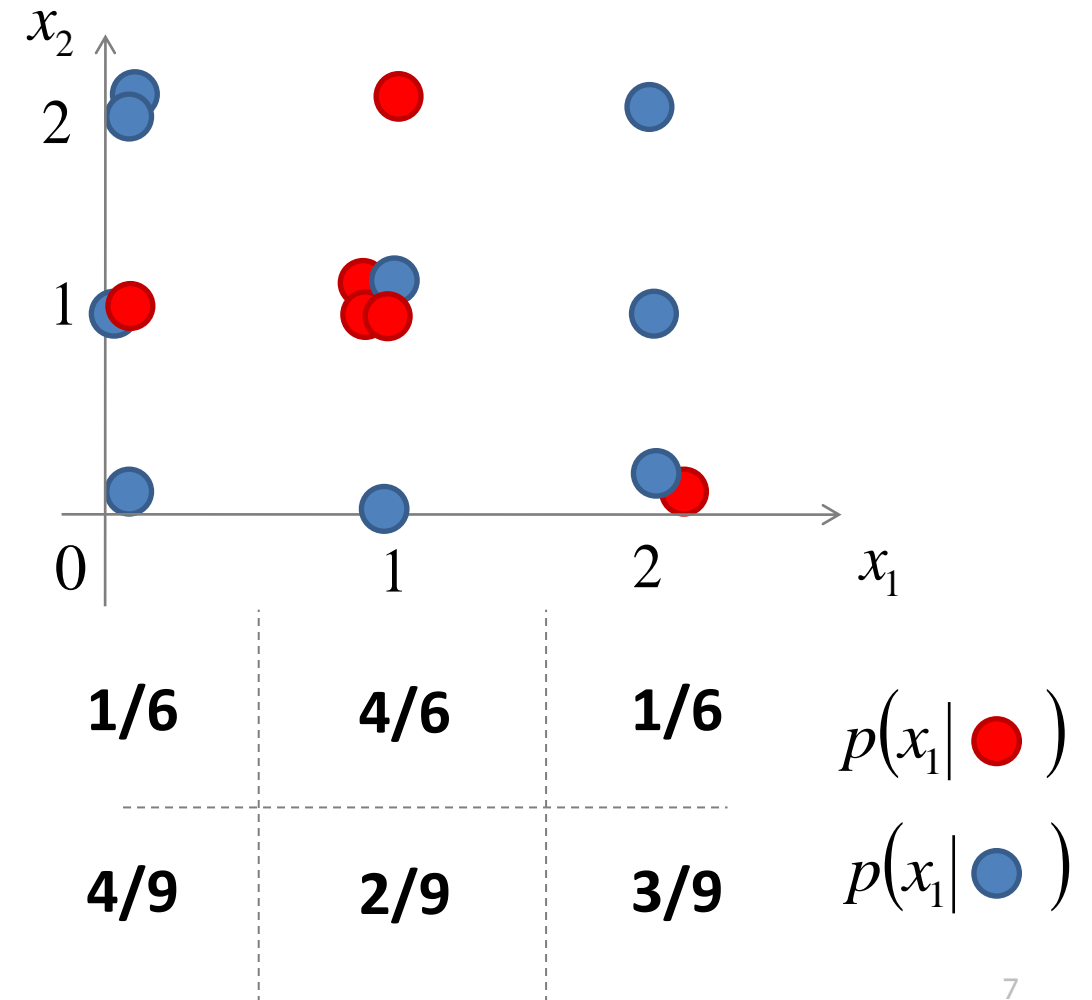
$$3/9$$

$$4/6$$

$$3/9$$

$$1/6$$

$$3/9$$



Naive Bayesian classifier

Example

$$p(x_2 | \bullet)$$

$$1/6$$

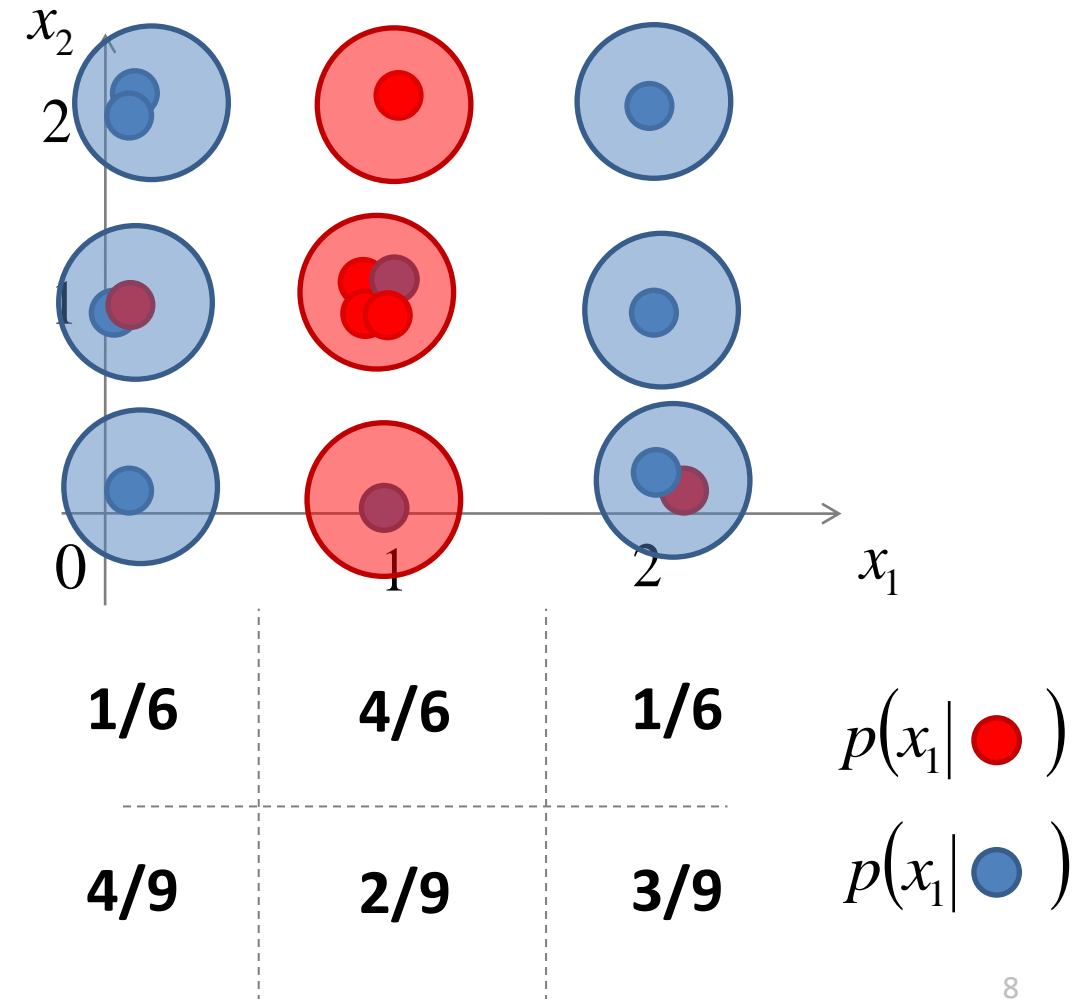
$$3/9$$

$$4/6$$

$$3/9$$

$$1/6$$

$$3/9$$



Gaussian classifier



Gaussian classifier

Definitions

$$\bar{x} = (x_1, x_2, \dots, x_N) \text{ feature}$$

