Machine Learning
Lecture 4: SVM, PCA

Harbour.Space University
February 2020

Radoslav Neychev

Outline

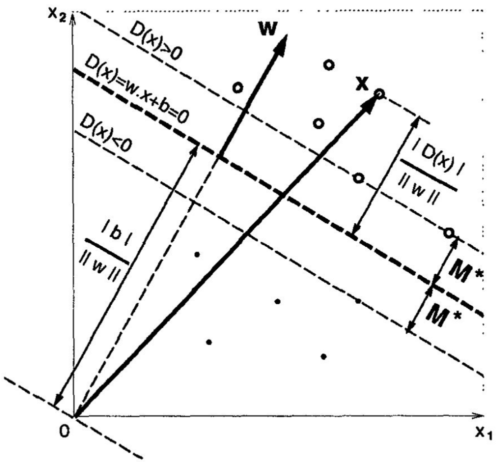
1. Maximum Likelihood Estimation (MLE)
2. Support Vector Machine (SVM)
3. Multiclass classification strategies
4. Dimensionality reduction and PCA
5. Bonus section: Validation strategies

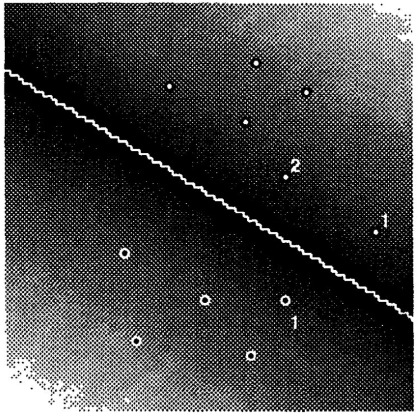
Support Vector Machine

Support Vector Machine

1. History
2. Motivation
3. Solution for separable design
4. Inseparable design, soft margin
5. Kernels
    a. Kernel definition (Hilbert spaces, inner product, positive semidefiniteness)
    b. Kernels properties (addition, infinite sums)
    c. Types of kernels (poly, exponential, gaussian)
6. Current state

Historical overview



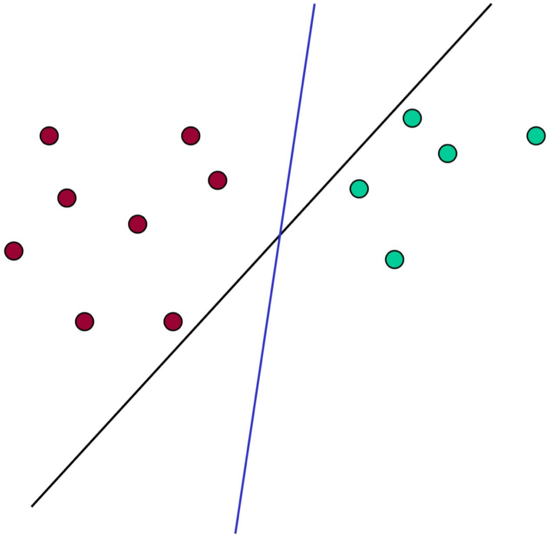


1963: SVM introduced by Soviet mathematicians Vladimir Vapnik and Alexey Chervonenkis

1992: kernel trick (Vapnik, Boser, Guyon)

1995: soft margin (Vapnik, Cortes)

