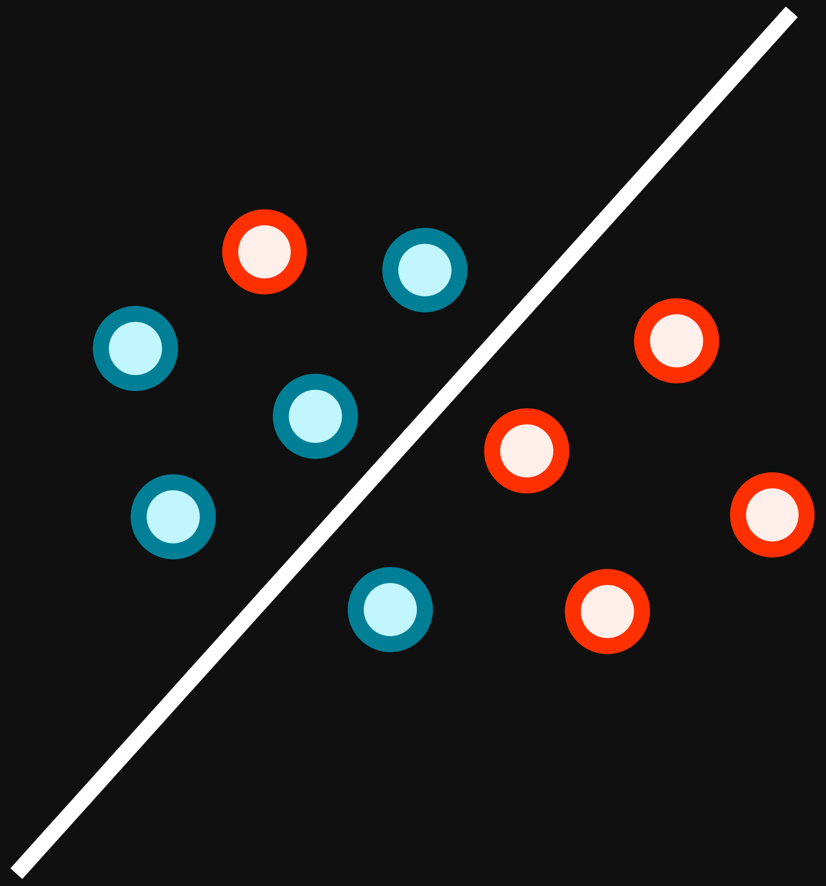
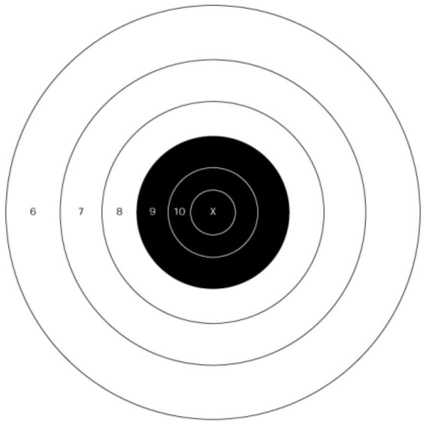
Basics of Machine Learning

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 regressionLinear discriminant analysisPerceptronNaive Bayes	K-nearest neighboursDecision trees (E.g.CART and C4.5)Support vector machines
Benefits	Easy to use Quick in functioningLess data	FlexibilityPowerPerformance
Limitations	Constrained Limited complexity Poor fit	More dataSlowerOverfit

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

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

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

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

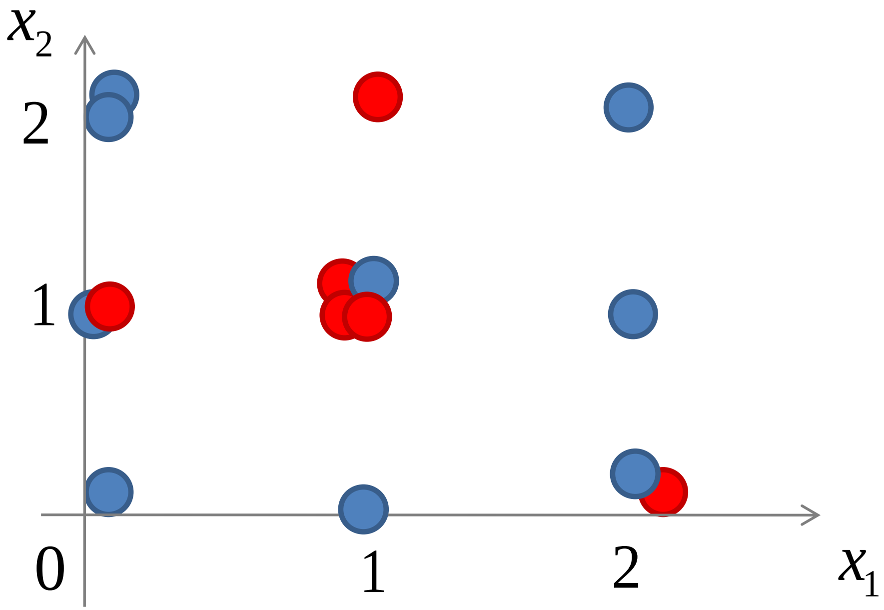
$$\frac{p(\bar{x}|k=1)}{p(\bar{x}|k=2)} \overset{1}{\underset{2}{\gtreqless}} \theta$$

