

Lecture 4:

SVM, PCA

Vladislav Goncharenko
Harbor, 2021





Outline

1. Support Vector Machine (SVM)
2. Dimensionality reduction and PCA
3. Validation strategies

Maximum Likelihood Estimation



Maximum Likelihood Estimation

What are reasons behind defining “best” linear estimator?

Maximize probability of particular parameter to explain given data

$$L(\theta|X, Y) = P(X, Y|\theta)$$

assuming i.i.d. observations

$$P(X, Y|\theta) = \prod_{i=1}^n P(x^i, y^i|\theta)$$

$$\log L(\theta|X, Y) = \sum_{i=1}^n \log P(x^i, y^i|\theta)$$

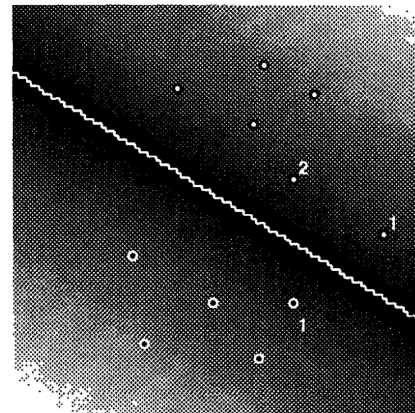
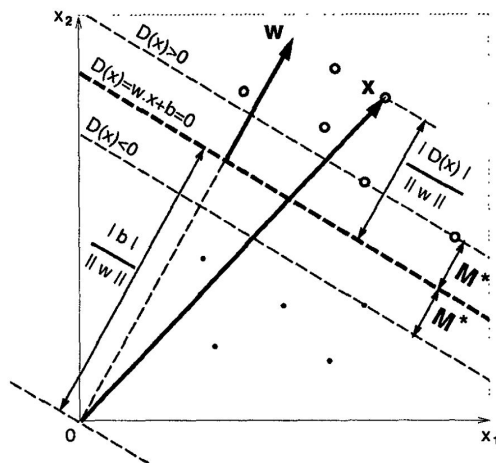
Support Vector Machine



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

History



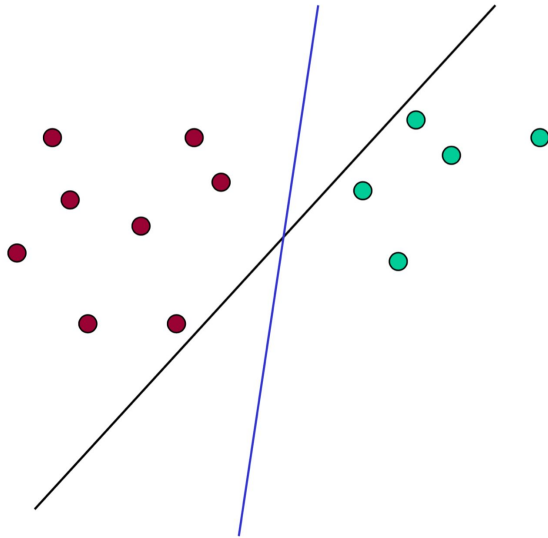
1963: SVM introduced by Soviet mathematicians Vladimir Vapnik and Alexey Chervonenkis

1992: kernel trick (Vapnik, Boser, Guyon)

1995: soft margin (Vapnik, Cortes)