Historical overview



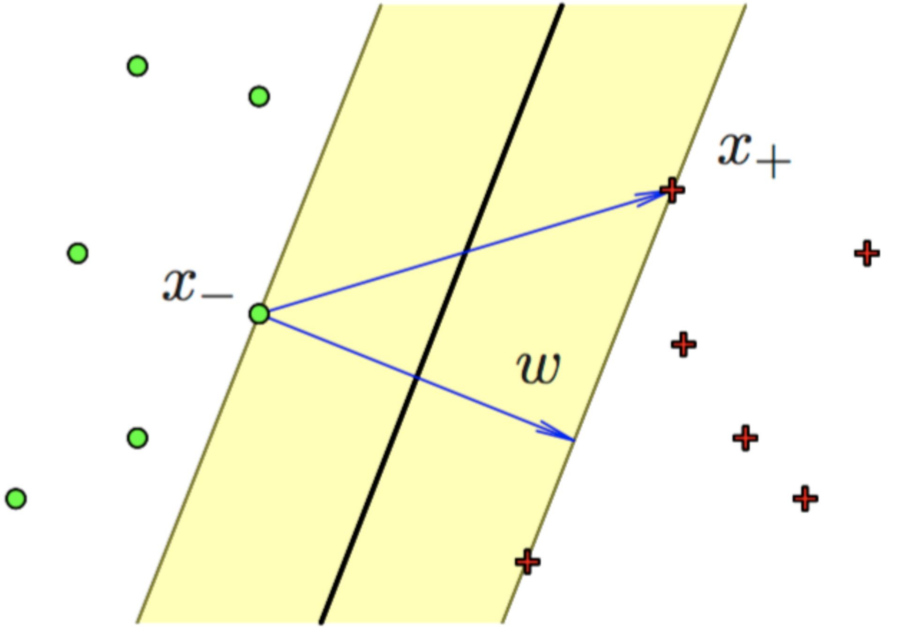
Linear separable case

Many separating hyperplanes exist

Maximize width

Margin recap

$$y \in \{1, -1\}$$
$$y_i = 1 : w^T x_i - c > 0$$
$$y_i = -1 : w^T x_i - c < 0$$
$$c_+(w) = \min_{y_i=1}(w^T x_i)$$
$$c_-(w) = \max_{y_i=-1}(w^T x_i)$$
$$\rho(w) = \frac{c_+(w) - c_-(w)}{2}$$



Theorem

$$\boxed{\rho\left(\frac{w_0}{||w_0||}\right) = \frac{1}{||w_0||}}$$

Optimization problem statement

$$y_i = 1 : w^T x_i - c > 0$$
$$y_i = -1 : w^T x_i - c < 0$$

$$\rho(w) = \frac{1}{||w||} \to \max_{w,c}$$

$$s.t.\ y_i(w^T x_i - c) \geq 1$$

$$M_i = y_i \cdot (w^T x_i - c)$$

Convex problem!

$$L(w, c, \alpha) = \frac{1}{2} w^T w - \sum_i \alpha_i (y_i (w^T x_i - c) - 1)$$

Some of them are zeros

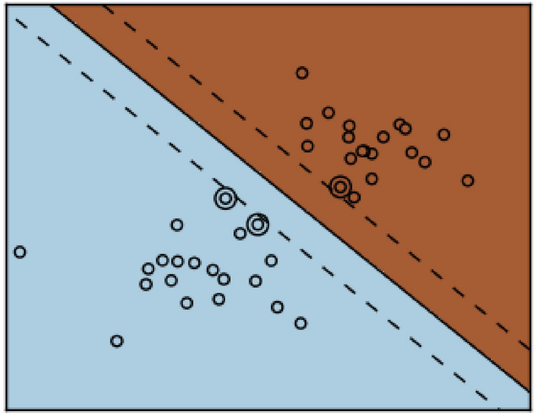
Inseparable case

Let our model mistake, but penalize that mistakes

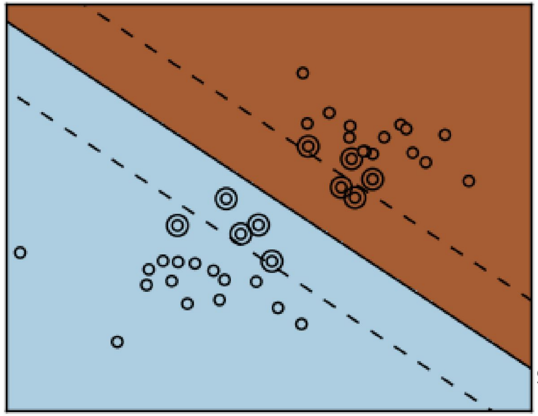
Implemented via margin slack

$$\begin{cases} \frac{1}{2}\langle w, w\rangle + C\sum_{i=1}^{\ell}\xi_i \to \min\limits_{w,w_0,\xi}; \\ y_i\big(\langle w, x_i\rangle - w_0\big) \geqslant 1 - \xi_i, \quad i = 1,\dots,\ell; \\ \xi_i \geqslant 0, \quad i = 1,\dots,\ell. \end{cases}$$