decision strategy

Naive Bayesian classifier

Example

$p(x_2 \mid \bullet)$	$p(x_2 \mid \bullet)$
1/6	**3/9**
4/6	**3/9**
1/6	**3/9**

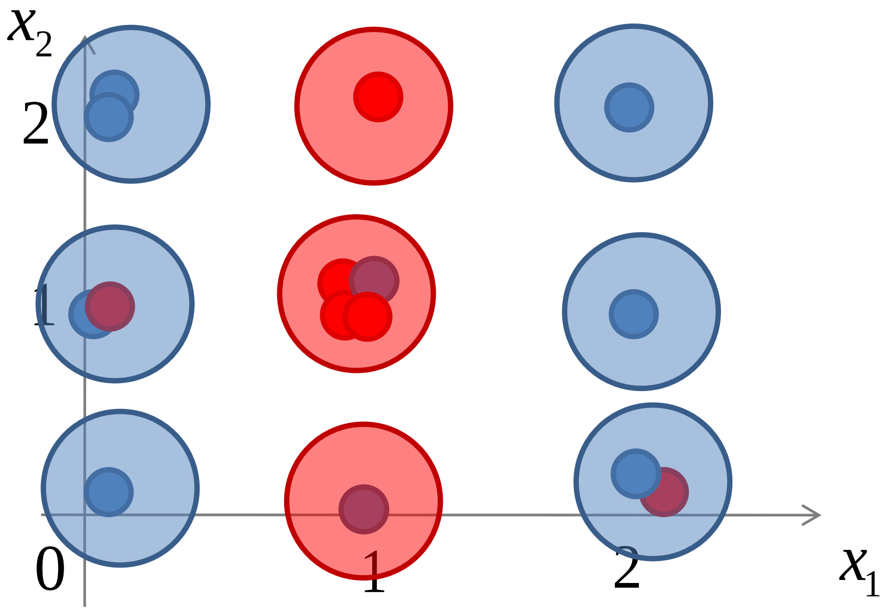


1/6	**4/6**	**1/6**	$p(x_1 \mid \bullet)$
4/9	**2/9**	**3/9**	$p(x_1 \mid \bullet)$