$$p(\bar{x}|k) \cong \exp\left(-0.5 \cdot \sum_{i=1}^n \sum_{j=1}^n a_{i,j}^{[k]} (x_i - \mu_i^{[k]})(x_j - \mu_j^{[k]})\right)$$

$$A^{[k]} = (B^{[k]})^{-1}$$

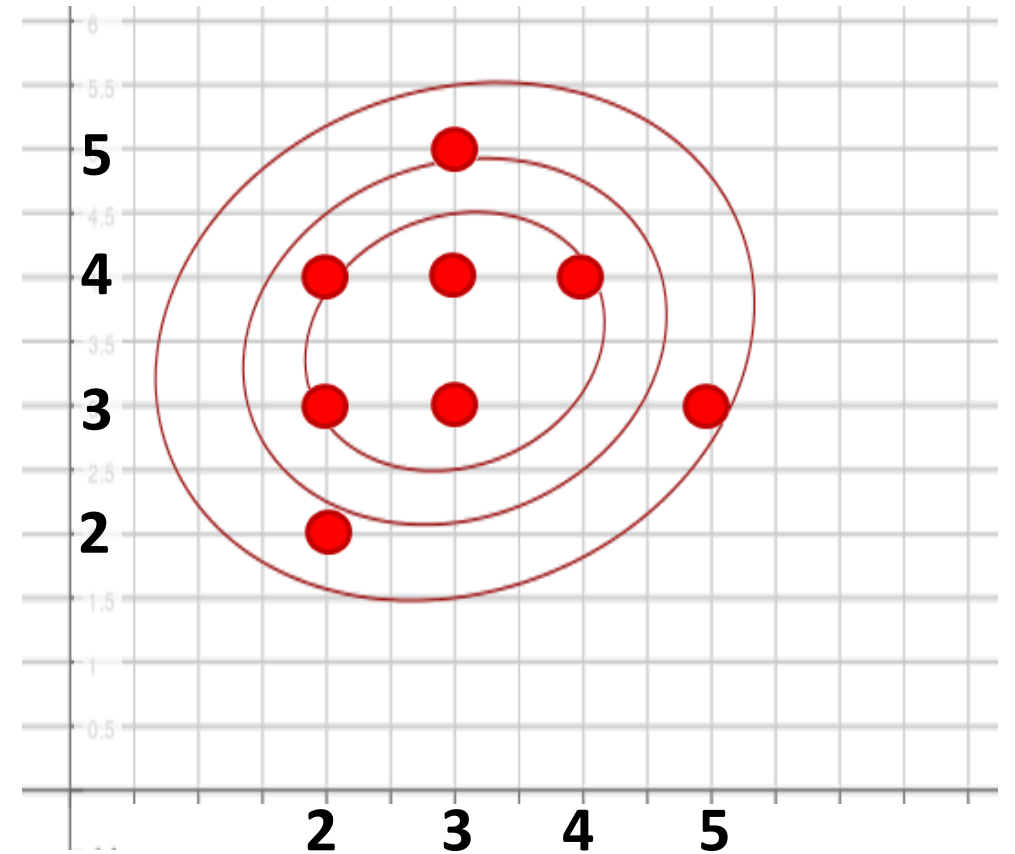
$$\mu_i^{[k]} = M.O.(x_i|k)$$

$$b_{ij}^{[k]} = M.O.(x_i - \mu_i^{[k]}) \cdot (x_j - \mu_j^{[k]})$$

Gaussian classifier

Example

$$p(\bar{x}|\bullet) \cong \exp\left(-0.5 \cdot \sum_{i=1}^n \sum_{j=1}^n a_{i,j} (x_i - \mu_i)(x_j - \mu_j)\right)$$