Big C



Small C



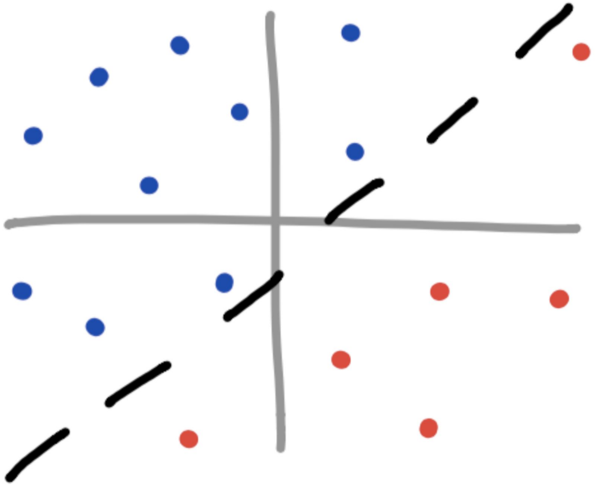
Kernel trick

$$y_i = 1 : w^T x_i - c > 0$$
$$y_i = -1 : w^T x_i - c < 0$$

$$\begin{matrix} x \mapsto \phi(x) \\ w \mapsto \phi(w) \end{matrix} \implies < w, x > \mapsto < \phi(w), \phi(x) >$$