Motivation



Linear separable case

Many separating hyperplanes exist

Maximize width

Margin

$$y \in \{1, -1\}$$

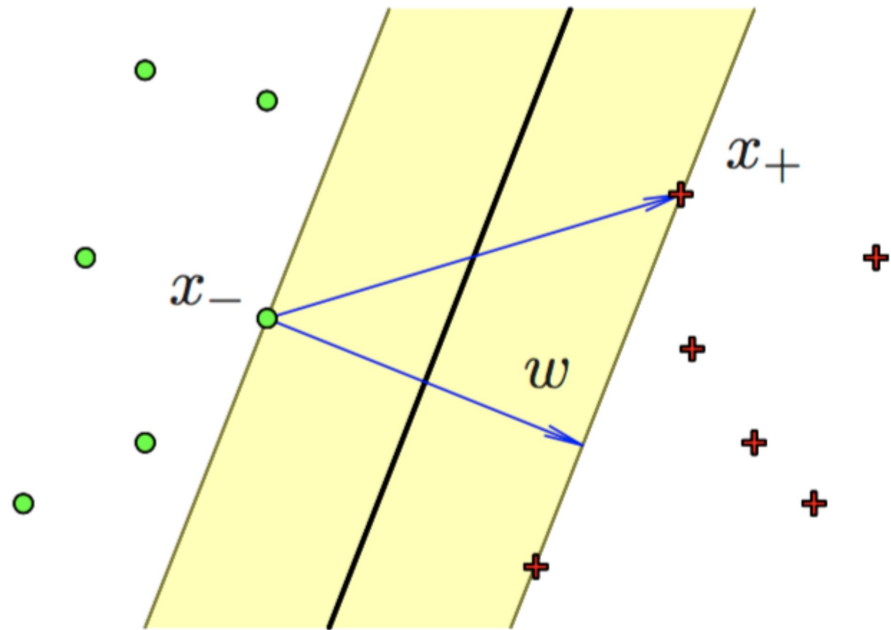
$$y_i = 1 : w^T x_i - c > 0$$

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

$$c_+(w) = \min_{y_i=1} (w^T x_i)$$

$$c_-(w) = \max_{y_i=-1} (w^T x_i)$$

$$\rho(w) = \frac{c_+(w) - c_-(w)}{2}$$



Theorem

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




Optimization problem

$$\begin{aligned} y_i = 1 & : w^T x_i - c > 0 \\ y_i = -1 & : w^T x_i - c < 0 \\ M_i = y_i \cdot (w^T x_i - c) & \quad s.t. \quad y_i(w^T x_i - c) \geq 1 \end{aligned} \quad \rho(w) = \frac{1}{||w||} \rightarrow \max_{w,c}$$

Convex problem!

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

 Many of them are zeros



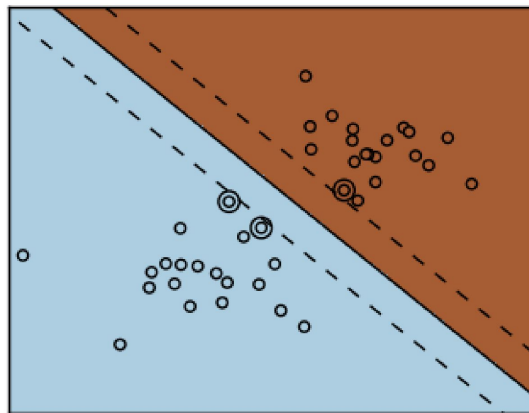
Inseparable case

Let our model mistake, but penalize that mistakes

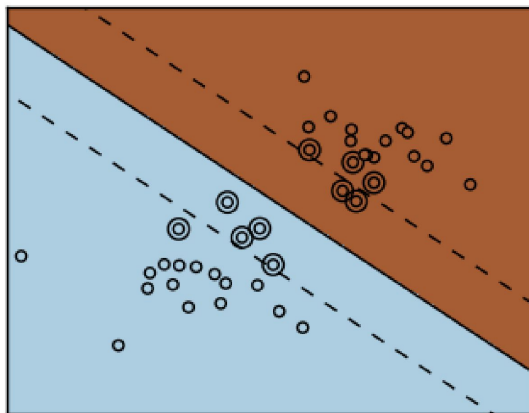
Implemented via margin slack variables

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

Big C



Small C





Kernel trick

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

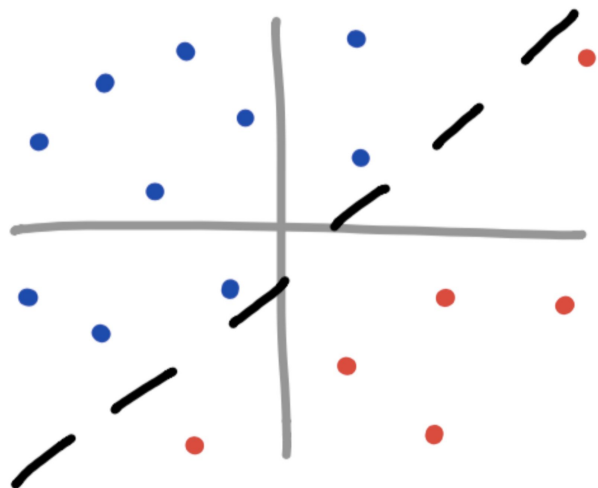
$$y_i = -1 : w^T x_i - c < 0$$

$$\begin{array}{l} x \mapsto \phi(x) \\ w \mapsto \phi(w) \end{array} \Rightarrow \langle w, x \rangle \mapsto \langle \phi(w), \phi(x) \rangle$$