Naive Bayesian classifier

Example

$p(x_2 \mid ●)$	$p(x_2 \mid ●)$
1/6	3/9
4/6	3/9
1/6	3/9



1/6	4/6	1/6	$p(x_1 \mid ●)$
4/9	2/9	3/9	$p(x_1 \mid ●)$

Gaussian classifier

Gaussian classifier

Definitions

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

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

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

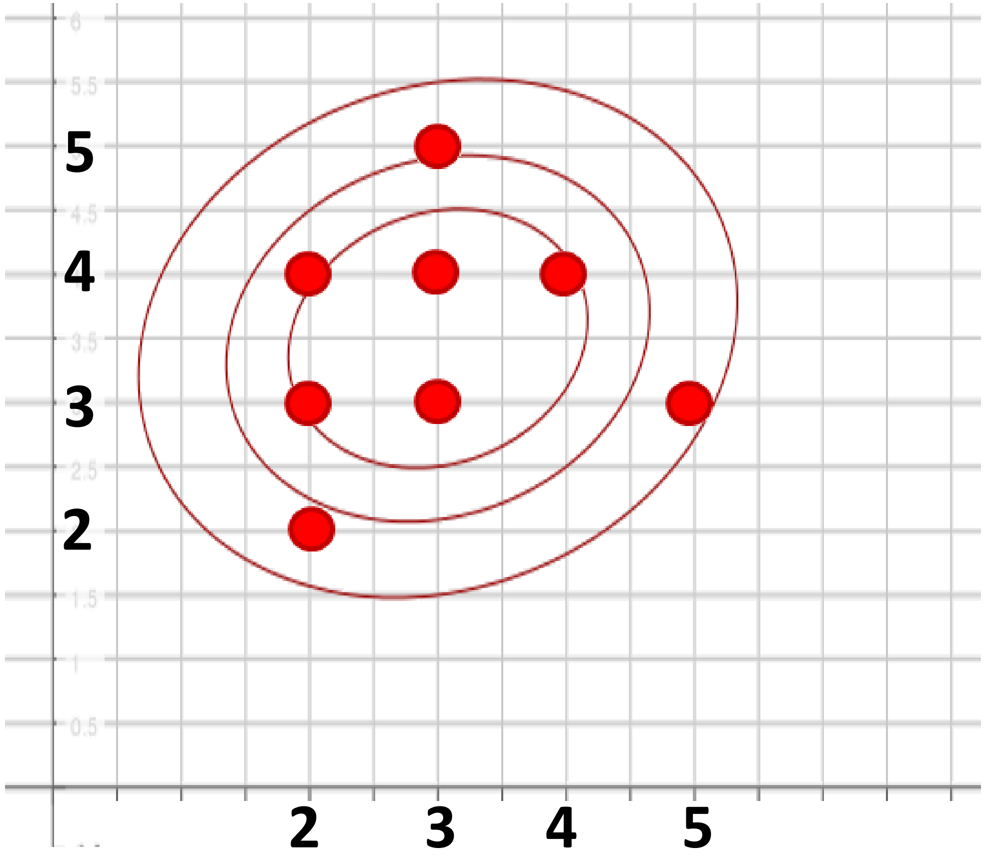
$$A^{[k]} = \left(B^{[k]}\right)^{-1}$$

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

Gaussian classifier

Example

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



Gaussian classifier

$$\mu_1 = MO\left(x_1 \middle| \bullet\right) = 3$$

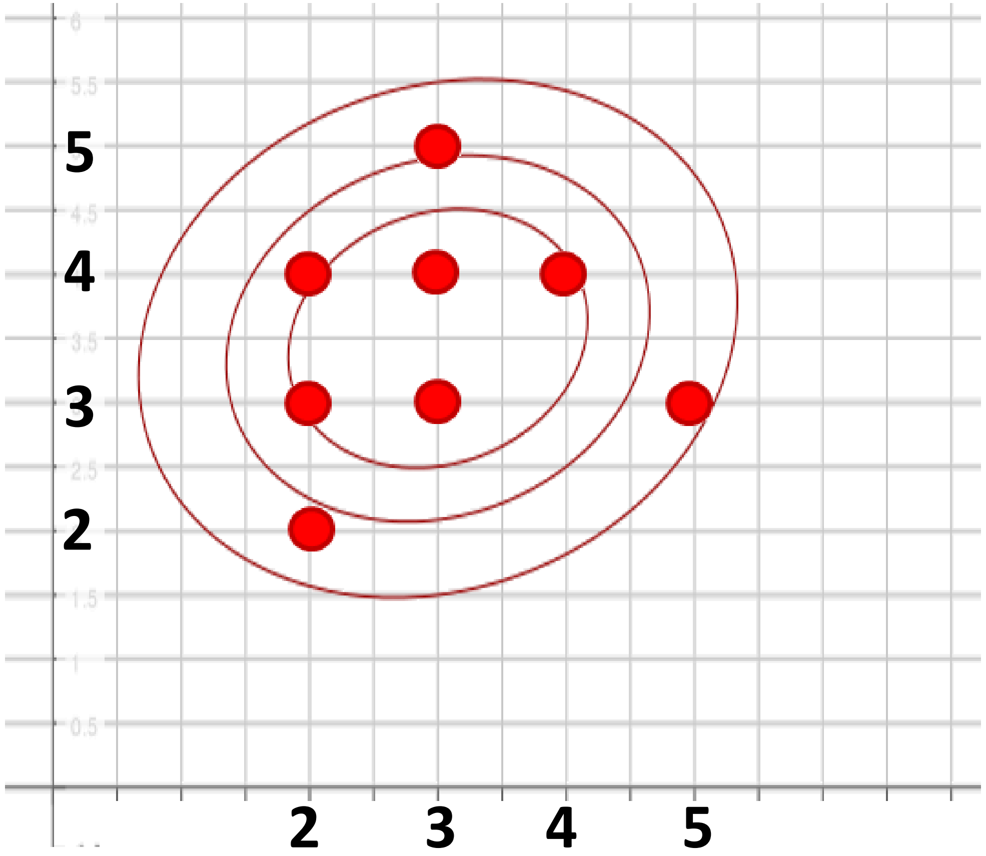
$$\mu_2 = MO\left(x_2 \middle| \bullet\right) = 3{,}5$$

$$b_{11} = MO\underbrace{\left(x_1 - \mu_1\right)\cdot\left(x_1 - \mu_1\right)}_{(1+1+1+0+0+0+1+4)/8} = 1$$