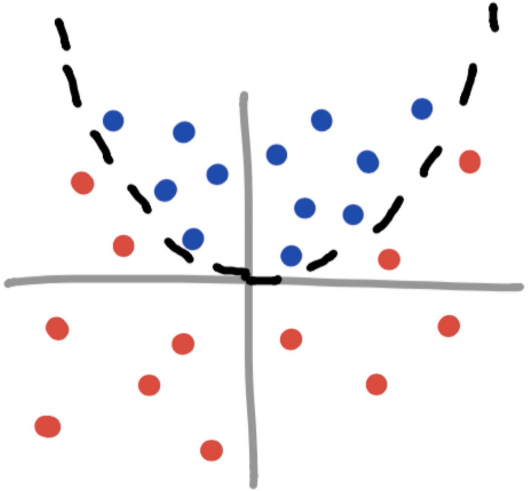
$$K(w, x) = < \phi(w), \phi(x) >$$

Kernel types



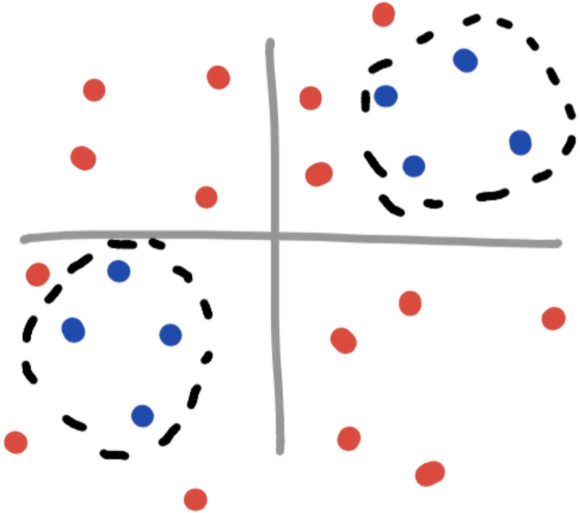
$K(w,x) = < \mathrm{w}, x >$

Linear



$K(w,x) = (\gamma < w,x > + r)^d$

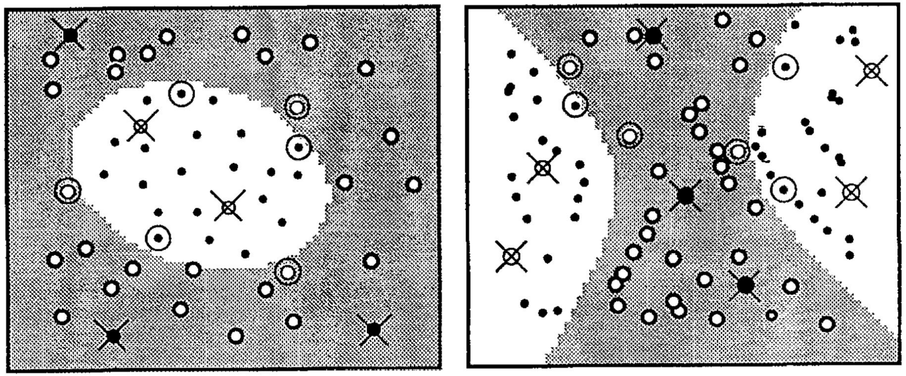
Polynomial



$K(w,x) = e^{-\gamma \|w-x\|^2}$

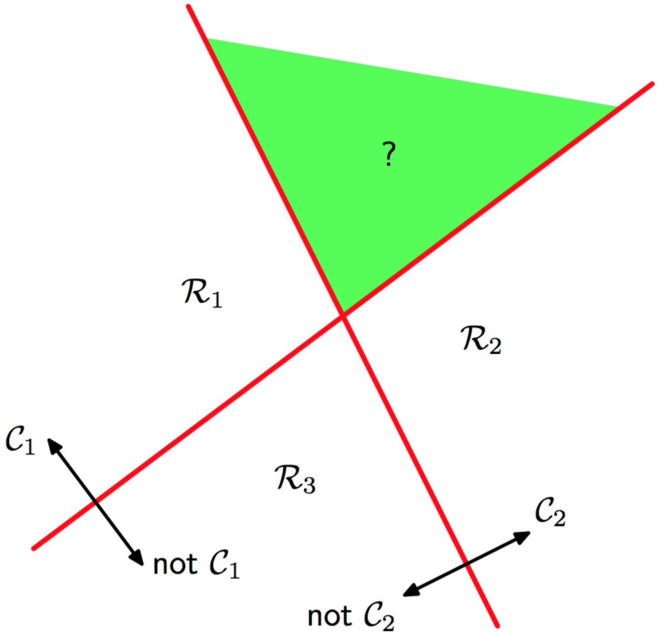
Gaussian radial basis function

Current state

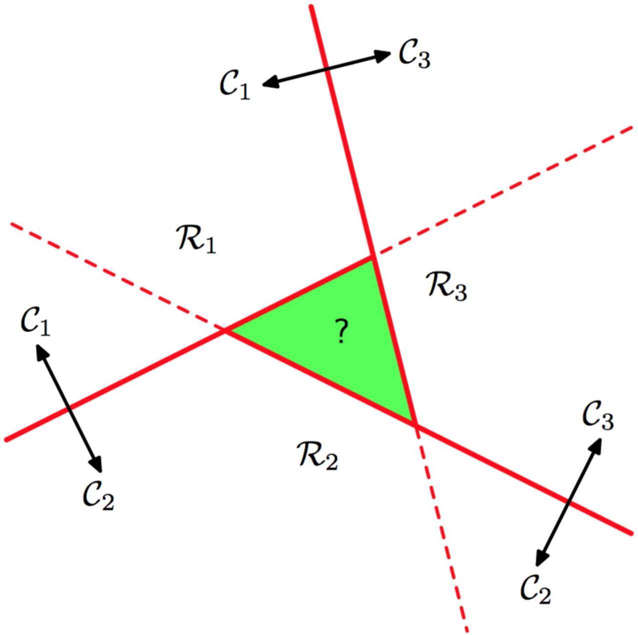


Multiclass classification

Multiclass strategies



One vs Rest



One vs One

Principal Component Analysis

Principal Component Analysis

$$x_1, \dots, x_n \to g_1, \dots, g_k, k \leq n$$

$$U : UU^T = I, G = XU, X = GU^T$$

$$\hat{X} = GU^T \approx X$$

$$||GU^T - X|| \to \min_{G,U} \; s.t. \; rank(G) \leq k$$

Singular value decomposition

$$||GU^T - X|| \to \min_{G,U} \;\; s.t.\; rank(G) \leq k$$

$$X = V\Sigma U^T : ||GU^T - V\Sigma U^T||_2 = ||G - V\Sigma||_2$$

$$G = V\Sigma' : ||V\Sigma' - V\Sigma||_2 = ||\Sigma' - \Sigma||_2$$

$$||A||_2 = \sigma_{max}(A) : ||\Sigma' - \Sigma||_2 = \sigma_k(\Sigma) = \sigma_k(X)$$