Gaussian classifier

$$\mu_1 = MO(x_1 | \bullet) = 3$$

$$\mu_2 = MO(x_2 | \bullet) = 3,5$$

$$b_{11} = MO(x_1 - \mu_1) \cdot (x_1 - \mu_1) = 1$$

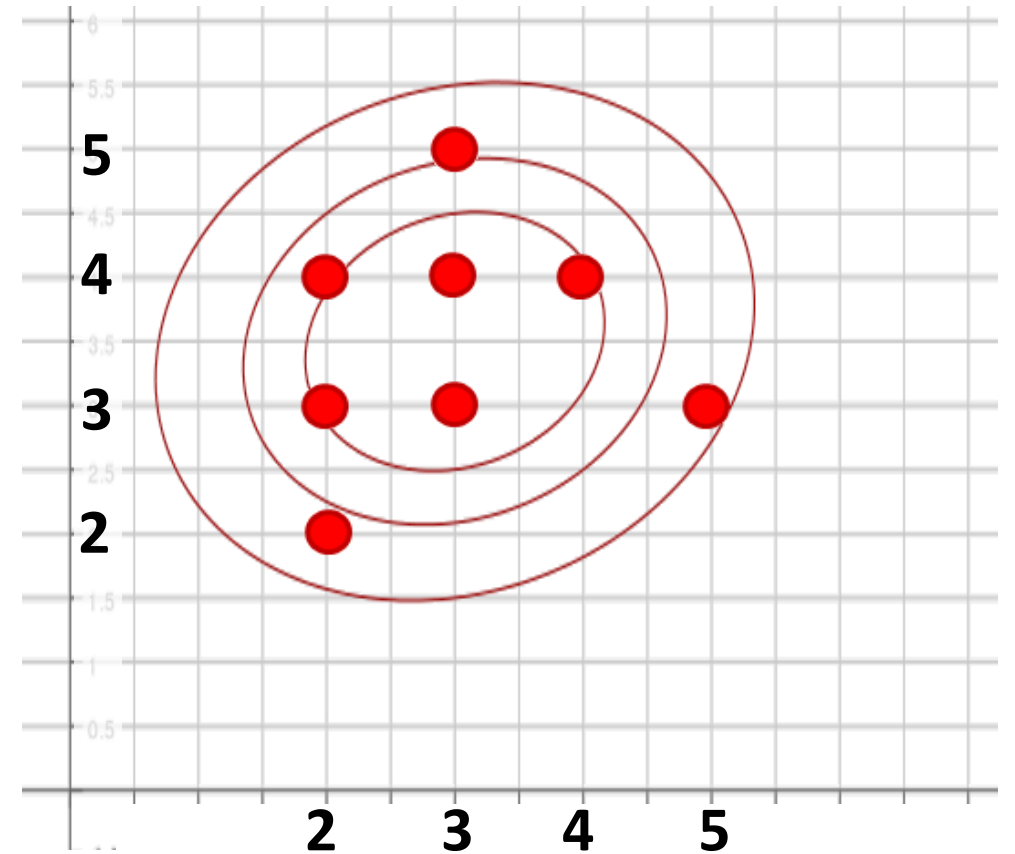
(1+1+1+0+0+0+1+4)/8

$$b_{22} = MO(x_2 - \mu_2) \cdot (x_2 - \mu_2) = 3/4$$

$$b_{12} = b_{21} = MO(x_1 - \mu_1) \cdot (x_2 - \mu_2) = 1/8$$

$$B = \begin{bmatrix} 1 & 1/8 \\ 1/8 & 3/4 \end{bmatrix}^{-1} = \frac{64}{47} \cdot \begin{bmatrix} 3/4 & -1/8 \\ -1/8 & 1 \end{bmatrix} = A$$

$$p(\bar{x} | \bullet) \cong \exp \left(\frac{64}{47} \cdot \left(-\frac{3}{4} \cdot (x_1 - 3)^2 + \frac{1}{4} (x_1 - 3)(x_2 - 3,5) - \frac{1}{1} (x_2 - 3,5)^2 \right) \right)$$



