Training and Test

Computational Social Intelligence - Lecture 14

Prof. Alessandro Vinciarelli
School of Computing Science &
Institute of Neuroscience and Psychology

http://www.dcs.gla.ac.uk/vincia
Alessandro.Vinciarelli@glasgow.ac.uk

Texts (see Moodle)

This lecture is based on the following text (available on Moodle):

- Chapter 5 of F.Camastra and A.Vinciarelli, "Machine Learning for Audio, Image and Video Processing", Springer Verlag, 2008.

Outline

- Training for the likelihood
- Training for the priors
- K-fold and performance measurement
- Conclusions

Outline

- Training for the likelihood
- Training for the priors
- K-fold and performance measurement
- Conclusions

Training

- Training is the mathematical process through which the parameters of statistical distributions are adapted to training data;
- The process is often referred to as "learning from data";
- In general, the training takes place by selecting the parameter values that are optimal according to a given criterion.

Sum over the standard deviations of the individual features

True when features statistically independent given the class

$$\log p(\vec{x}|\mathcal{C}_k) = -\sum_{i=1}^{D}\left[\log\sqrt{2\pi\sigma_{ik}} + \frac{(x_i - \mu_{ik})^2}{2\sigma_{ik}^2}\right]$$

Euclidean distance between feature vector and class average

Every element of the training set is a pair of a feature vector and class

The training set

$$\mathcal{X} = \{(\vec{x}_1, c_1), \ldots, (\vec{x}_N, c_N)\}$$

$$c_i \in \mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_M\}$$

The class of training feature vector "i"

The number of possible decisions (classes)

The subset of the training set that includes the feature vectors belonging to class “k”

$$\mathcal{X}^{(k)} = \{\vec{x}_1, \ldots, \vec{x}_K\}$$

Feature vectors that belong to the training set and belong to class “k”

The likelihood of the training set

The product of the likelihoods of the individual feature vectors

$$p(\mathcal{X}^{(k)}|\mathcal{C}_k) = \prod_{i=1}^{K} p(\vec{x}_i|\mathcal{C}_k)$$

The log-likelihood of the training set

$$\mathcal{L} = \log p(\mathcal{X}^{(k)}|\mathcal{C}_k) = \sum_{i=1}^{K} \log p(\vec{x}_i|\mathcal{C}_k)$$

Dimensionality of the feature vectors

Component "j" of vector "i"

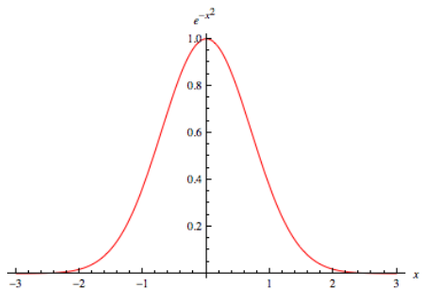
Sample mean of component "j" in the feature vectors belonging to class "k"

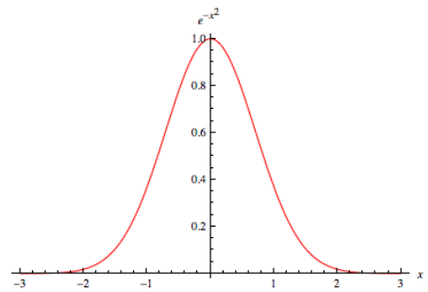
$$\mathcal{L} = -\sum_{i=1}^{K}\sum_{j=1}^{D}\left[\log\sqrt{2\pi}\sigma_{kj} + \frac{(x_{ij}-\mu_{kj})^2}{2\sigma_{kj}^2}\right]$$

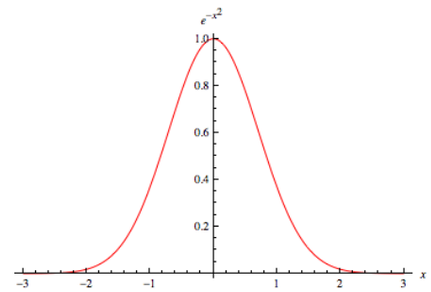
Number of feature vectors belonging to class "k"