$$K(w, x) = \langle \phi(w), \phi(x) \rangle$$

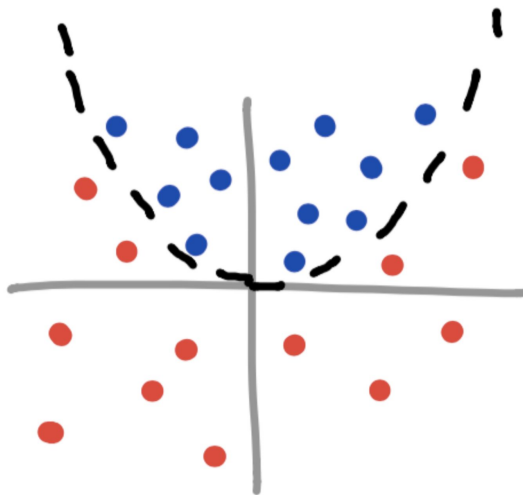


Kernel types



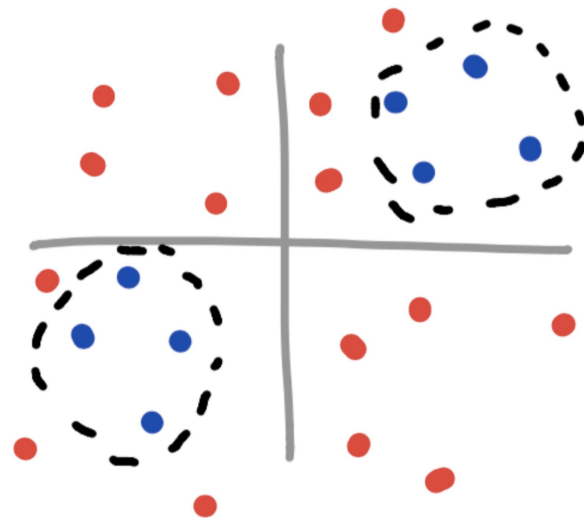
$$K(w, x) = \langle w, x \rangle$$

Linear



$$K(w, x) = (\gamma \langle w, x \rangle + r)^d$$

Polynomial

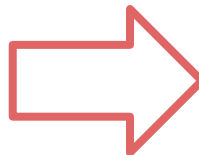
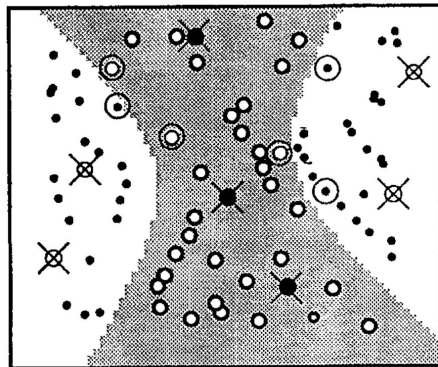
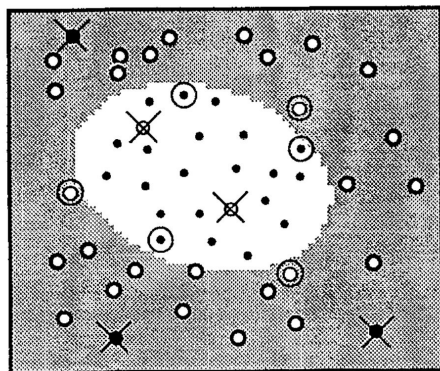


$$K(w, x) = e^{-\gamma \|w - x\|^2}$$

Gaussian radial basis function

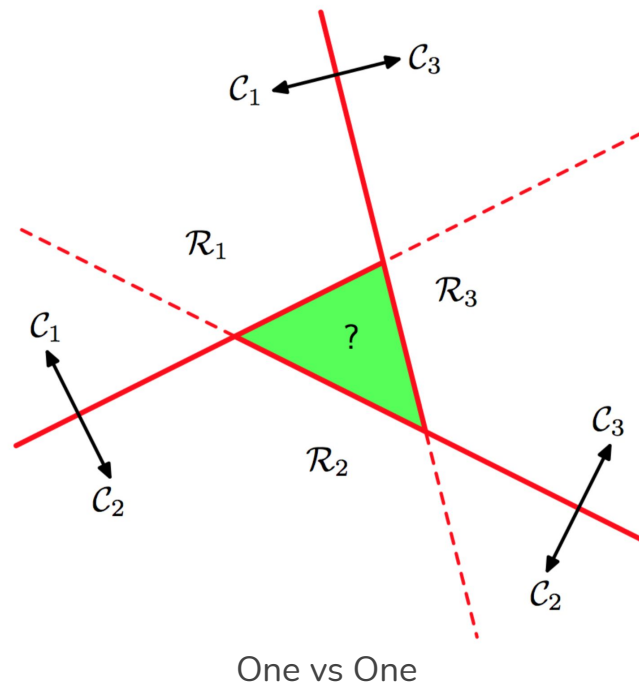
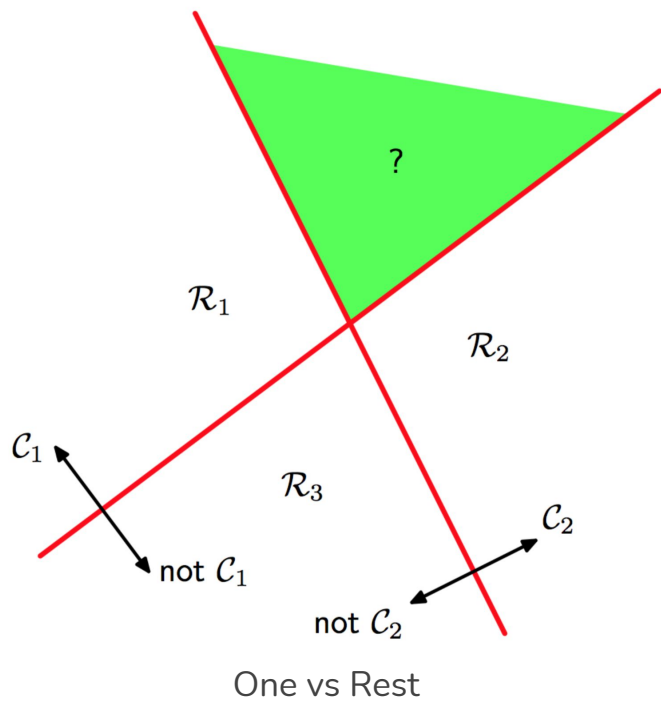