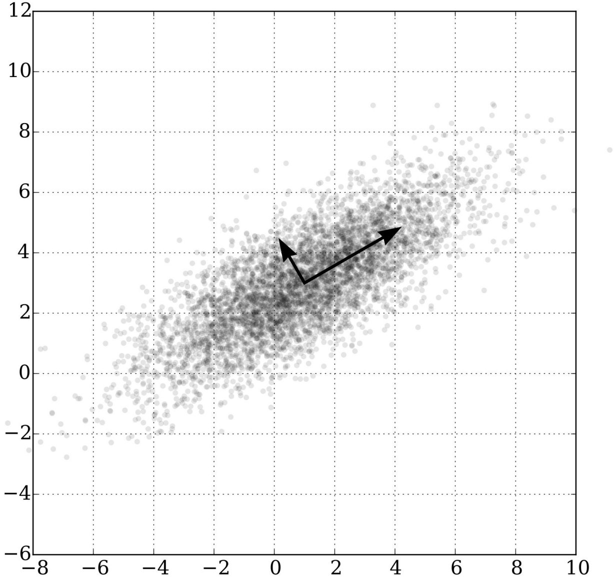
Eckart–Young–Mirsky theorem

Another point of view: variance maximization

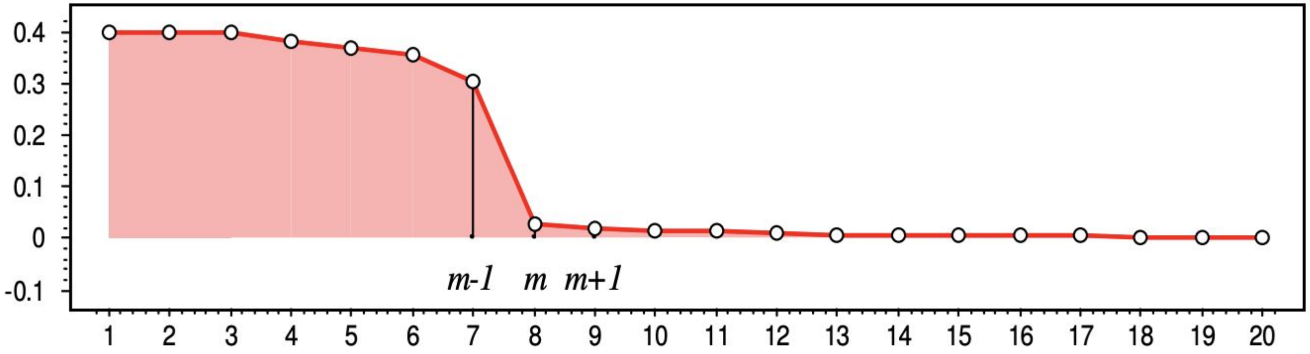
Residual variance maximization

Take new basis vectors greedy

Same result for G and U