Standard deviation of component "j" in the feature vectors belonging to class "k"

$$\mathcal{X}^{(k)} = \begin{pmatrix} x_{11} & \dots & x_{1l} & \dots & x_{1D} \\ x_{21} & \dots & x_{2l} & \dots & x_{2D} \\ \dots & \dots & \dots & \dots & \dots \\ x_{K1} & \dots & x_{Kl} & \dots & x_{KD} \end{pmatrix}$$







$$p(x_1|\mathcal{C}_k) \qquad p(x_l|\mathcal{C}_k) \qquad p(x_D|\mathcal{C}_k)$$

Setting the derivative to zero allows to find the value of the parameters that maximise the log likelihood

Sample mean of component "l" in the feature vectors belonging to class "k"

$$\frac{\partial \mathcal{L}}{\partial \sigma_{kl}} = \sum_{i=1}^{K} \left[\frac{(x_{il} - \mu_{kl})^2}{2\sigma_{kl}^3} - \frac{1}{\sigma_{kl}} \right] = 0$$

Standard deviation of component "j" in the feature vectors belonging to class "k"

Setting the derivative to zero allows to find the value of the parameters that maximise the log likelihood

$$\frac{\partial \mathcal{L}}{\partial \mu_{kl}} = 2 \sum_{i=1}^{K} \frac{(x_{il} - \mu_{kl})}{2\sigma_{kl}^2} = 0$$

The maximum likelihood estimate of the parameter

The sample mean of component “l” in the feature vectors belonging to class “k”

$$\mu_{kl} = \frac{1}{K} \sum_{i=1}^{K} x_{il}$$

Number of feature vectors belonging to class “k” in the training set

The maximum likelihood estimate of the parameter

The sample variance of component "l" in the feature vectors belonging to class "k"

$$\sigma_{kl}^2 = \frac{1}{K} \sum_{i=1}^{K} (x_{il} - \mu_{kl})^2$$

Number of feature vectors belonging to class "k" in the training set

$$\begin{pmatrix} x_{11} & \dots & x_{1l} & \dots & x_{1D} \\ x_{21} & \dots & x_{2l} & \dots & x_{2D} \\ \dots & \dots & \dots & \dots & \dots \\ x_{K1} & \dots & x_{Kl} & \dots & x_{KD} \end{pmatrix}$$

$$p(x_l|\mathcal{C}_k) = \frac{1}{\sqrt{2\pi}\sigma_{kl}} e^{\frac{(x_{kl}-\mu_{kl})^2}{2\sigma_{kl}^2}}$$

Recap

- For every component of the feature vectors and for every class there is a different Gaussian distribution;
- Means and standard deviations of the Gaussians have been set to maximise the likelihood over the training data;

Outline

- Training for the likelihood
- Training for the priors
- K-fold and performance measurement
- Conclusions

Number of samples belonging to class "j"

$$n(\mathcal{C}_1) = n_1, \ldots, n(\mathcal{C}_M) = n_M$$

$$\sum_{j=1}^{M} n_j = N$$

Sum over all numbers of samples belonging to the classes

Total number of samples

Probability of observing the number of training samples per class

$$p(n_1, \ldots, n_M) = \prod_{j=1}^{M} p(C_j)^{n_j}$$

Prior of class "j"

Log likelihood

$$\mathcal{L} = \sum_{j=1}^{M} n_j \log p(\mathcal{C}_j) + \lambda \left(1 - \sum_{j=1}^{M} p(\mathcal{C}_j) \right)$$

Lagrange multiplier

The content of the parenthesis is null

Setting the derivative to zero allows to find the value of the parameters that maximise the log likelihood

$$\frac{\partial \mathcal{L}}{\partial p(\mathcal{C}_l)} = \frac{n_l}{p(\mathcal{C}_l)} - \lambda = 0$$

$$n_l = \lambda p(\mathcal{C}_l)$$

The value of the Lagrange multiplier must still be found

$$N = \sum_{l=1}^{M} n_l = \lambda \sum_{l=1}^{M} p(\mathcal{C}_l) = \lambda$$

The value of the Lagrange multiplier

$$n_l = \lambda p(\mathcal{C}_l)$$

$$p(\mathcal{C}_l) = \frac{n_l}{N}$$

The maximum likelihood estimate of the prior of class “l”

Recap

- The maximum likelihood estimate of the priors is the percentage of samples belonging to the classes in the training set;
- A possible alternative is to use a-priori knowledge about the problem under exam (e.g., it is known that men and women are 50% of the population);