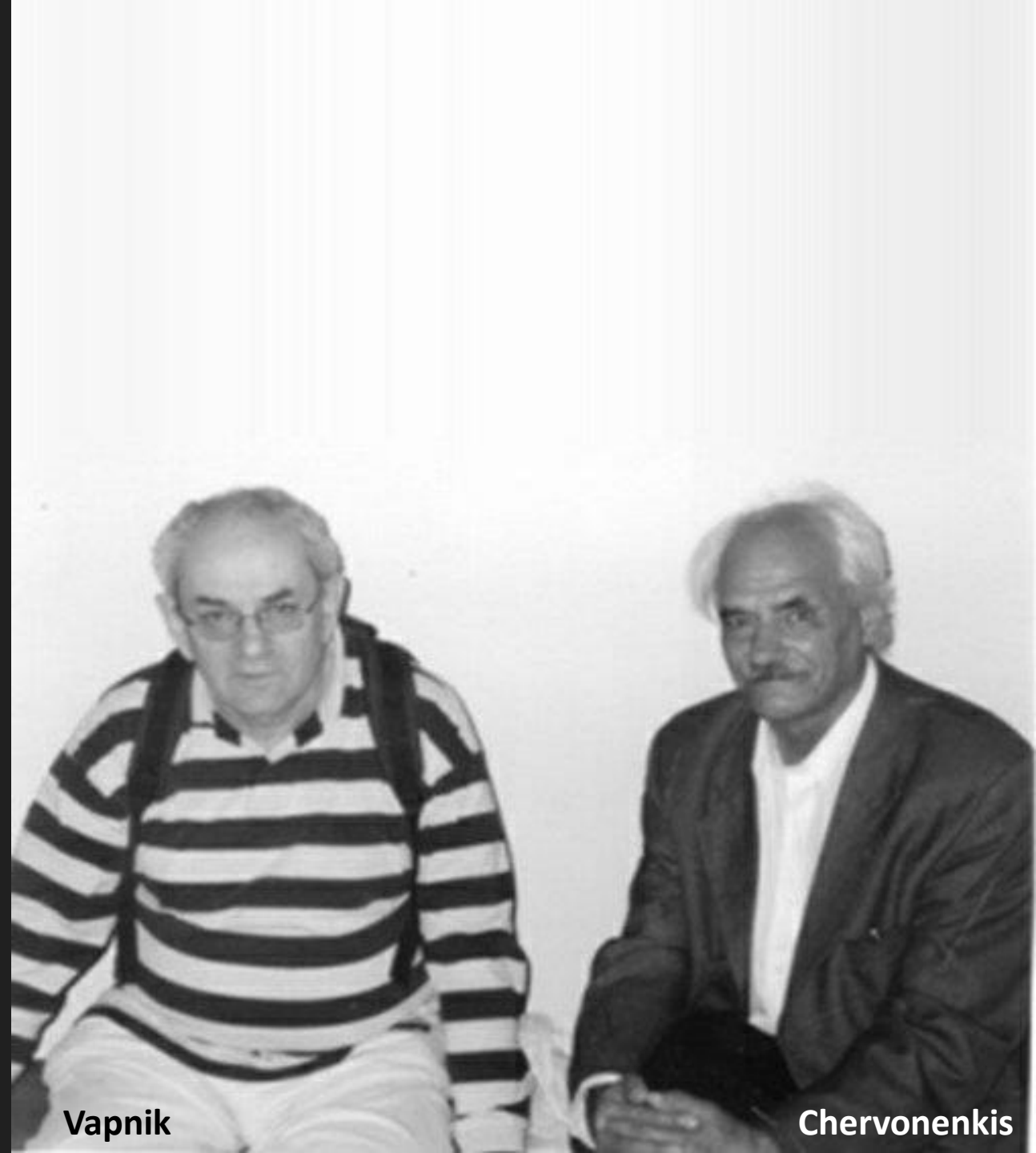
Gaussian classifier

Decision strategy

$$\log \frac{p(\bar{x}|k=1)}{p(\bar{x}|k=2)} = \frac{\sum_{i=1}^n \sum_{j=1}^n a_{i,j}^{[1]} (x_i - \mu_i^{[1]})(x_j - \mu_j^{[1]})}{\sum_{i=1}^n \sum_{j=1}^n a_{i,j}^{[2]} (x_i - \mu_i^{[2]})(x_j - \mu_j^{[2]})}$$

$$\sum_i \sum_j \alpha_{ij} \cdot x_i x_j + \sum_i \beta_i \cdot x_i \begin{matrix} 1 \\ \geq \\ \leq \\ 2 \end{matrix} \theta$$

Linear discrimination



Vapnik

Chervonenkis

Linear discrimination

Perceptron

$$\text{input} \left\{ \begin{array}{l} X = \{x_1, x_2, \dots, x_r\} \\ X' = \{x'_1, x'_2, \dots, x'_s\} \end{array} \right.$$

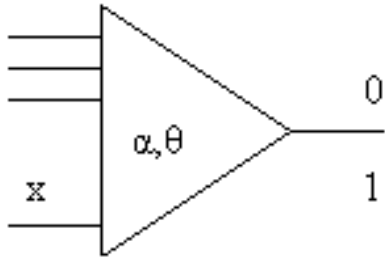
$$\begin{array}{l} \text{output} \\ \alpha \in R^n \\ \theta \end{array} \left\{ \begin{array}{l} \forall x \in X \quad (x, \alpha) = \sum_{i=1}^n x_i \cdot \alpha_i > \theta \\ \forall x' \in X' \quad (x', \alpha) = \sum_{i=1}^n x'_i \cdot \alpha_i < \theta \end{array} \right.$$

Rosenblatt



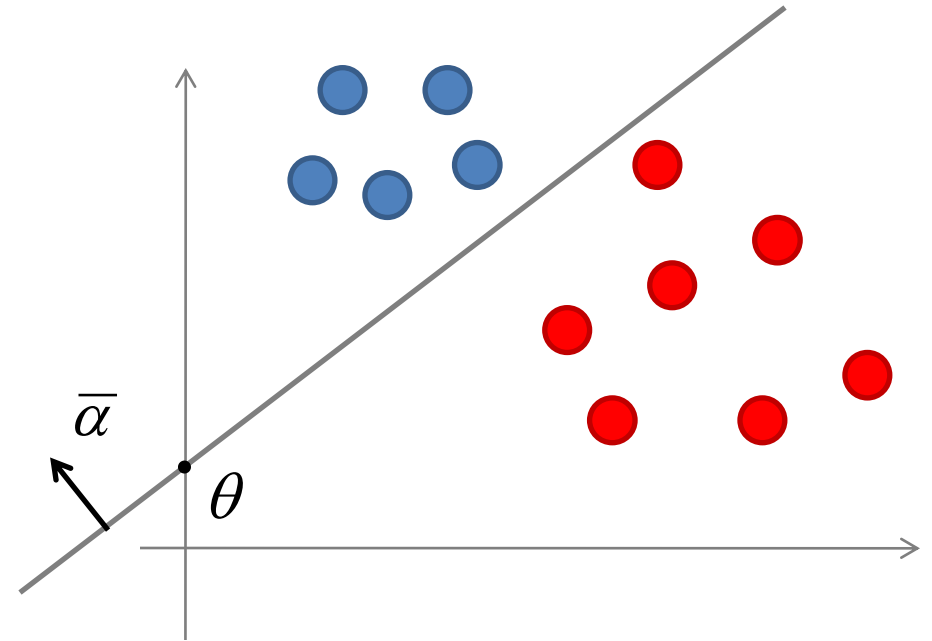
Linear discrimination

Perceptron



$$\text{output} \left\{ \begin{array}{l} \forall x \in X \quad (x, \alpha) = \sum_{i=1}^n x_i \cdot \alpha_i > \theta \\ \forall x' \in X' \quad (x', \alpha) = \sum_{i=1}^n x'_i \cdot \alpha_i < \theta \end{array} \right.$$