Current state



Multiclass classification

Multiclass strategies



Principal Component Analysis



Principal Component Analysis

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

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

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

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



Singular value decomposition

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

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

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

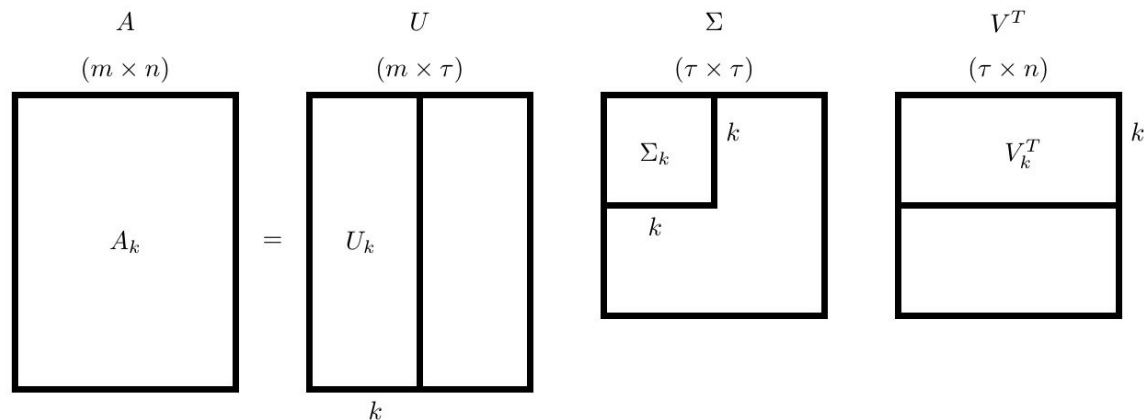
$$\|A\|_2 = \sigma_{\max}(A) : \|\Sigma' - \Sigma\|_2 = \sigma_k(\Sigma) = \sigma_k(X)$$