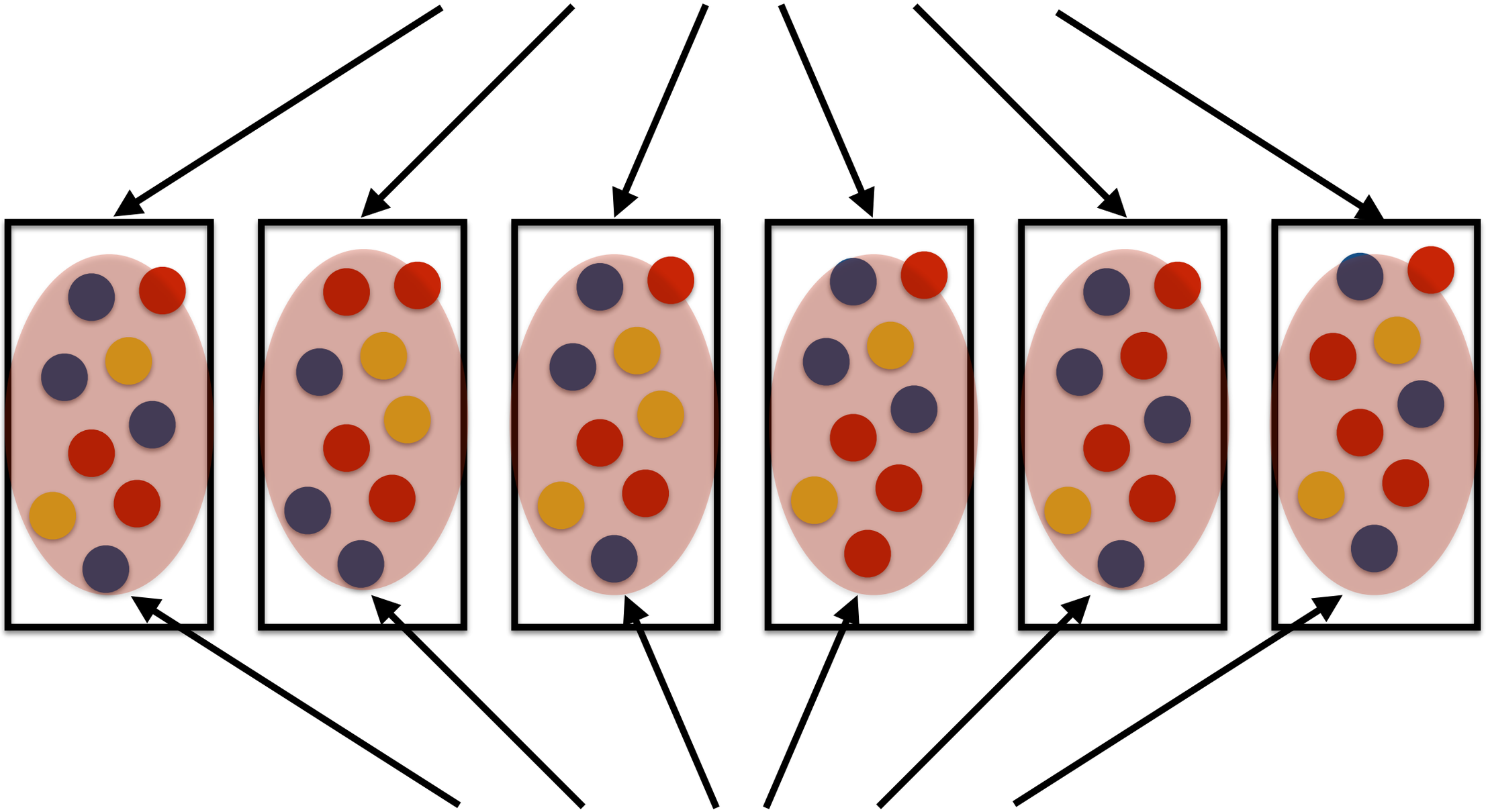
Outline

- Training for the likelihood
- Training for the priors
- K-fold and performance measurement
- Conclusions