$\alpha \in R^n$
 θ



Linear discrimination

Perceptron

$$\left\{ \begin{array}{l} \forall x \in X \quad \sum_{i=1}^n x_i \cdot \alpha_i > \theta \\ \forall x' \in X' \quad \sum_{i=1}^n x'_i \cdot \alpha_i < \theta \end{array} \right. = \left\{ \begin{array}{l} \forall x \in X \quad \sum_{i=1}^n x_i \cdot \alpha_i + 1 \cdot \alpha_{n+1} > 0 \\ \forall x' \in X' \quad \sum_{i=1}^n x'_i \cdot \alpha_i + 1 \cdot \alpha_{n+1} < 0 \end{array} \right.$$

Linear discrimination

Perceptron: tuning

$$t = 0$$

$$\alpha_t = 0$$

while (sets are not separated by hyperplane)

{

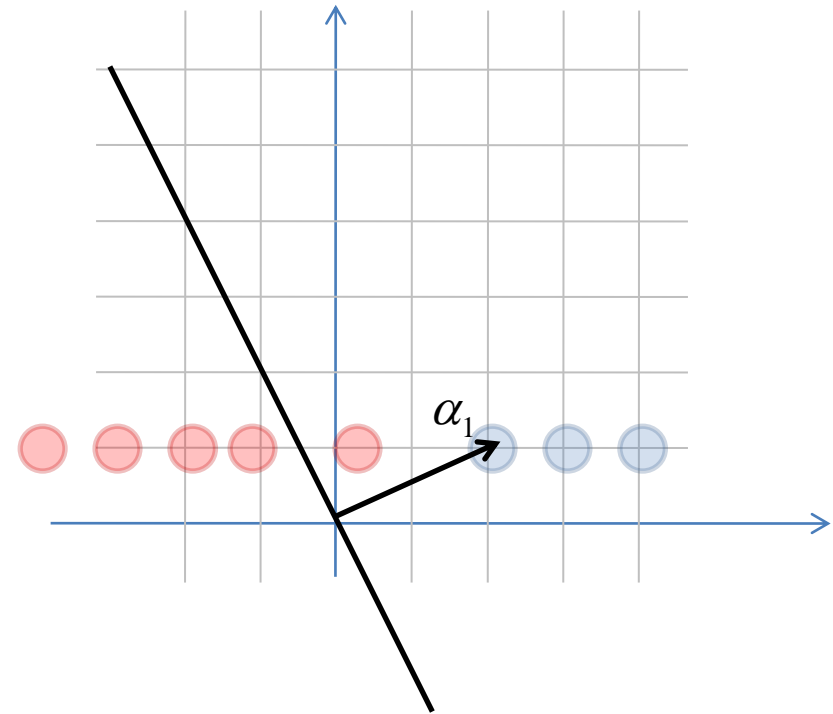
$$\text{if } (\exists x \in X | (x, \alpha_t) < 0) \quad \alpha_{t+1} = \alpha_t + x;$$

$$\text{if } (\exists x' \in X' | (x', \alpha_t) > 0) \quad \alpha_{t+1} = \alpha_t - x';$$