Eckart–Young–Mirsky theorem

Singular value decomposition

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

$$X = V\Sigma U^T \quad \sigma_k(\Sigma) = \sigma_k(X)$$



Eckart–Young–Mirsky
theorem



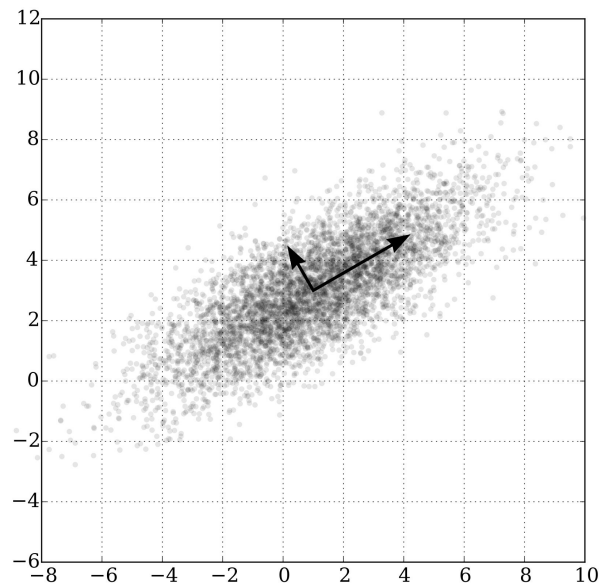
Another approach

Residual variance maximization

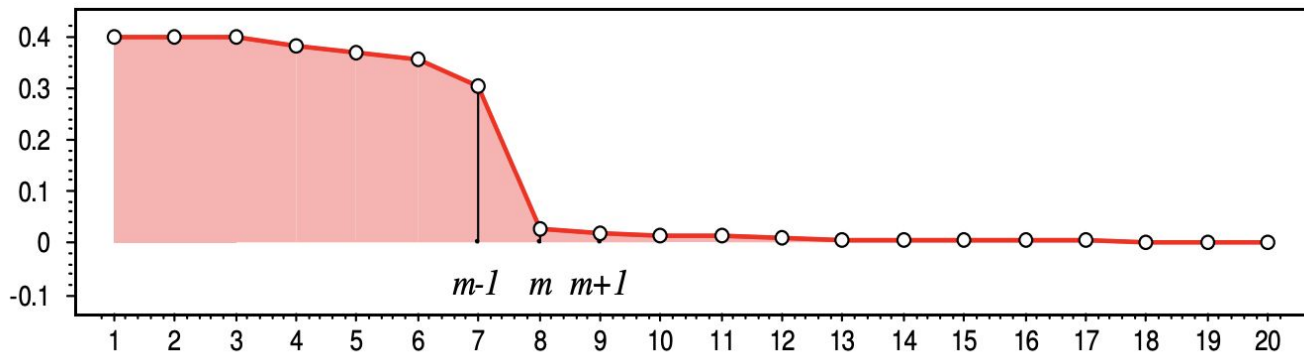
Take new basis vectors greedy

Same result for G and U

Always normalize data before PCA!!!



Dimensionality reduction



Get rid of low-variance components

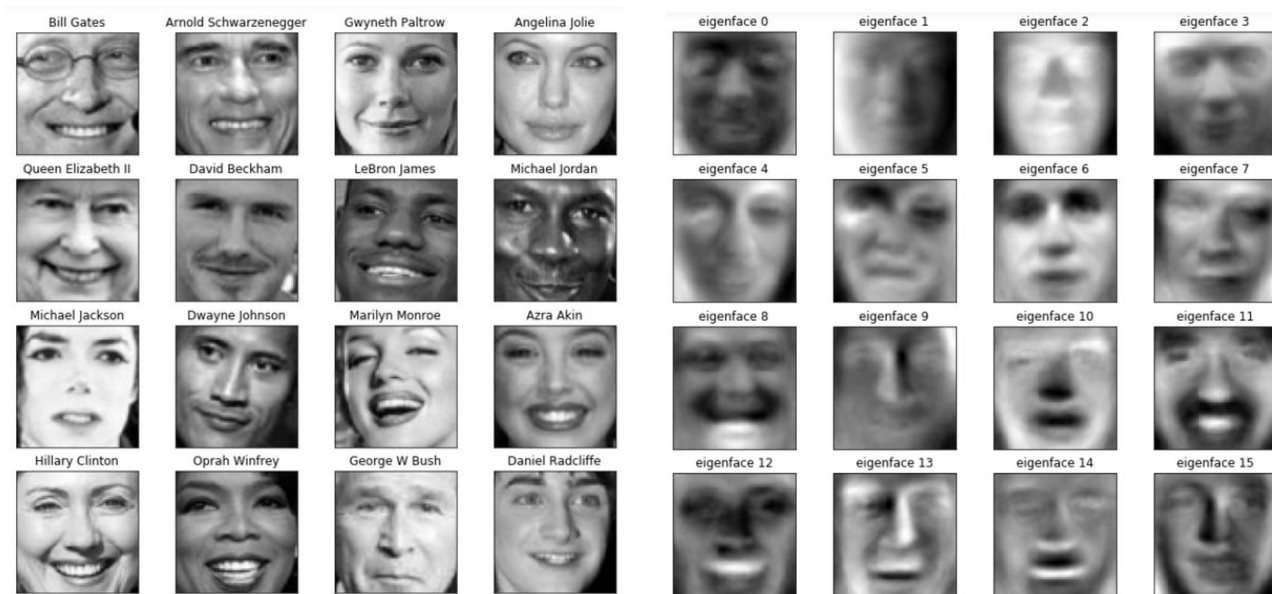
$$E_m = \frac{\|GU^T - F\|^2}{\|F\|^2} = \frac{\lambda_{m+1} + \dots + \lambda_n}{\lambda_1 + \dots + \lambda_n} \leq \varepsilon.$$