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

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

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



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

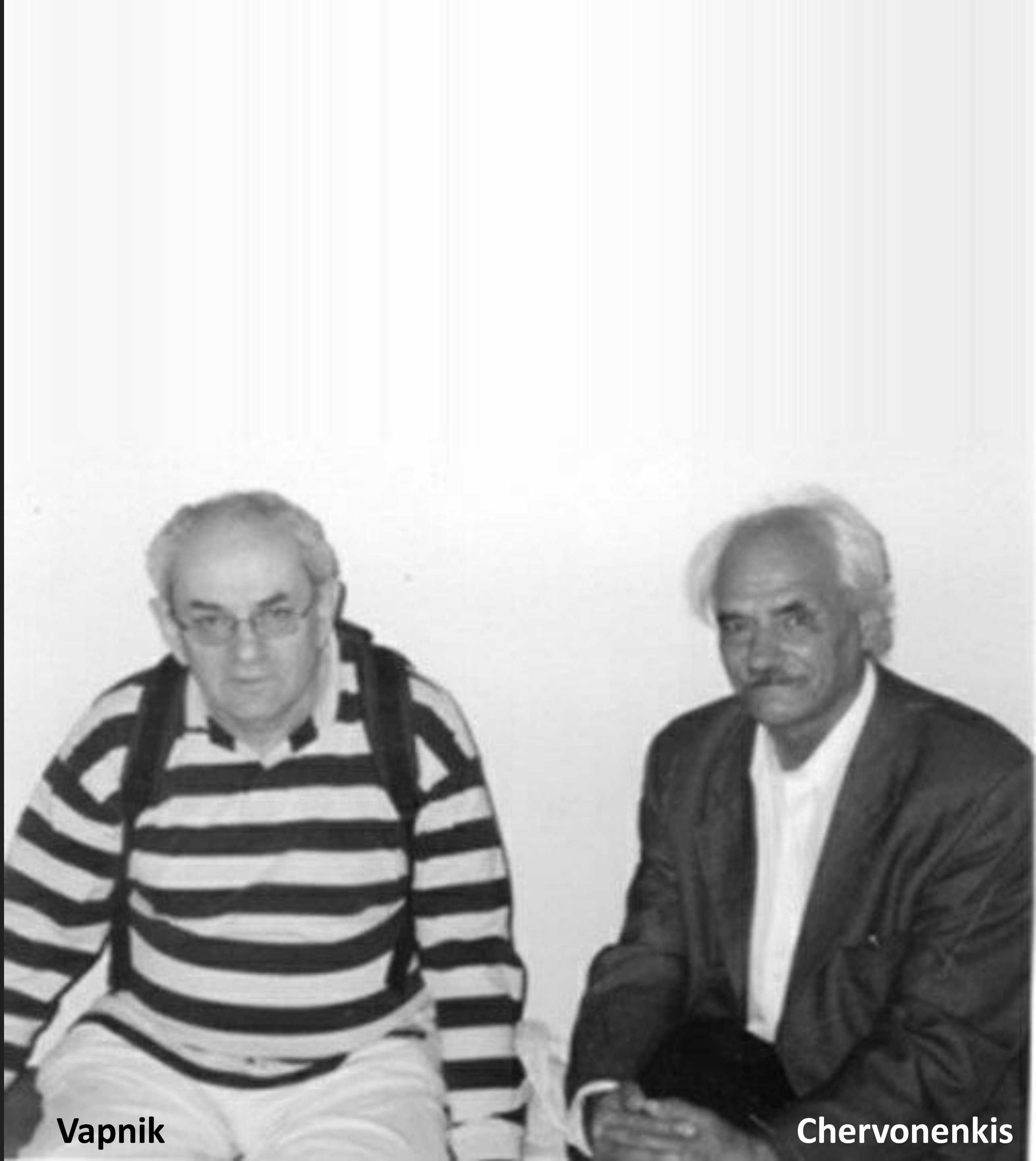
Gaussian classifier

Decision strategy

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

$$\boxed{\sum_i \sum_j \alpha_{ij} \cdot x_i x_j + \sum_i \beta_i \cdot x_i \underset{2}{\overset{1}{\gtreqless}} \theta}$$

Linear discrimination



Linear discrimination

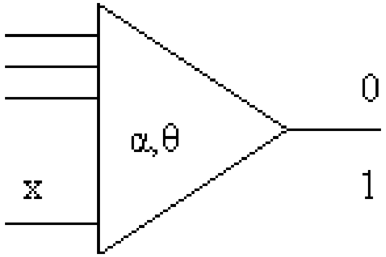
Perceptron

input
$$\begin{cases} X = \{x_1, x_2, \ldots, x_r\} \\ X' = \{x'_1, x'_2, \ldots, x'_s\} \end{cases}$$

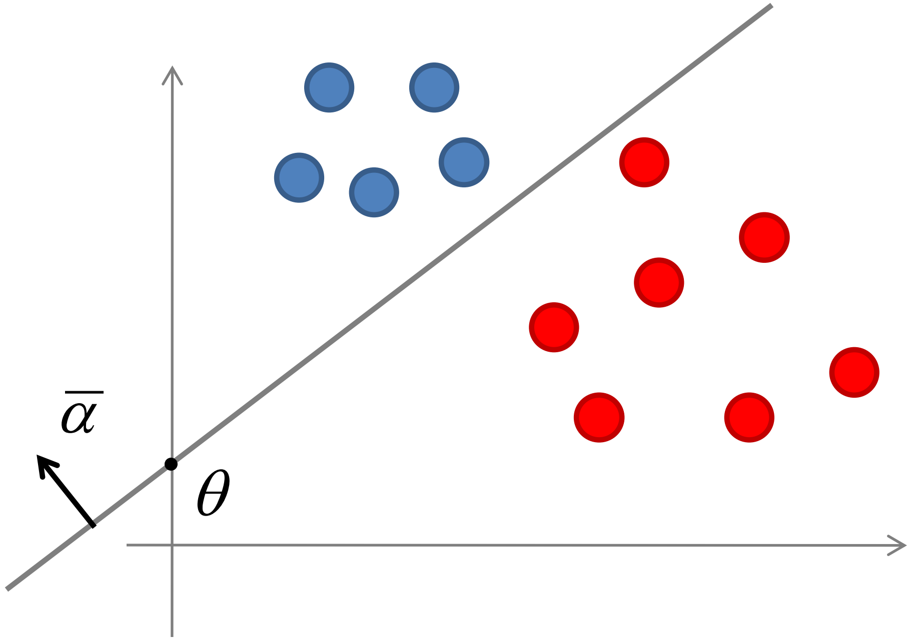
output
$\alpha \in R^n$
θ

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

Perceptron



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



Linear discrimination

Perceptron

$$\begin{cases} \forall x \in X & \sum_{i=1}^{n} x_i \cdot \alpha_i > \theta \\ \forall x' \in X' & \sum_{i=1}^{n} x'_i \cdot \alpha_i < \theta \end{cases} = \begin{cases} \forall x \in X & \sum_{i=1}^{n} x_i \cdot \alpha_i + 1 \cdot \alpha_{n+1} > 0 \\ \forall x' \in X' & \sum_{i=1}^{n} x'_i \cdot \alpha_i + 1 \cdot \alpha_{n+1} < 0 \end{cases}$$