}

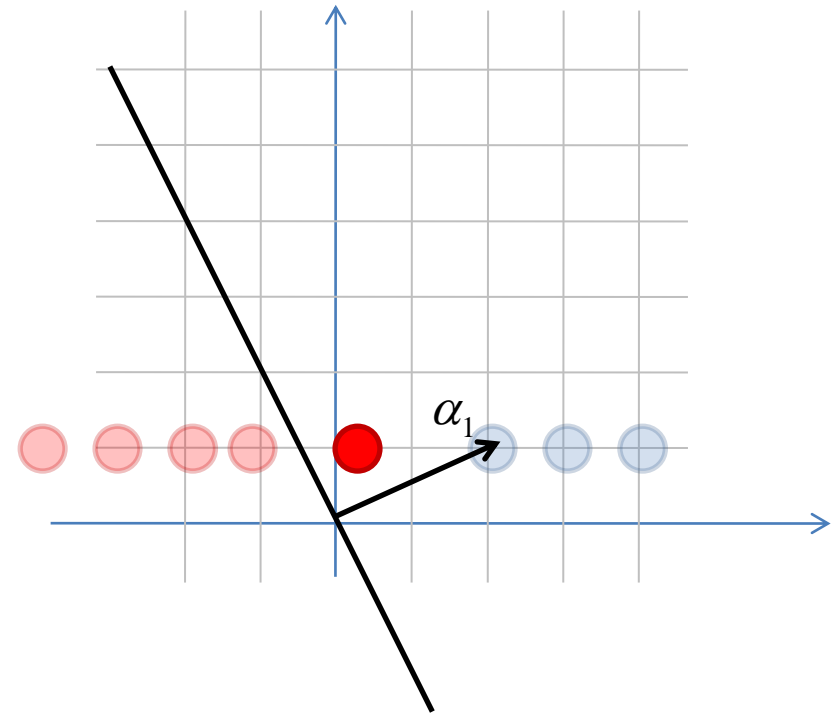
Linear discrimination

Perceptron: tuning



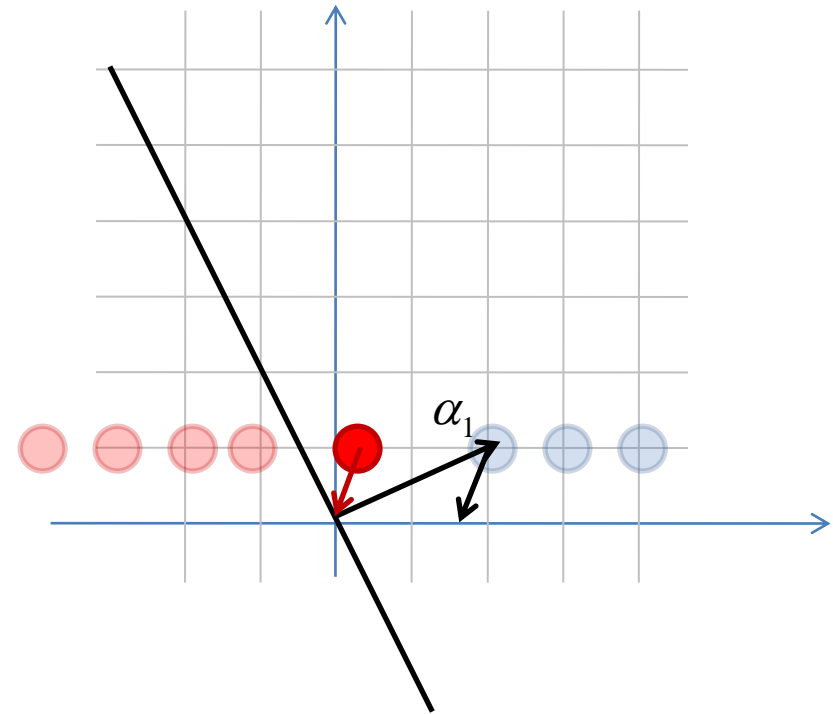
Linear discrimination

Perceptron: tuning



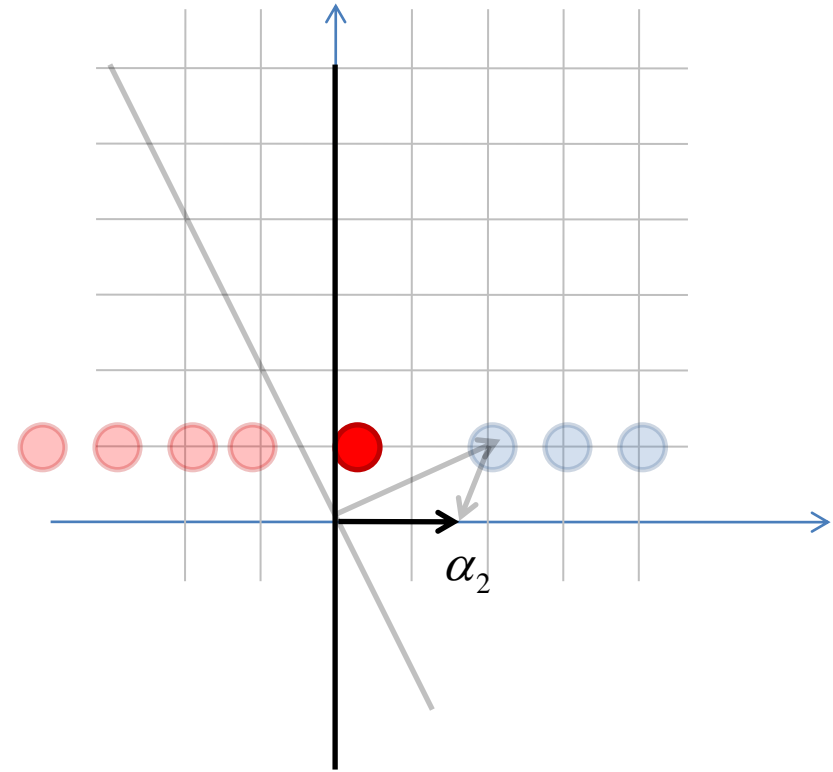
Linear discrimination

Perceptron: tuning



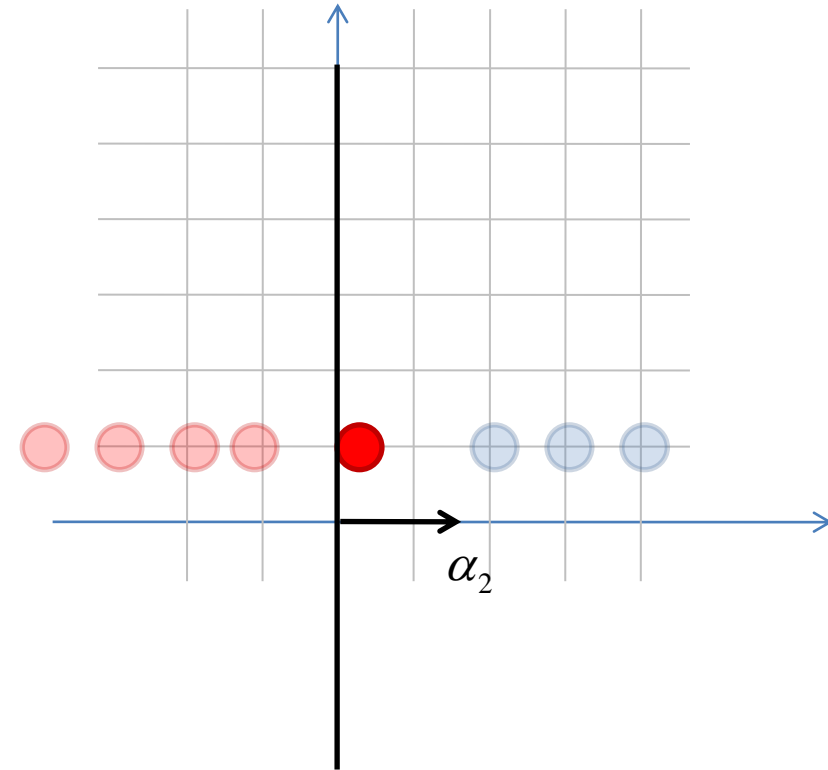
Linear discrimination

Perceptron: tuning



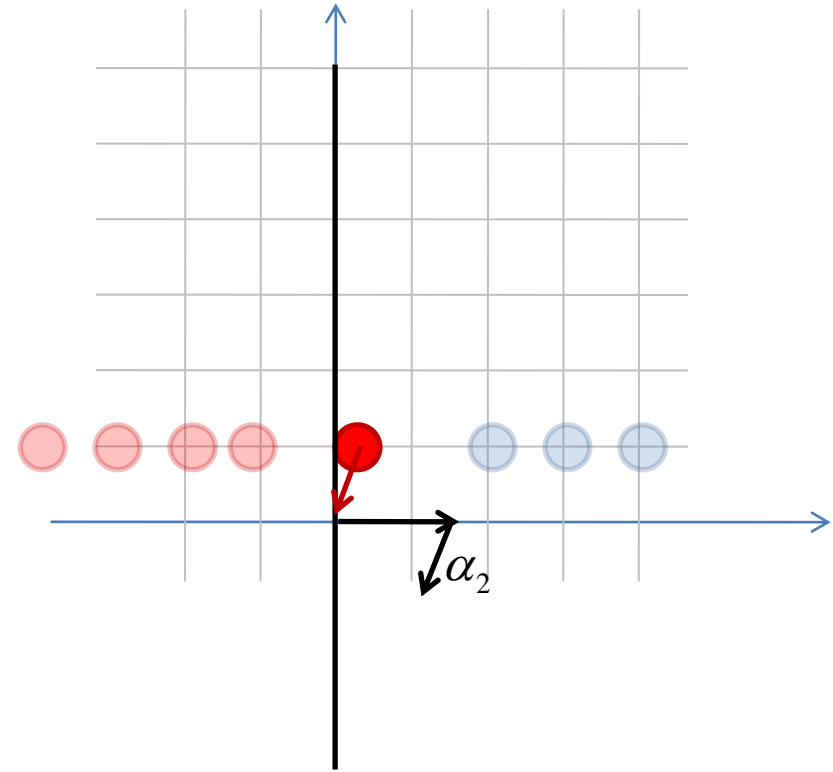
Linear discrimination

Perceptron: tuning



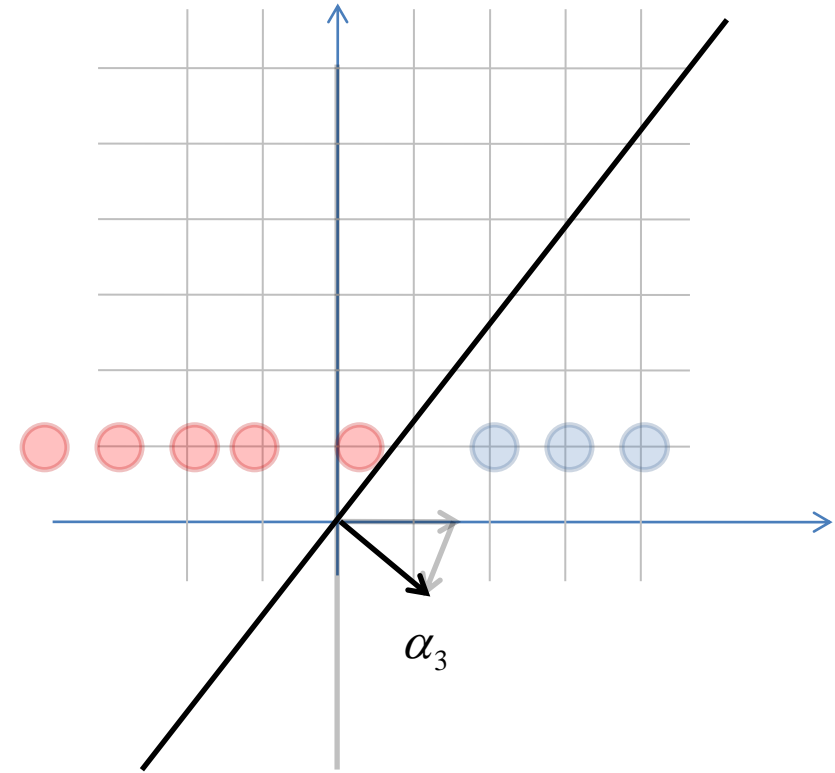