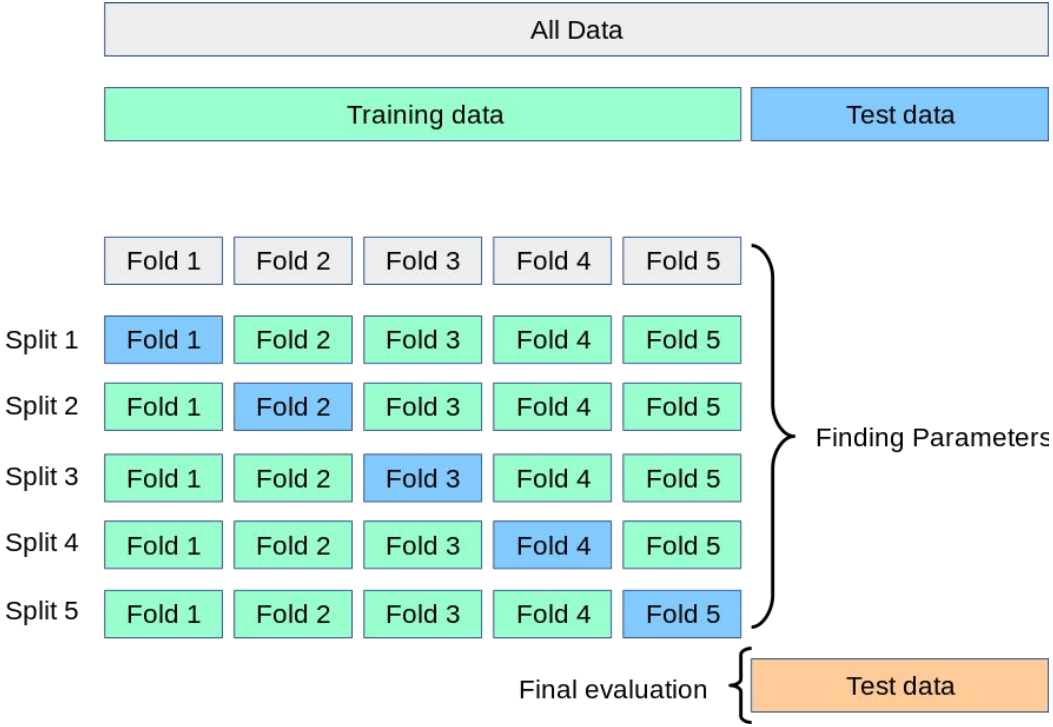
Dimensionality reduction



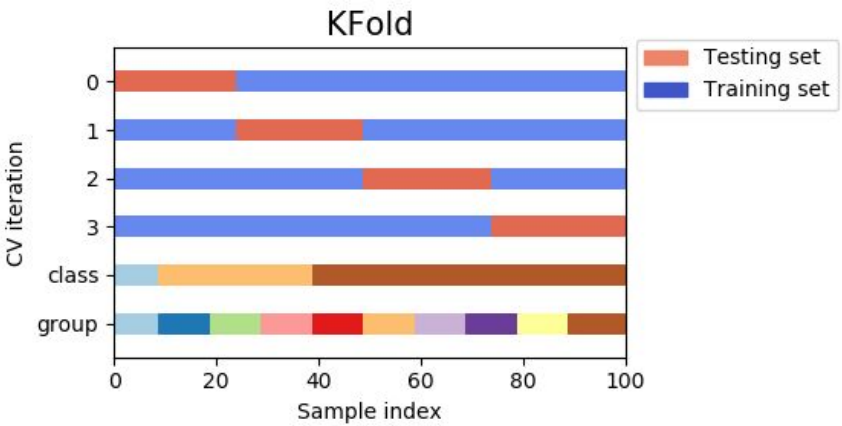
Get rid of low-variance components

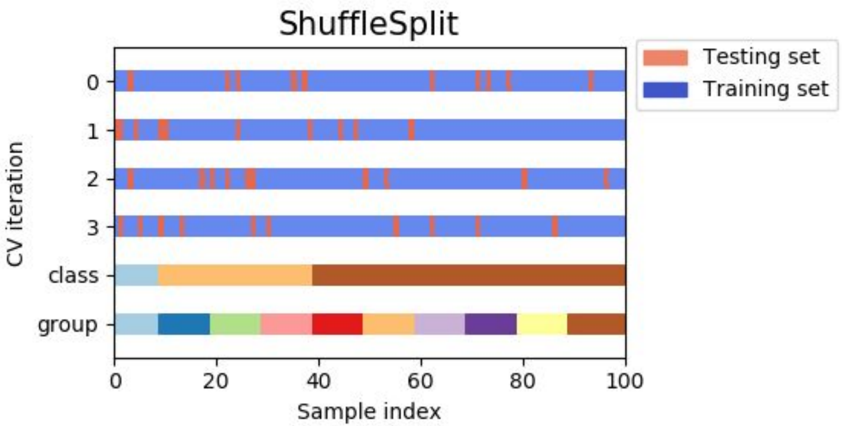
Bonus section:
More validation strategies