Linear discrimination

Perceptron: tuning

$t = 0$

$\alpha_t = 0$

while (sets are not separated by hyperplane)

$$\{$$

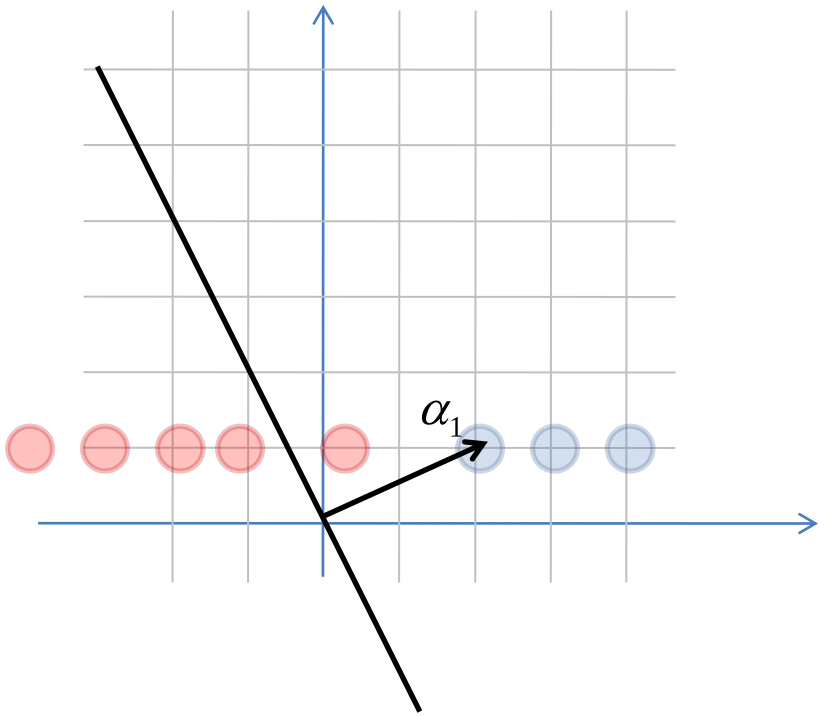
$$if\left(\exists x \in X \middle| (x, \alpha_t) < 0\right) \quad \alpha_{t+1} = a_t + x;$$