Dimensionality reduction

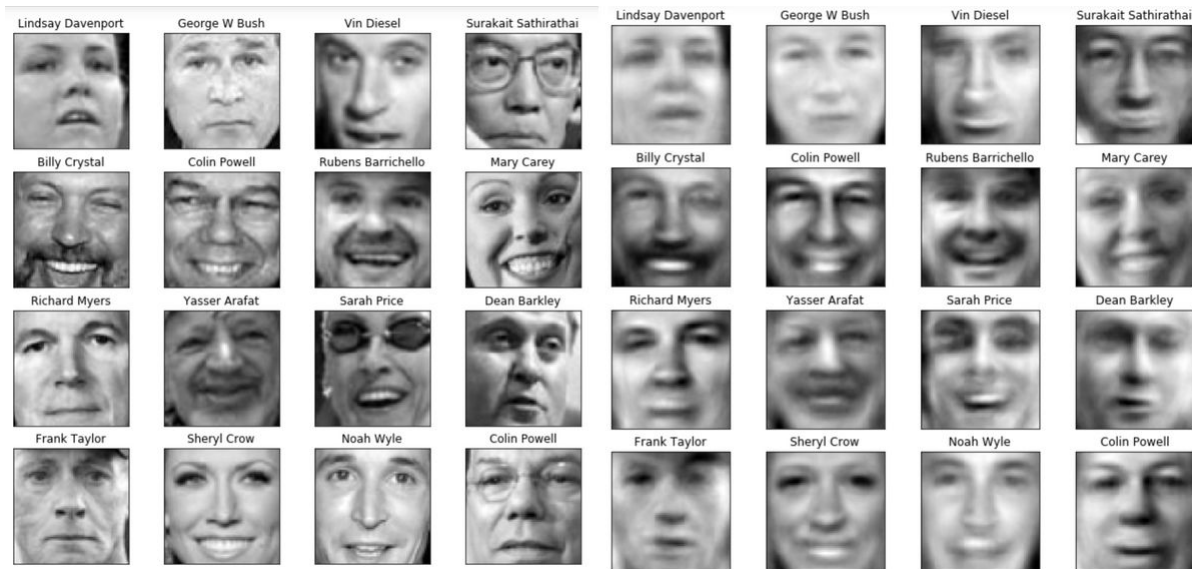
**Let's walk through
space...**

Dimensionality reduction



16 components

Dimensionality reduction



50 components

Dimensionality reduction

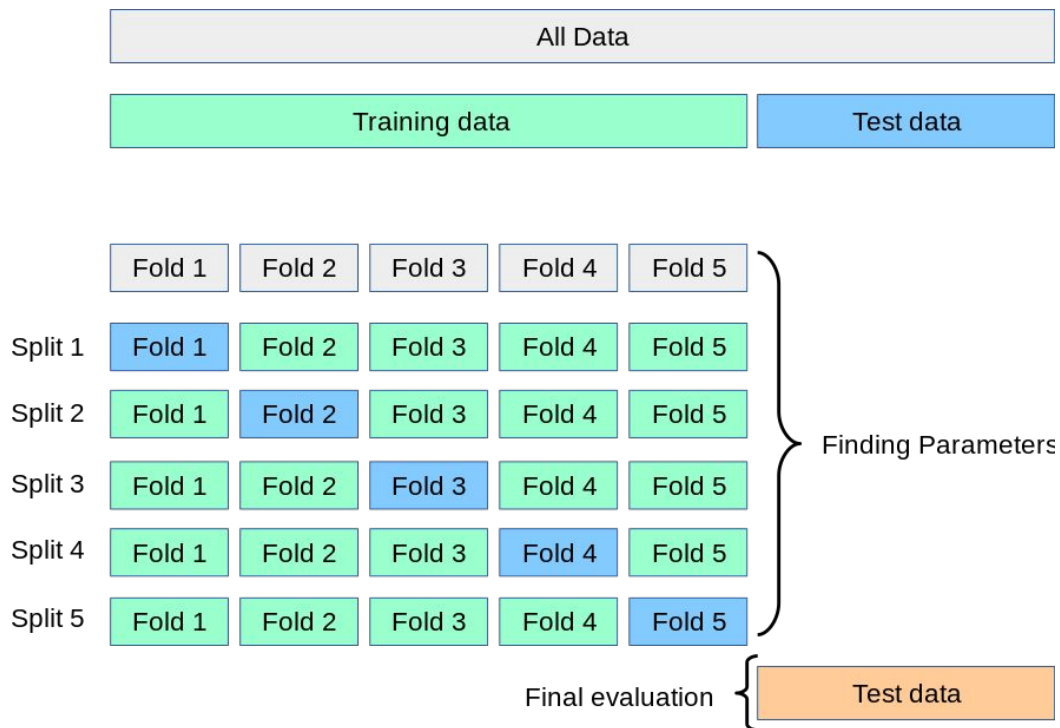