Linear discrimination

Perceptron: tuning



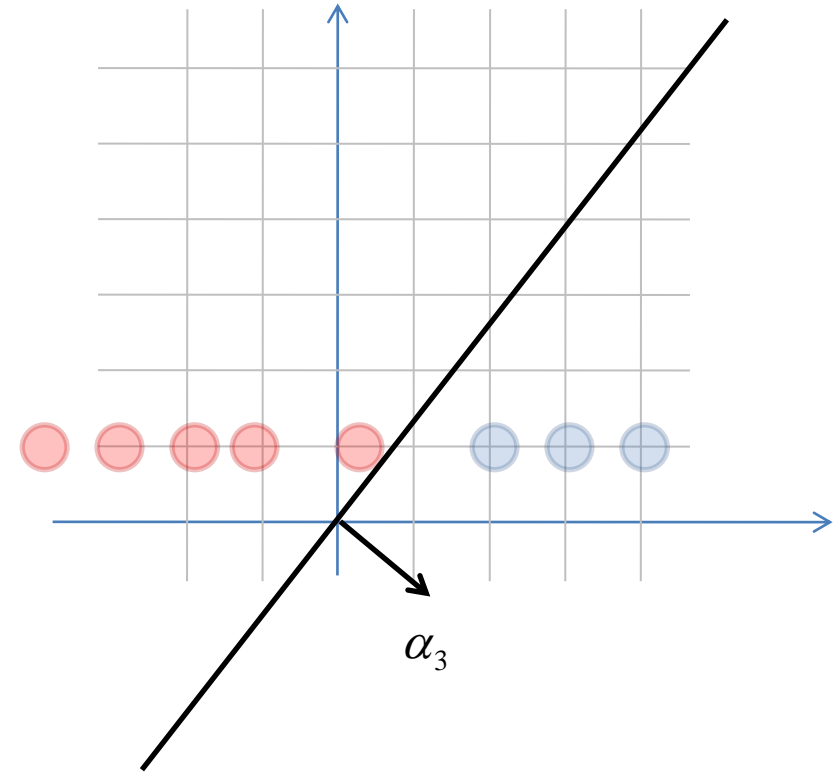
Linear discrimination

Perceptron: tuning



Linear discrimination

Perceptron: tuning



Linear discrimination

Perceptron: convergence

$$\max_{\substack{x \in X \\ x \in X'}} \|x\| = D \quad \|\alpha^*\| = 1$$

$$\begin{cases} \forall x \in X & (x, \alpha^*) \geq \varepsilon > 0 \\ \forall x \in X' & (x', \alpha^*) \leq -\varepsilon < 0 \end{cases}$$