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

$$\}$$

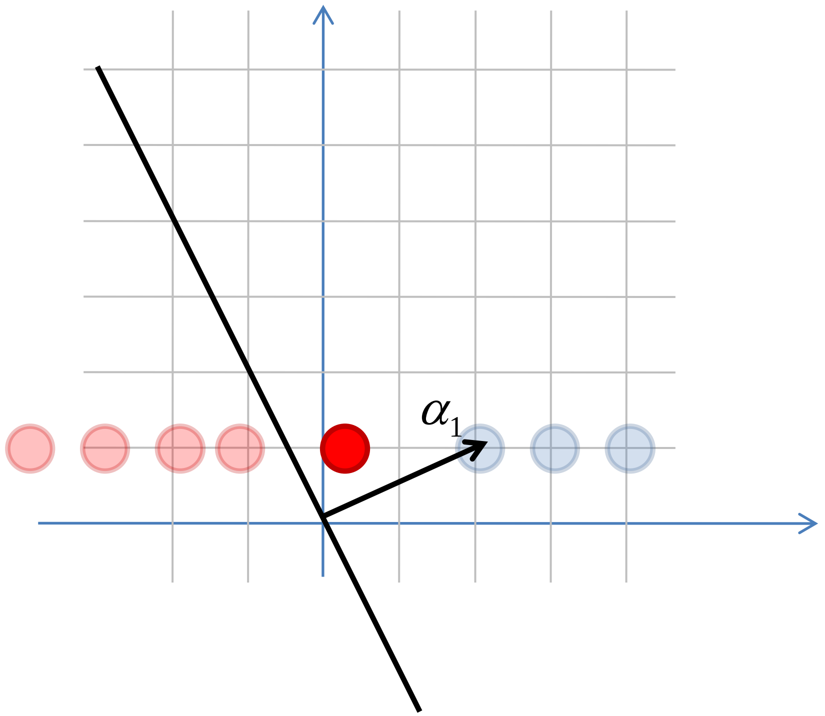
Linear discrimination

Perceptron: tuning

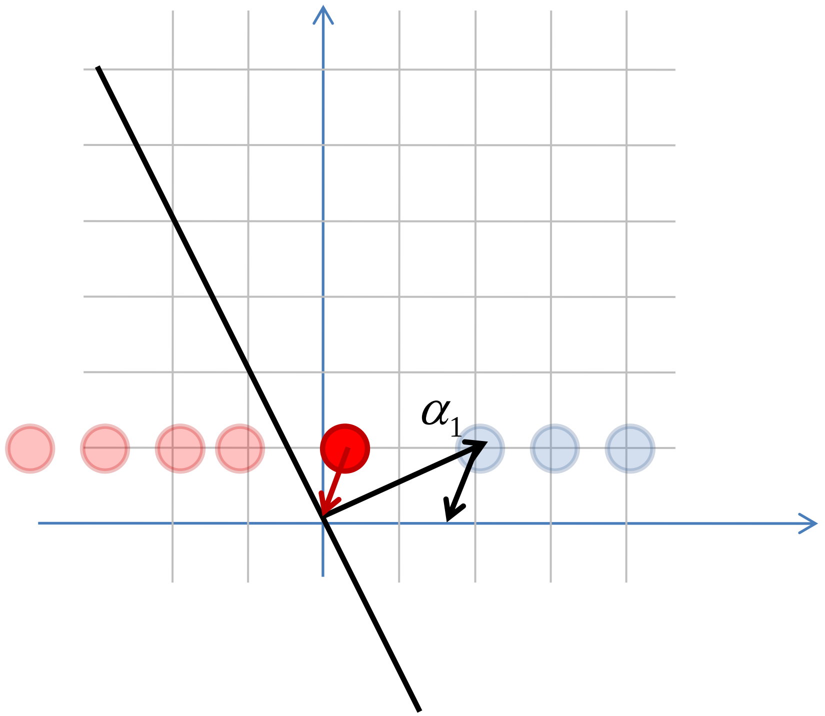


Linear discrimination

Perceptron: tuning

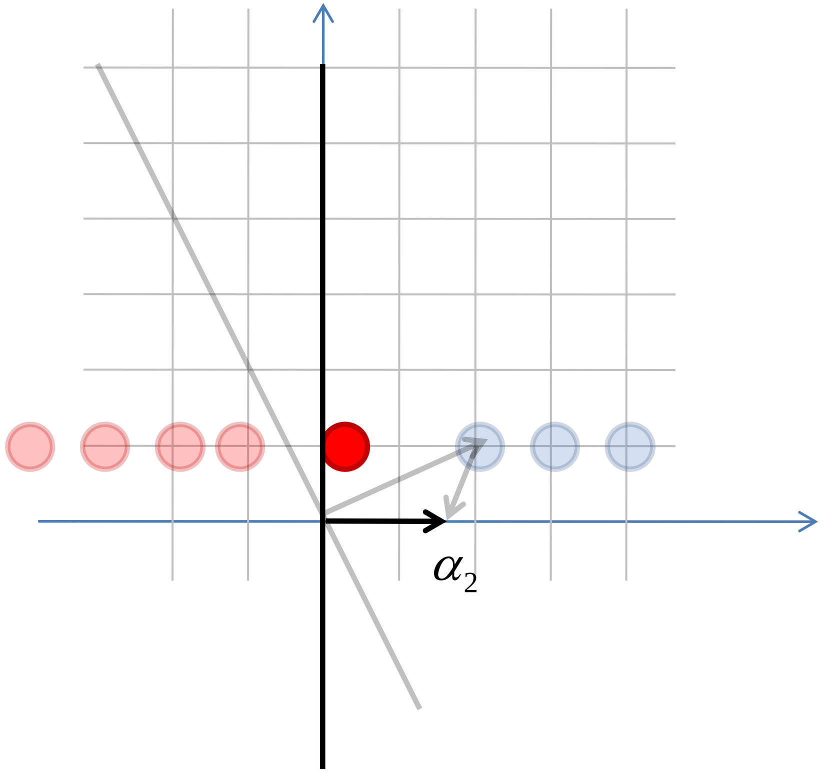


Perceptron: tuning



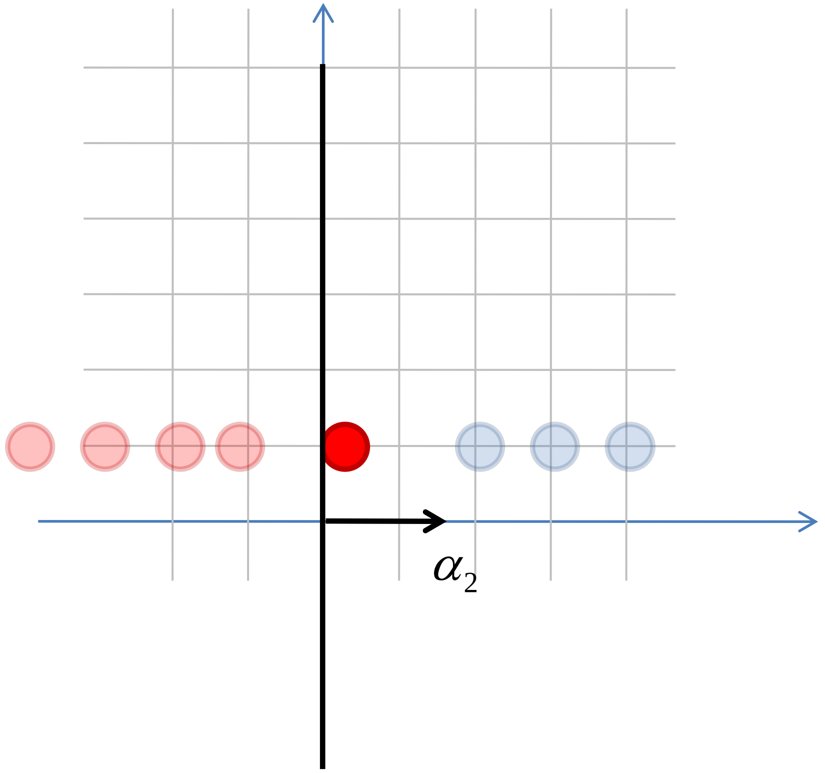
Linear discrimination

Perceptron: tuning



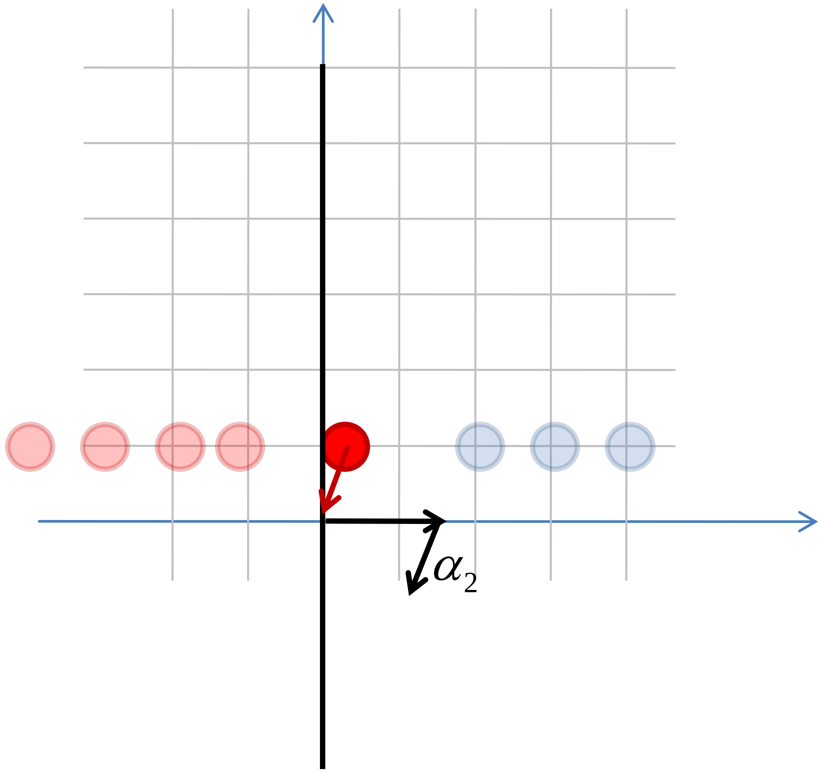
Linear discrimination

Perceptron: tuning



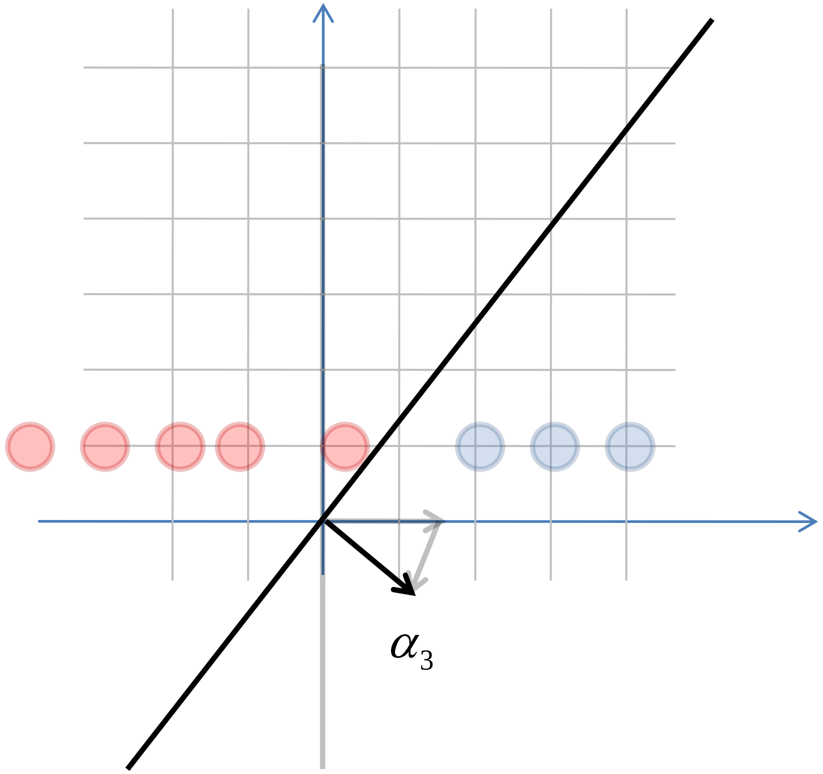