250 components

Validation strategies

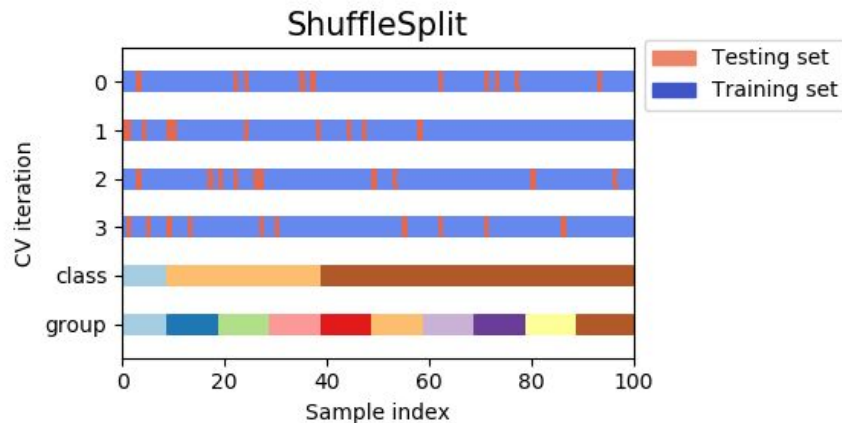
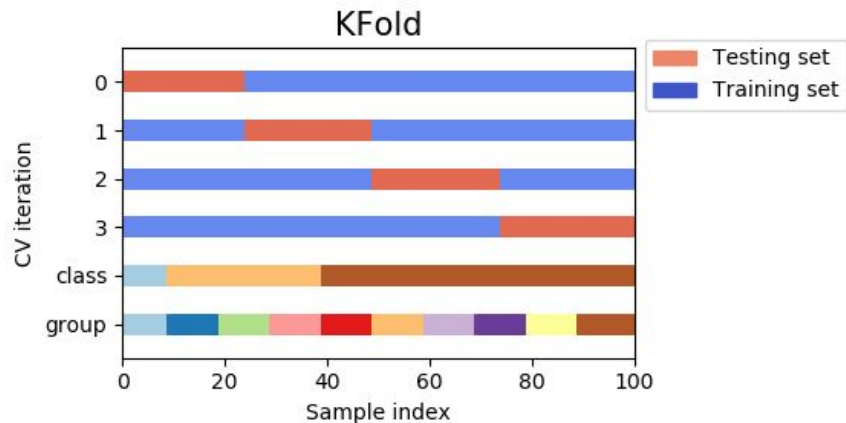


Validation strategies





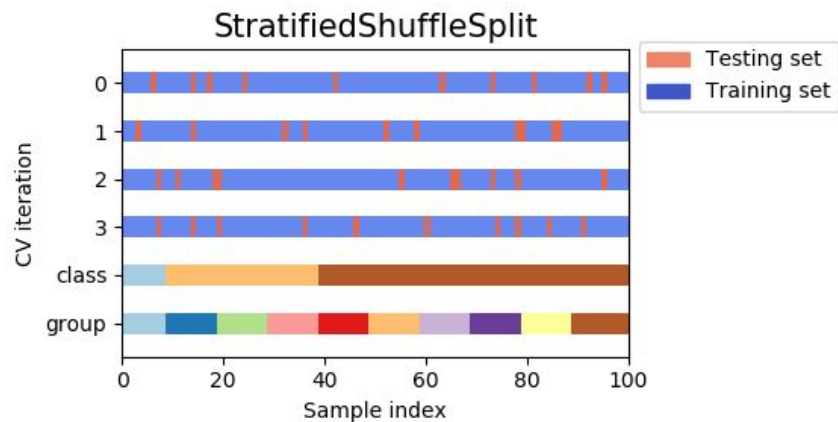
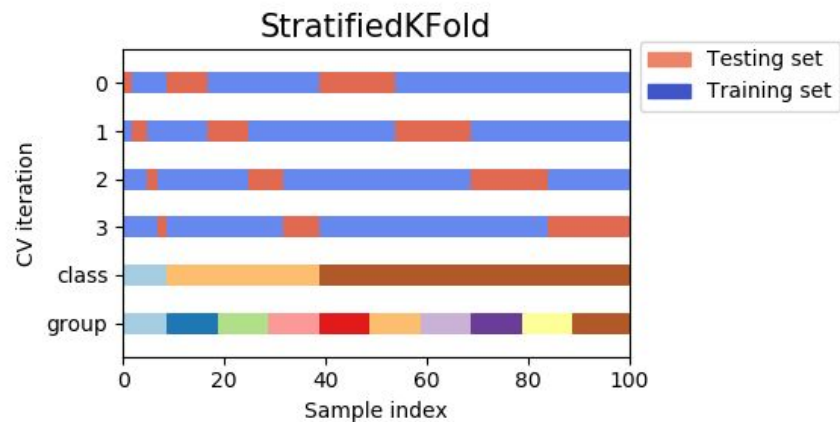
Validation strategies