Validation strategies



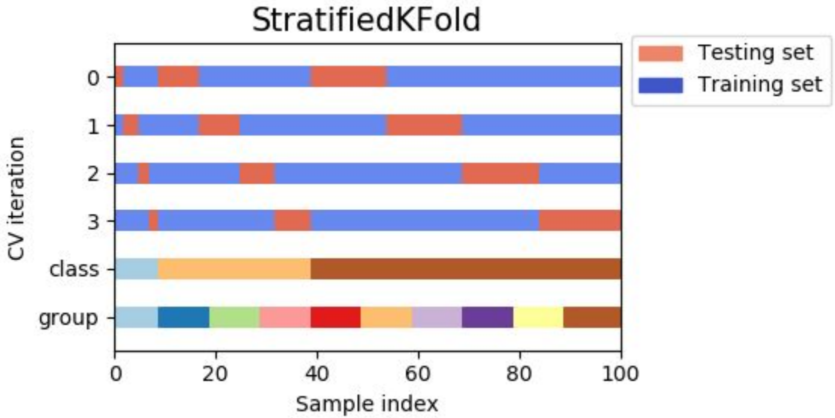
Validation strategies

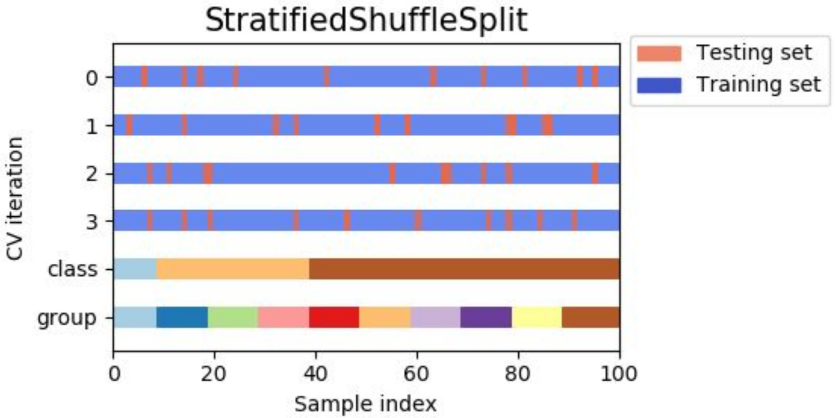




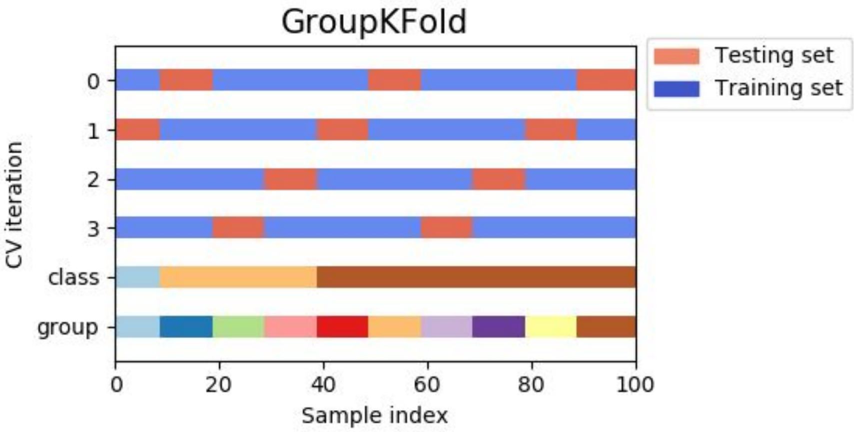
Special case: Leave One Out (LOO) - good for tiny datasets