Linear discrimination

Perceptron: tuning



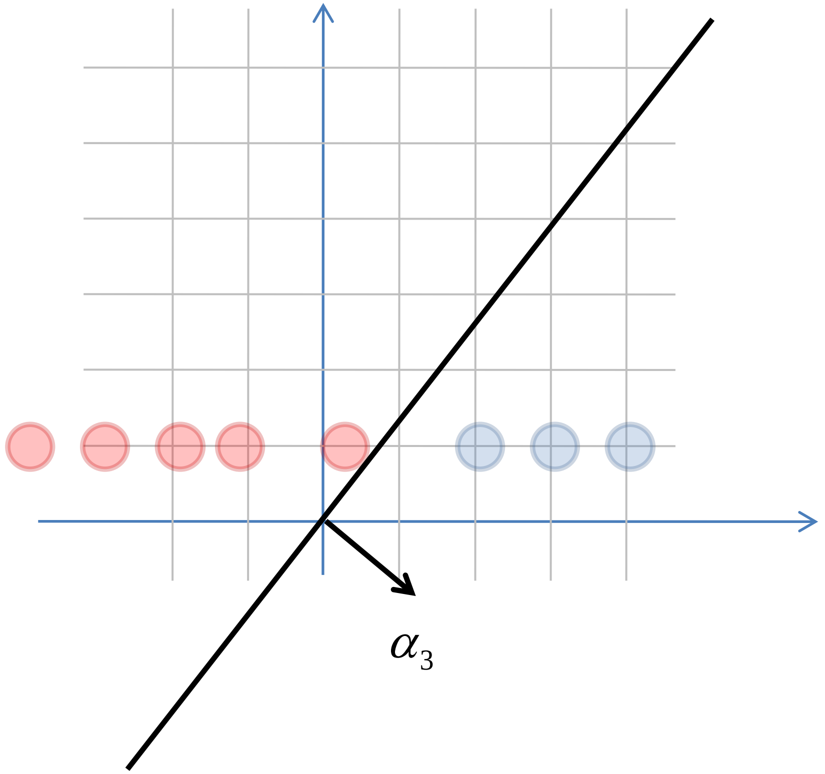
Linear discrimination

Perceptron: tuning



Linear discrimination

Perceptron: tuning



Linear discrimination

Perceptron: convergence

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

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

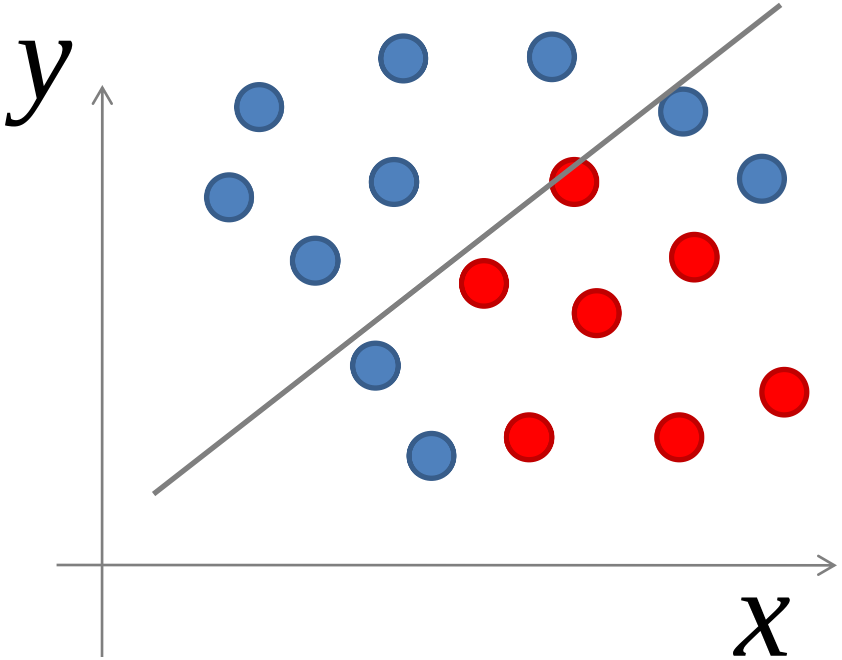
$$\boxed{t \leq \frac{D^2}{\varepsilon}}$$

Linear discrimination

Transformation of the feature space

$$Ax + By \gtrless C$$