Training and Test Set

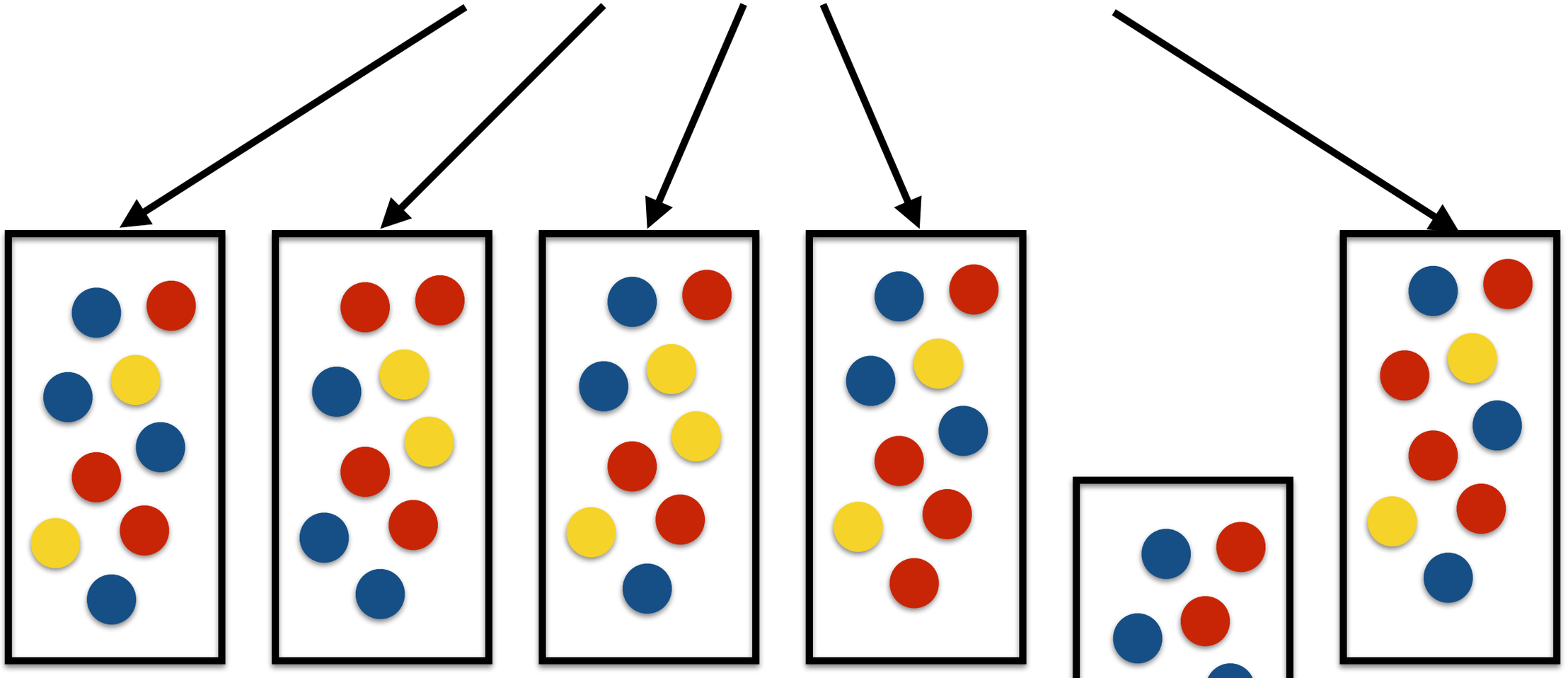
- There must be a separation between the data used for training and the data used for testing;
- The reason is that an approach must generalise, i.e., it must be capable to make the right decision for previously unseen data;
- The k-fold approach allows one to test over the whole dataset at disposition while keeping separated training and test set.