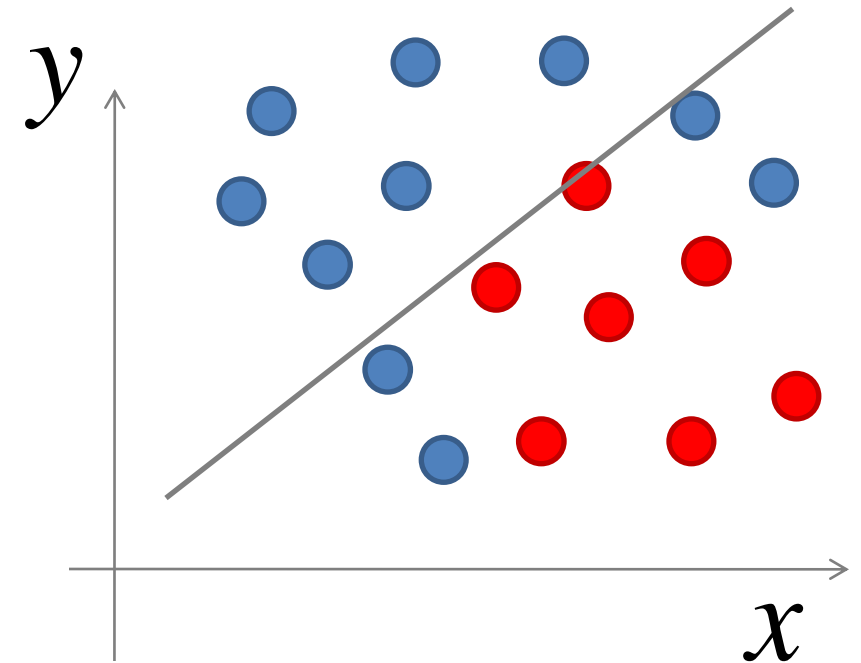
$$t \leq \frac{D^2}{\varepsilon}$$

Linear discrimination

Transformation of the feature space

$$Ax + By \gtrless C$$

a line in \mathbb{R}^2



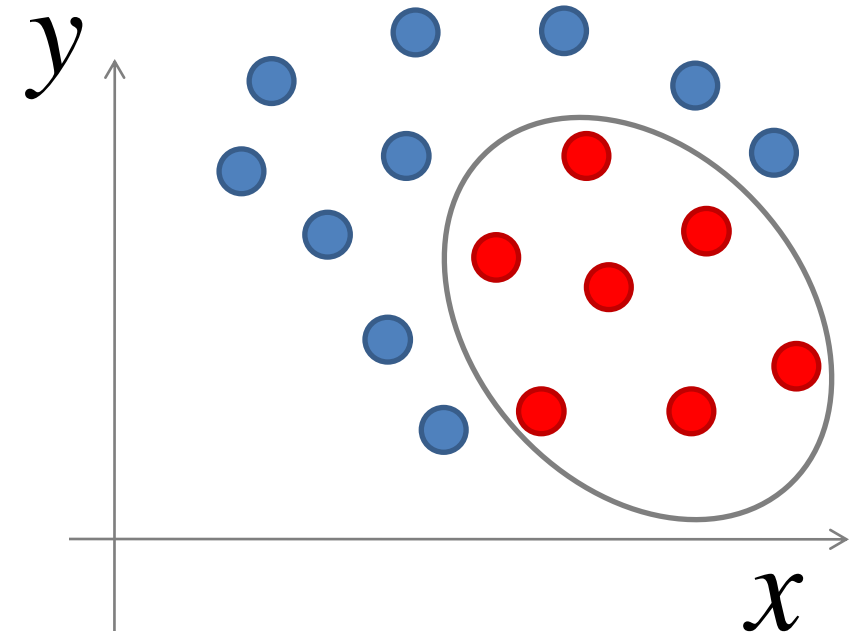
Linear discrimination

Transformation of the feature space

$$Ax^2 + Bxy + Cy^2 + Dx + Ey \gtrless F$$

hyperplane in \mathbb{R}^5

elliptic curve in \mathbb{R}^2



Linear discrimination

SVM: support vector machine

$$\begin{cases} \forall x \in X & (x, \alpha) = \sum_{i=1}^n x_i \cdot \alpha_i > \theta \\ \forall x' \in X' & (x', \alpha) = \sum_{i=1}^n x'_i \cdot \alpha_i < \theta \end{cases}$$

$$(\alpha, \alpha) \rightarrow \min$$

Separate by the
widest band

Linear discrimination

SVM: support vector machine

$$\begin{cases} \forall x \in X & (x, \alpha) = \sum_{i=1}^n x_i \cdot \alpha_i > \theta - \xi(x) \\ \forall x' \in X' & (x', \alpha) = \sum_{i=1}^n x'_i \cdot \alpha_i < \theta + \xi(x') \end{cases}$$

penalty

$$(\alpha, \alpha) + \text{const} \cdot \sum_{X, X'} \xi(x) \rightarrow \min$$

Separate by the
widest band