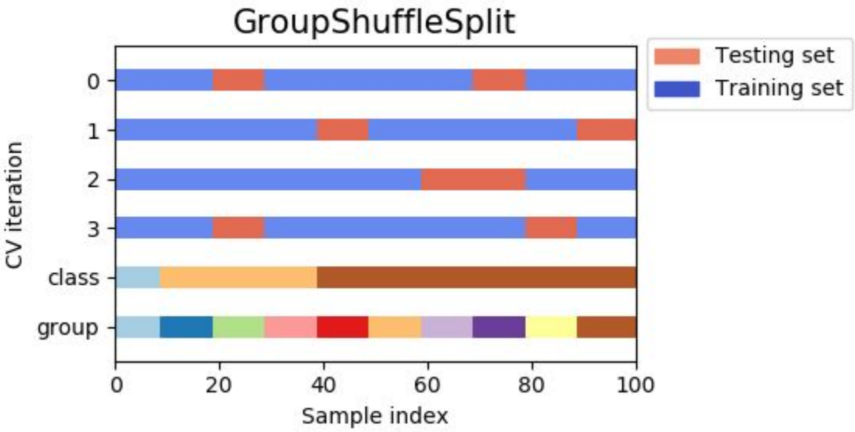
Validation strategies



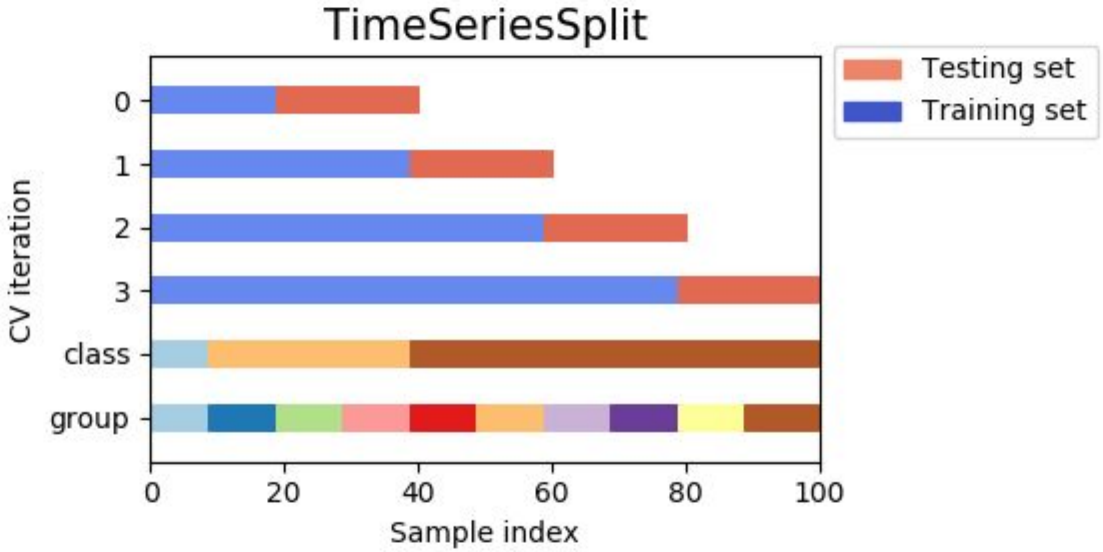


Validation strategies





Special case: time series



Never use `train_test_split` in this case!