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

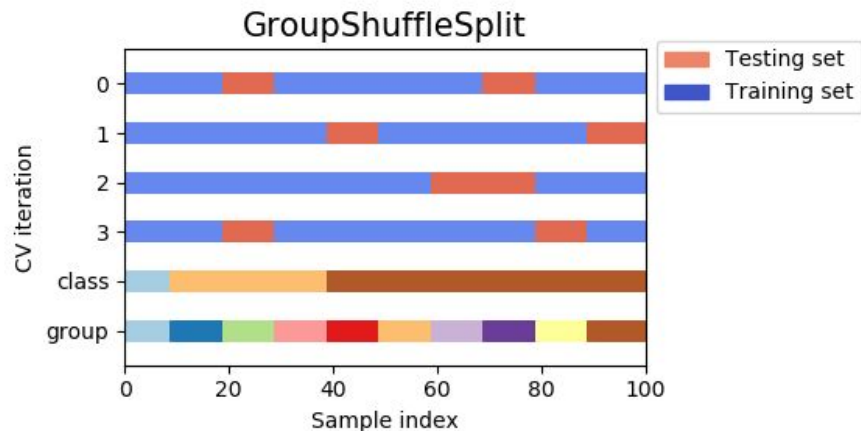
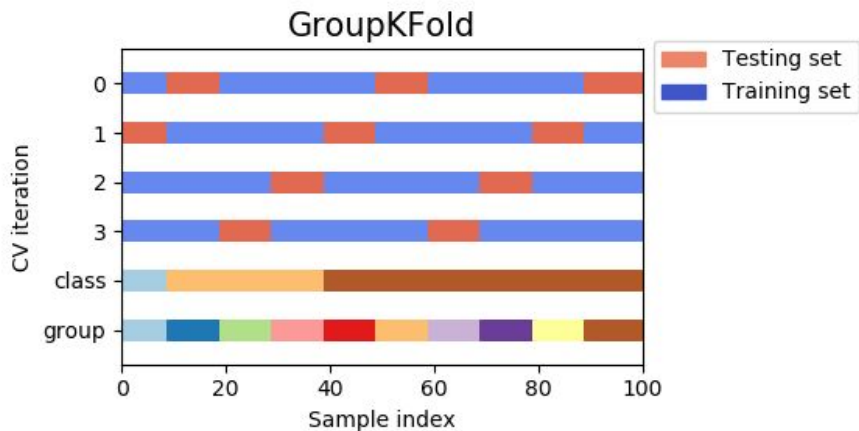


Validation strategies



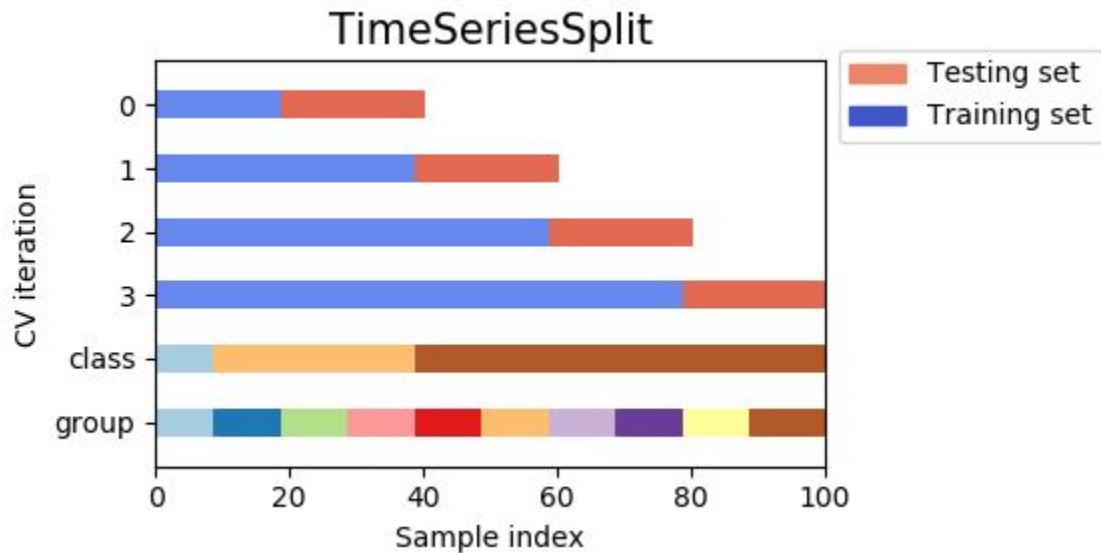


Validation strategies





Special case: timeseries



Never use `train_test_split` in this case!!!

Thanks for attention!