The folds are subsets of
the data at disposition



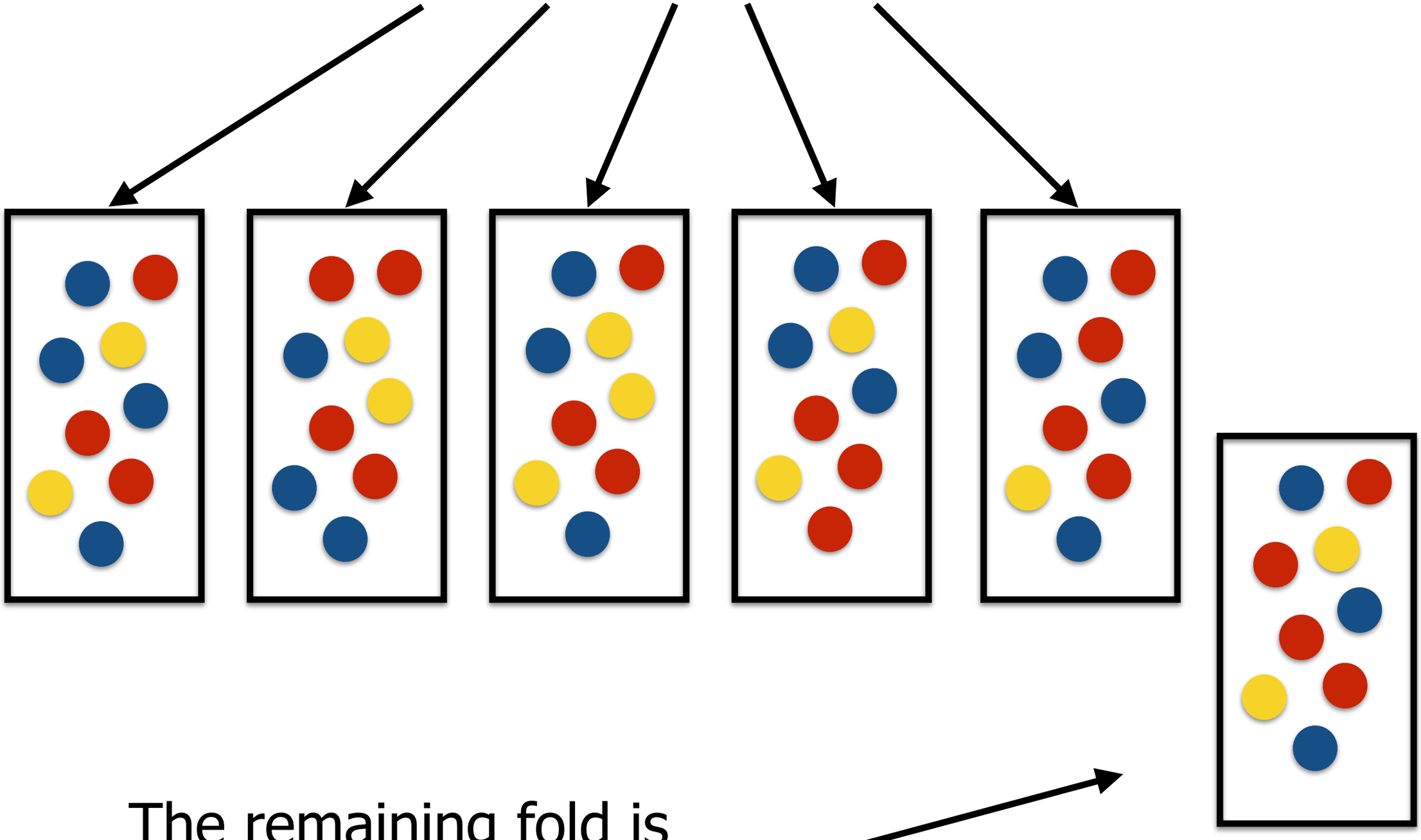
Number of samples and
class distribution are
roughly the same

K-1 folds are used for training



The remaining fold is used for testing

K-1 folds are used for training



The remaining fold is used for testing

K-Fold

- The dataset at disposition is split into K disjoint subsets (the folds) of roughly the same size;
- The class distribution is roughly the same in all folds (it is sufficient to select the folds through a random process);
- Every fold is used once as a test set and K-1 times as a part of the training set.

Performance measured in terms of average loss

Action that minimises the Bayes Risk

$$\alpha = \frac{1}{N}\sum_{i=1}^{N} \pi(\mathcal{B}_i^*|\vec{x}_i)$$

Loss associated to the action that minimises the Bayes Risk

Fraction of times the approach makes the wrong decision (error rate)

Action that minimises the Bayes Risk

$$\alpha = \frac{1}{N} \sum_{i=1}^{N} \pi(\mathcal{B}_i^* | \vec{x}_i)$$

In the Zero-One loss case, it is 1 when the action (the decision is wrong) and zero otherwise

Outline

- Training for the likelihood
- Training for the priors
- K-fold and performance measurement
- Conclusions

Conclusions

- The parameters of the discriminant functions are set by optimising a criterion;
- The maximisation of the likelihood is one of the training approaches most commonly adopted;
- Once the models are trained, it is possible to test them over unseen data and measure their performance.