a line in R^2

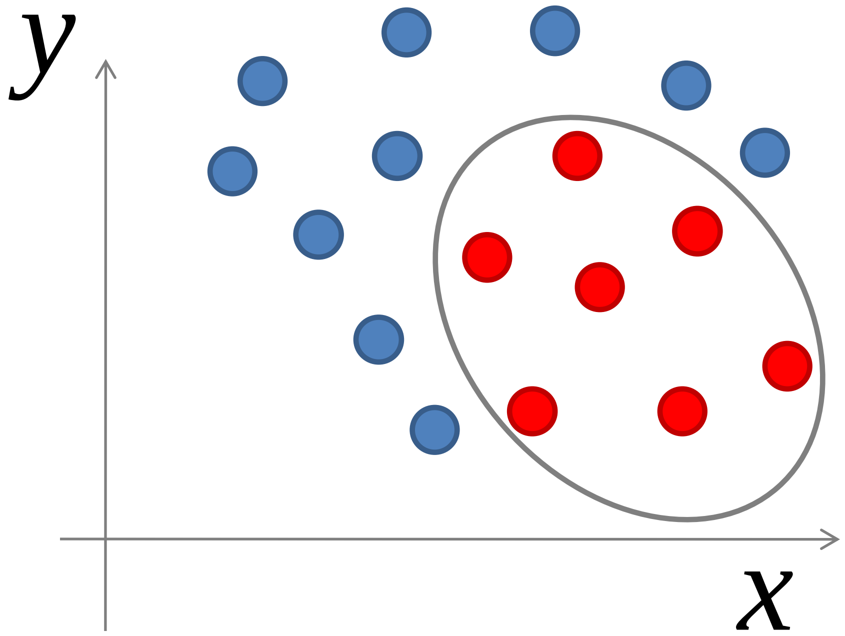


Transformation of the feature space

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

hyperplane in R^5

elliptic curve in R^2



Linear discrimination

SVM: support vector machine

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

$$(\alpha,\alpha) \to \min$$

> Separate by the widest band

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

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

Separate by the widest band